AUTOMATIC CAMERA CONTROL FOR CAPTURING COLLABORATIVE MEETINGS

by

Abhishek Ranjan

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto

Abstract

Automatic Camera Control for Capturing Collaborative Meetings

Abhishek Ranjan

Doctor of Philosophy, 2009

Graduate Department of Computer Science

University of Toronto

The growing size of organizations is making it increasingly expensive to attend meetings and difficult to retain what happened in those meetings. Meeting video capture systems exist to support video conferencing for remote participation or archiving for later review, but they have been regarded ineffective. The reason is twofold. Firstly, the conventional way of capturing video using a single static camera fails to capture focus and context. Secondly, a single static view is often monotonous, making the video onerous to review. To address these issues, often human camera operators are employed to capture effective videos with changing views, but this approach is expensive.

In this thesis, we argue that camera views can be changed automatically to produce meeting videos effectively and inexpensively. We automate the camera

view control by automatically determining the visual focus of attention as a function of time and moving the camera to capture it. In order to determine visual focus of attention for different meetings, we conducted experiments and interviewed television production professionals who capture meeting videos. Furthermore, television production principles were used to appropriately frame shots and switch between shots.

The result of the evaluation of the automatic camera control system indicated its significant benefits over conventional static camera view. By applying television production principles various issues related to shot stability and screen motion were resolved. The performance of the automatic camera control based on television production principles also approached the performance of trained human camera crew. To further reduce the cost of the automation, we also explored the application of computer vision and audio tracking.

Results of our explorations provide empirical evidence in support of the utility of camera control encouraging future research in this area. Successful application of television production principles to automatically control cameras suggest various ways to handle issues involved in the automation process.

To my parents

Sushila and Suresh

Acknowledgement

People (including myself) often asked me this question: is it worth doing a Ph.D.? Now I affirm that the sheer experience of meeting, working with, and being supported in diverse ways by so many amazing people makes this whole journey worthwhile.

It was an honor and pleasure to work with Prof. Ravin Balakrishnan (my Ph.D. advisor). Without him always encouraging, providing the best research facility a graduate student could possibly get, and pushing for excellence in research (and also in the art of food tasting), this thesis would have been impossible.

I would like to thank my committee members Prof. Mark Chignell for constantly showing the bigger picture and pointing to low level statistical analysis errors and Prof. Karan Singh for asking deep philosophical questions. I would also like to thank my youngest committee member Prof. Khai Truong for his insightful comments, and the external examiner Prof. Carl Gutwin (University of Saskatchewan) for carefully reading the dissertation and providing his expert feedback that made the final defense a unique experience. I cannot forget to thank Prof. Jeremy Birnholtz (Cornell University) for showing a sincere interest in this research project, actively helping me through his social science expertise, and be-

ing a remarkable collaborator.

All the members of the DGP lab definitely deserve a huge credit. In particular: Alex for $1.50 subs and working till late in the lab during his early Ph.D. days; Anand and Eron for proving that the non-academic job market has more hope than the academic one; Anastasia for always providing me with numerous tips and documents for the checkpoints; Gerry for all the utterly novel design and non-design ideas; Jack for interesting (and, of course, cynical) discussions about almost all worldly matters; Joe for being there with tonnes of research ideas and elaborate Tango tips; John for being a fun and easily accessible system administrator; Mike Wu and Nigel for giving me a company during the entire grad school journey; Pierre for introducing me to climbing and Matt for getting me back to climbing; Ryan for willingly volunteering in several of my system test-runs; Shahzad for squash games and regular discussions on how to finish a Ph.D.; Tomer for showing how to have a relaxed attitude towards everything; and Xiang for being a helpful officemate.

I would also like to thank my neighboring lab dwellers CM and Nilesh for letting me in the Hotel Database where late night ping-pong tourneys, fresh Cappuccino from the expensive coffee machine, and Jhapak's series of bizarre questions always recharged my spirits during long deadline hours. I also cannot forget to thank Marina for showing interest in my studies, pointing me to interesting psychology references, and organizing all the entertaining Salsa outings. And many thanks to Daniela, John D, and Vish for being regulars who occasionally accompanied in those outings. I am also fortunate to have met Mei who helped me recuperate after deadlines through fun tree climbing, canoing, and camping trips. I am also thankful to Celine for being one of those few mortals who have thoroughly read this entire dissertation!

Finally, I am grateful to my parents for supporting me in all my decisions that

led to the successful completion of my graduate studies. I dedicate this thesis to them. I am thankful to my sisters, Anjali, Jyoti, and Suman, and my brother, Rajesh, for always filling me with confidence and making all my vacation trips home purely rejuvenating.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Thesis statement

This dissertation provides evidence in support of the following thesis: *It is useful and feasible to change camera views automatically while capturing visual information from collaborative meetings.*

In this dissertation, we analyze the behavior of trained human operators controlling cameras to capture different types of meetings. Using the results of the analysis, we detect cues to determine visual focus of attention. Further, we derive heuristics from the cues and the analysis to control cameras automatically. We argue that this approach to automate camera control is appropriate under the following assumption: *Trained humans can capture visual information more effectively than any existing automatic camera control system.*

1.2 Motivation

Collaborative meetings are a frequent and necessary aspect of work in most organizations, with a 1999 white paper reporting that 37% of employee time in the United States is spent attending meetings, and that there are over 11 million business meetings held daily [Ver99]. Moreover, in a recent survey, 50% of the respondents indicated that attending face-to-face meetings was a waste of their time [Ver03]. These statistics highlight two problems. First, the amount of time that some employees spend in meetings makes it difficult for them to either attend all of them or get anything else done. Second, it can be difficult to recall what was said or accomplished in each one.

These problems arise primarily from how participants attend meetings and how archives of meetings are created for later recall. As participants are frequently traveling to attend face-to-face meetings, the cost-benefit ratio of attending meetings is also soaring. This in turn gives rise to a feeling of "wasted time". Videoconferencing technologies exist to support remote participation, but they have been repeatedly regarded unsuccessful in making remote participation as engaging as face-to-face meeting. Moreover, even if participants are collocated in meetings, the lack of effective meeting archiving makes it increasingly difficult to retain all the useful information for later reference.

At its roots, both videoconferencing and archiving systems face the same issue of capturing and presenting the "right" information at the "right" time. In the case of videoconferencing, this information needs to be exchanged among remote sites, and for archiving, this information needs to be made available offline for later access. Although there are various modalities of information present in any meeting room, our focus in this dissertation is on visual information captured in videos.

Capturing and presenting the "right" information at the "right" time can be

difficult since it requires understanding of various activities happening in the room and determining which activity is more important than others. Furthermore, this does not preclude the situation when multiple activities hold equal importance. Most current approaches use a single wide angle camera to capture the visual information from all possible activities, disregarding the varying importance of the activities. While this provides an overview of all the activities to remote participants in a videoconferencing or to reviewers of the archive, it fails to provide details of more important activities.

Alternative approaches employ a single pan-tilt-zoom (PTZ) camera or multiple static cameras pointed to different regions of the scene. This approach allows capturing details of various activities at the cost of significant human effort. Not only the importance of activities has to be determined by a human operator, but also the camera-view needs to be adjusted (for PTZ cameras) or selected (out of multiple static views) to capture the details of the activity deemed important.

In order to reduce the amount of human effort required, recently there has been increasing interest in automating the process of visual information capture. In this thesis, we will refer to this automation process as *automating camera control* and the result of this process as *automatic camera control*. We propose that this process essentially consists of the following three stages:

1. Detecting various activities and assigning importance to them,

2. Capturing visual details of as many activities as possible keeping into account the cost of capture, and

3. Presenting the captured visual information in an *engaging* manner.

These three stages of automation can be applied to a range of collaborative activities. In the next section we specify the type of activities we tackle in this dissertation.

1.3 Problem specification

From a camera control design perspective, collaborative meetings can be considered to vary in the *role of visual information*, *complexity of the scene* and *type of remote participation*.

1.3.1 Role of visual information

The role of visual information is characterized by how critical the visual information is in the smooth progress of the task. For example, activities with heavy use of physical artifacts (machine equipment, whiteboards, papers, *etc*.) can be considered as tasks for which visual information is critical since removal of visual information will severely impair the performance.

1.3.2 Complexity of the scene

The complexity of the scene can be determined by the possible number of simultaneous and spatially distinct foci of attentions. It should be noted that each of these foci of attention may or may not have critical visual component; they might involve verbal, tactile or other possible components. For example, a meeting room scene with three participants could be considered to have higher scene complexity than that with only one participant. However, if the meeting involves mainly verbal communication then the visual information associated with it is non-critical. Various collaborative meetings can be categorized along these two dimensions. In table 1.1, we provide examples of activities for each of the combinations.

	Visual information	
	Non-critical	**Critical**
Simple scene	Meeting involving a single participant and mostly verbal communication, single person repairing a simple machine	*Single person repairing a complex machine*, meeting in a foreign language involving gestures
Complex scene	*Meeting with multiple participants and mainly verbal communication*	Multiple participant repair task, single participant repair task with multiple spatially distinct components, Meeting with multiple participants and heavy use of artifacts

Table 1.1: A classification of collaborative meetings. The example meeting scenarios discussed in this document are *italicized*.

1.3.3 Type of remote participation

Collaborative meetings might involve two or more sites, with each site having single or multiple participants. Current videoconferencing systems often capture the visual information from a site and send it to all the remote sites. In a typical two site setting, each site is equipped with a screen to show the visual information captured as video from the other site.

As the number of sites involved increases, presenting visual information captured at each site becomes an increasingly difficult problem. Most current systems show the videos from all the sites tiled in some predetermined order.

The level of participation at the sites involved also influences how the visual information should be captured and presented. For example, in a two-site scenario, one site may have multiple active participants (complex scene and possibly critical visual information) and the other site may have only one participant (simple scene and non-critical information). In this case, an effective camera control will be more critical for the first site.

In this dissertation, we develop the notion of the three staged camera control and explore how it can be applied to various meetings. In particular, we consider two categories out of the four possible ones as shown in Table 1.1: (1) meetings with low scene complexity and critical visual information, (2) meetings with high scene complexity and non-critical visual information. We focus on the low-level problem of automatically capturing video of a single site effectively and presenting it to a passive or minimally active remote viewer. This problem is an essential first step in developing systems to tackle more complex scenarios.

1.4 Our approach

1.4.1 Understanding desired visual information

We explore what visual information is important for the viewer. Our approach is to obtain this information from the trained humans proficient in capturing different types of meetings.

Television production crews have expertise in capturing several types of collaborative meetings in the form of talk shows, cooking shows, and news shows. We analyze the television production literature (in Chapter 3) and consult with expert television directors (in Chapter 6) to gain knowledge of how they decide what is important visual information for the viewers.

Secondly, we set up a lab experiment to observe how a trained camera operator controls camera views to capture meetings with critical visual information (Chapter 4).

1.4.2 Using television production to capture and to show visual information

Once the desired visual information is determined, capturing and showing it is the next step in automating camera control. Once again we turn to experts in this field. We learn the various activities and principles involved in television production, such as setting up the cameras, framing shots, and switching between shots.

We used heuristics derived from the basic production principles of zooming and shot stability to design a camera control for visually critical tasks with simple scene complexity (in Chapter 5). We further applied more advanced principles to design a camera control for capturing a complex scene of a meeting room (Chapter 6 and Chapter 7).

1.4.3 Dissertation contributions

Chapter 3

- We interviewed two professional television directors and a professional television editor. We studied the television production literature and present a summary of the various aspects of television production that could be used in capturing collaborative meetings.
- For each of these aspects, we suggest how it can be used in meeting capture.

Chapter 4

- We experimentally demonstrate that performance in collaborative tasks can be improved by automating camera control.
- For collaborative tasks involving use of physical artifacts, we propose that hand movement can be used as an effective cue to control the camera.

- We demonstrate that participants modify their behavior to adapt to the camera control and propose implications of this behavior modification for both theory and practice.

Some of the results from this chapter were also published as a full paper in ACM CSCW 2006 [RBB06].

Chapter 5

- Building upon the prior work in Chapter 3, we design a prototype automatic camera control system that uses participant hand motion as a cue to automatically control the camera.

- We experimentally demonstrate that participants' performance significantly improves when using our automatic camera control system as compared to conventional fixed scene camera.

- Based on the results of the experiment, we propose that a loose coupling between activity and camera view can provide a stable shared visual space without losing the desired visual information.

Some of the results from this chapter were also published as a full paper in ACM CHI 2007 [RBB07].

Chapter 6

- We combine TV production principles with our three stage automatic camera control design process and present a camera control system for capturing small group meetings.

- We demonstrate that visual information captured by our system was comparable to that captured by experienced TV production crew.

Some of the results from this chapter were also published as a full paper in ACM CHI 2008 [RBB08].

Chapter 7

- In the concluding chapter of the thesis, we develop a practical automatic camera control system that uses unobtrusive vision and audio tracking to capture small informal meetings.

- We identify possible tracking errors that an automatic camera control system will face and propose graceful degradation and recovery techniques. These techniques are inspired from TV production principles and grounded in the principle of loose coupling between activity and camera view as proposed in Chapter 4.

Chapter 2

Background

2.1 Video for remote collaboration

One of the early beliefs about technology mediated communication hypothesizes that as the bandwidth of a communication channel increases, the efficiency of communication increases. This hypothesis was termed the *Bandwidth Hypothesis*. Despite its intuitive formulation, this hypothesis has been challenged several times in later studies [Whi03]. These studies have shown that the introduction of a visual channel (video) along with audio does not improve the time to solution or the quality of solution as compared to an audio only channel, for various collaborative tasks [WO97]. However, communication among humans is a complex process and cannot be measured only by the quality of the result.

McCarthy *et al.* [MM94] identified various different factors influencing communication and suggested some new measures for communication efficacy, such as common ground, mutual orientation, and content of conversation. Recently, Kramer *et al.* introduced linguistic features as a new measure of presence [KOF06]. They observed that participants' perceived sense of presence was highly correlated with the use of the pronoun *we* and local deixis (this, these, here), whereas

the pronoun *you* correlated with a lower sense of presence. Furthermore, these linguistic features were also shown to predict the sense of presence. Several other studies considered various such factors to explore the potentials of video as a channel for both formal and informal communication [FKRR92, Sel92, IT93, WO97, HL91, DJMW98, Fin97, Sch00, WS98].

2.1.1 Advantages

Isaacs and Tang [IT93] observed a group of people communicating over both audio only and audio-video channels and analyzed common benefits of the visual channel. The most common use of the video channel was to convey non-verbal responses, such as nodding the head to show agreement, looking away to show consideration, and leaning forward to show a request for further elaboration.

Furthermore, the visual channel made it possible for the participants to forecast responses and enhance their verbal description with gestures. In a separate study, Heath *et al.* [HLS97] observed that speakers continue to gesture and produce a range of body motions during the delivery of talk in video-mediated communication without caring if the remote participants could see them. These findings indicate that the presence of a visual channel helps speakers express themselves more naturally using hand gestures and rely less on intricate verbal descriptions. Later studies have further analyzed this difference in verbal description in communication with or without video [FKS00, Kuz92].

The visual channel was also found to be effective for interpreting the meaning of pauses which is difficult to do without a visual channel. Thus, video helped in maintaining awareness among remote participants. However, this finding assumes that the visual channel used for the communication can actually show enough information to explain the meaning of a pause. For example, Isaacs and Tang described a case where one of the participants (say A) spent two minutes

to search for an email and during that time other participants paused their conversation. If the visual channel could not show what *A* was doing for those two minutes then such a long pause would have been confusing for other participants.

Nardi *et al.* explored the utility of video from a perspective different from the traditional 'talking head video' concept [NKW+95]. They observed that in an operating room, a live video of the inside of a patient's brain is often used to coordinate activities. For example, using this video the scrub nurse could anticipate which instruments and supplies the surgeon would need. This view of video as data to convey information has tremendous design implications in collaborative physical tasks.

2.1.2 Limitations

By comparing the affordances of video technology for communication with that of everyday media (*e.g.*, light for vision, air for hearing, intrinsic up and down for orientation) used in face-to-face communication, Gaver pointed out various limitations of video medium [Gav92]. One of the main limitations was identified to be a restricted field of view that limits peripheral vision and perceptual exploration.

Peripheral cues play an important role in face-to-face communication. They are often used to notice changes in other's body, head or eye position and to coordinate actions accordingly [IT93]. In the video media, either a restricted field of view or a low resolution makes these cues hard to discern.

Another limitation of a video channel is the difficulty in dealing with physical objects. Not only is sharing or manipulating physical objects across a video link physically impossible, but the information about the objects pointed to in ones own environment is also difficult to convey through video channel. In face-

to-face communication participants often refer to some artifacts by using bodily movements or gestures. Furthermore, these objects or artifacts often become the subject of discussion. Although video could potentially convey such visual information to the remote site, the realization of this possibility is still a research challenge [IT93, HLS97].

These limitations of video medium give rise to a *communicative asymmetry* in video conferencing [HL92]. In two-way telephone communications, both participants know that they can interact only through a voice channel at all times, so they assume visual interactional conducts to be useless. In a video channel, however, a participant can see parts of the environment of the remote co-participant, but at times this view is not enough to convey all the information. This leads to an asymmetric communication where one participant is not always aware of what the other participant can see.

Heath *et al.* [HL92] studied these asymmetries in detail and also suggested that some measures such as explicit notifications and audio cues can minimize this asymmetry. Furthermore, a context aware camera control system could also address this issue.

A large body of research has shown that several limitations of video-mediated communication can potentially be addressed through better understanding of the task at hand, technological advancements, and innovative designs for a range of tasks. For example, multiple camera-views could possibly convey peripheral information more effectively [GSHL93, YCNB96] or a head-orientation based camera could support subtle communication using eye gaze [VWSC03, NC05, TOY05, Che02b, GM03]. Furthermore, a number of studies have also addressed privacy and awareness in their design [MRS+91, CFKL92, GMM+92].

In brief, researchers have expressed mixed opinion about the role of video in remote communication. However, there is growing evidence in support of utility of video. It is believed that visual information from collaborative meetings can be

captured more effectively if designers of the systems have a better understanding of meeting structures [Bux92]. In this chapter we focus on understanding two types of meetings: meetings with critical visual information and meetings with complex scene.

2.2 Understanding meetings with simple and complex scene

An extensive amount of research has been done to study various types of collaborative meetings. Despite variations in the appearance, they share a number of common aspects and are all guided by some implicit rules based on the norms of organizations [McG90]. In this section we identify three major aspects of meetings and discuss them in detail. They primarily involve major activities in a meeting, behavior of participants, and control-flow. These aspects point to some non-verbal cues that could be used to detect potential focus of attention in a meeting.

2.2.1 Coarse level activities

Participants perform a wide range of activities during meetings depending on the purpose of the meeting. However, conceptually, their activities are similar across most of the meetings; they present information, review information, generate and analyze solutions, plan, schedule, make decisions, track actions items, and prepare reports. Poltrock *et al.* analyzed the meetings of several physically collocated teams and divided their activities into three categories [PE97]:

- **Work-centered activities**: These activities include presenting, reviewing, creating, editing and annotating work products.

- **People-centered activities**: These activities include members greeting each other, interacting socially before settling down, introducing new people to the group *etc*.

- **Meeting-centered activities**: These include scheduling a meeting room, preparing for the meeting, notifying members, starting equipment, managing resources for new members during the meeting, distribution and archiving of meeting notes *etc*.

Thus, coarse level activities form the shape of a meeting and determine the workflow. Most of them require some amount of conscious effort from the participants. Apart from these major activities, meetings also involve some activities which are subtle and performed effortlessly by the participants.

2.2.2 Subtle activities

In face-to-face collaborations participants communicate with one another by coordinating, most often effortlessly, various vocal and visual activities. Heath *et al.* explored these subtle activities which make any collaboration successful [HLS97]. They could be studied under the following three categories:

- **Alignment toward a focal area**: Collaborators are always focusing on a common object or artifact (*e.g.*, a document or a computer screen) by coordinating their bodily movements, gestures, facial expressions, and gaze.

- **Peripheral awareness**: Participants are always aware of their surroundings even when they are concentrating on a focal area. Awareness is essential for maintaining a continuous communication throughout a collaboration. Monk and Watts further studied the role of awareness in the context of the peripheral participation in communication [MW98].

A peripheral participant can overhear or see what is going on but is not engaged in the current task. During the course of a communication, participants often exchange the role of the peripheral participant by using peripheral awareness.

- **Transition from individual to collaborative**: Most of the activities during a collaborative meeting are either individual or collaborative, but what makes it a coherent event is the smooth transition between these two activities. In face-to-face meetings, participants often make such transitions by getting involved in multiple, interrelated activities.

Thus, subtle activities primarily involve low level communication activities which are ubiquitous and performed during the course of the entire meeting. They form a major factor to influence the control flow of the meeting. Another major factor is the shared resource control.

2.2.3 Shared resource Control

Collaborative meetings often have shared resources, such as a whiteboard or a projector, and all participants coordinate to make maximum use of that. Austin *et al.* [ALM90] observed that such shared resources often become communication channels and the control of the resource becomes a means to influence other group members. For example, in a collaborative writing task, meeting members assuming the role of a scribe do not participate with as high a verbal frequency in the meeting as when they are not acting as a scribe [Man88].

Similarly, in the case of a whiteboard, the person using the whiteboard assumes the role of the discussion leader. Several factors have been identified to determine who controls the shared resources, such as the group's usage strategy, social influence, and technology proficiency [ALM90].

In short, meetings have been extensively studied to explore both its machinery and the functioning. Most of them share some common aspects and that is what makes it feasible to design systems to support them. On the one hand, coarse level activities indicate major stages of a meeting. On the other hand, subtle level activities and shared resource control outline the process through which the stages are reached. Furthermore, the nature of a meeting depends on the dominance of one or more of these aspects. For example, in a lecture room meeting, coarse level activities (2.2.1) play an important role, but shared resource control (2.2.3) does not. However, in a boardroom meeting or a team meeting, subtle activities (2.2.2) and shared resource control play major roles.

The understanding of the structure and process of meetings formalizes the design requirements to support them over a video channel. For example, the role of a whiteboard as a shared resource indicates that the remote participants must be provided with the information about the person who is using the whiteboard. Similarly, transitions from individual to collaborative are subtle but very important for the work flow. Therefore, the system to capture collaborative meetings must capture this transition event and convey this information to the viewers. However, while these studies help us determine the information that needs to be captured and conveyed to the remote participants, they do not clearly suggest how this could be done. In order to explore the later problem, in the following section we discuss its technological and design aspects with respect to camera control.

2.3 Camera control for meeting capture

Existing camera control strategies vary from one another in various aspects. Here we identify two major dimensions along which we can place most of the strate-

gies. The first is camera and view setup, and the second is detection and capture of events in a meeting.

2.3.1 Camera and view setup

Camera and view setup for a camera control system includes the type and number of cameras used, number of views captured, and number of views selected for the final video. Various existing systems can be divided into the following categories based on their camera and view setup.

Single camera, single view

The most basic video conferencing systems use a single camera to capture the scene at the local site and send that single view to the remote site. Due to its simplicity and low cost, this setup is also the most commonly used setup [Pol08]. This is also known as the "talking head" form of video conferencing. Fish *et al.* used this common setup to explore informal communication across a distance in the VideoWindow system [FKC90]. By capturing a high aspect ratio video from a specially designed camera and displaying it at the remote site on a large display (8 feet wide and 3 feet high) they realized a sense of co-presence without physical proximity. This was one of the first video conferencing systems to show strong evidence that *"technology can provide, to a degree, an increased sense of shared space between remote coworkers"* [FKC90].

Two cameras, two views

Single static or pan-tilt-zoom (PTZ) camera based systems fail to provide peripheral information and do not scale well to multiple users. Therefore, some systems use two cameras: a static camera for a wide view and a controllable camera for close-ups [YCNB96, KKT94, Bux92, Bux95, Kuz92]. In this case, the

remote user is shown both of these views either in a picture-in-picture format [KKT94, Bux95, Kuz92] or separately with links [YCNB96, LKF+02, LLK+03].

Foote *et al.* engineered a panoramic video camera, called FlyCam, by combining various inexpensive video cameras in an array [FK00]. They also proposed methods to correct lens distortion and merge videos to create a panoramic video. The FlySpec system used this camera in combination with a PTZ camera to show both wide and detailed views of meetings to remote participants [LKF+02]. The system shows a fixed wide view of the scene and allows users to select any circular or rectangular region in that view. Once selected, the PTZ camera moves and zooms to show the detail of the selected region below the overview window (see Figure 2.1).

Figure 2.1: Left-top: Panoramic video camera FlyCam, Left-bottom: FlySpec system consisting of a PTZ camera and a FlyCam, Right-top: A wide view captured by FlyCam with rectangular region selected by a user, Right-bottom: A zoomed-in view of the rectangular region captured by the PTZ camera. Images taken from [FK00, LKF+02].

Despite the presence of cues (*e.g.*, a link or a highlighted window) for showing the contextual relationship between the overview and the zoomed-in view, the use of two separate windows always causes some amount of context switch. In the CoMedi system [CBC+99], Coutaz *et al.* explored distortion based visualization techniques to view focus and context using various computer vision algorithms. The system shows a high-resolution focal view of remote participant's face (captured by a PTZ camera) as a circular or rectangular region inside a low-resolution overview (captured by a static camera) and blends these two views using alpha channel coding (see Figure 2.2).

While this blended visualization allows the user to view focus with context without switching attention between two separate views, its practical utility is limited since only the face of a remote user could be seen in high-resolution. In real life scenarios various parts of the remote participant's body and environment could be of focal interest. However, this idea could be extended to facilitate focusing on different parts of the remote environment.

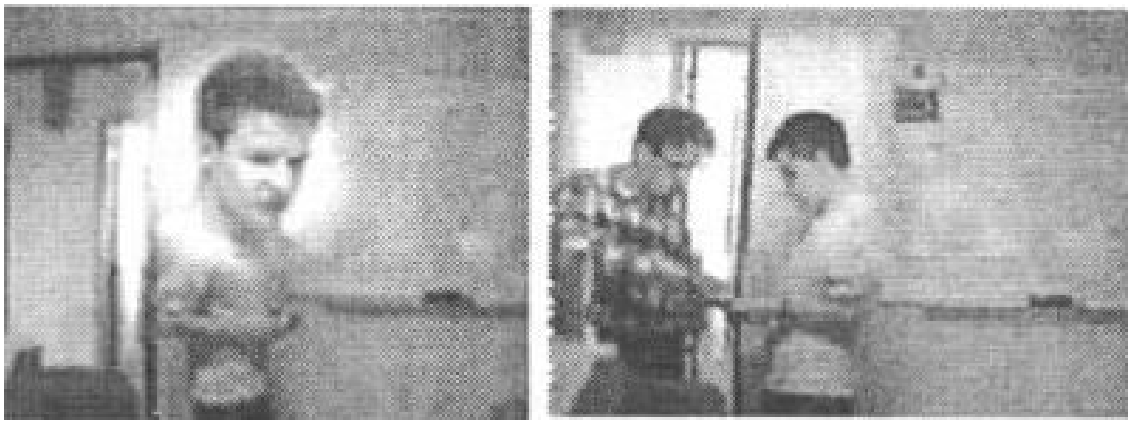


Figure 2.2: Left: Zoomed-in circular view of the face blended with the surrounding, Right: Zoomed-in rectangular view of two users. Images taken from [CBC+99].

Multiple views

Numerous systems have used multiple cameras to capture a small meeting and send several views of the scene to the remote site. Multiple views from different cameras have also been associated with multi-party video conferencing systems in which the videos from different remote sites occupy different locations on the screen [Che01, GMR95, DB92]. The users can view all the sites or some selected ones at once and can also interact with one or more sites. Here we limit our focus on the utility of multiple cameras and multiple views for two-party video conferencing systems only. This could later be extended to include multiple sites.

Nguyen and Canny's MultiView system utilized multiple cameras and a special display screen to communicate gaze information of multiple participants to the remote site [NC05]. The life-sized images, captured by cameras at the remote site, are projected onto the screen and the screen's main function is to display the image produced by a projector only to a person in a specific viewing zone (see Figure 2.3). Previously, MAJIC system has achieved gaze preservation in video conferencing for single user using two cameras [OMIM94].

Recently, the emergence of omni-directional cameras have made it possible to capture a 360-degree view of the scene using a single camera. Rui *et al.* used this camera to capture a single image of all the participants sitting around a table [RGC01]. Using vision techniques, an image of each participant was separated from this single image. Thus, remote participants could see all the participants using a single omni-directional camera (see Figure 2.4). They also used this technology later in a more elaborate meeting capture and broadcasting system [CRG+02].

In the HERMES system, Inoue *et al.* proposed a specific spatial arrangement of multiple cameras and monitors to integrate the images of the remote users with the local users [IOM97]. Figure 2.5 shows the arrangement of chairs and

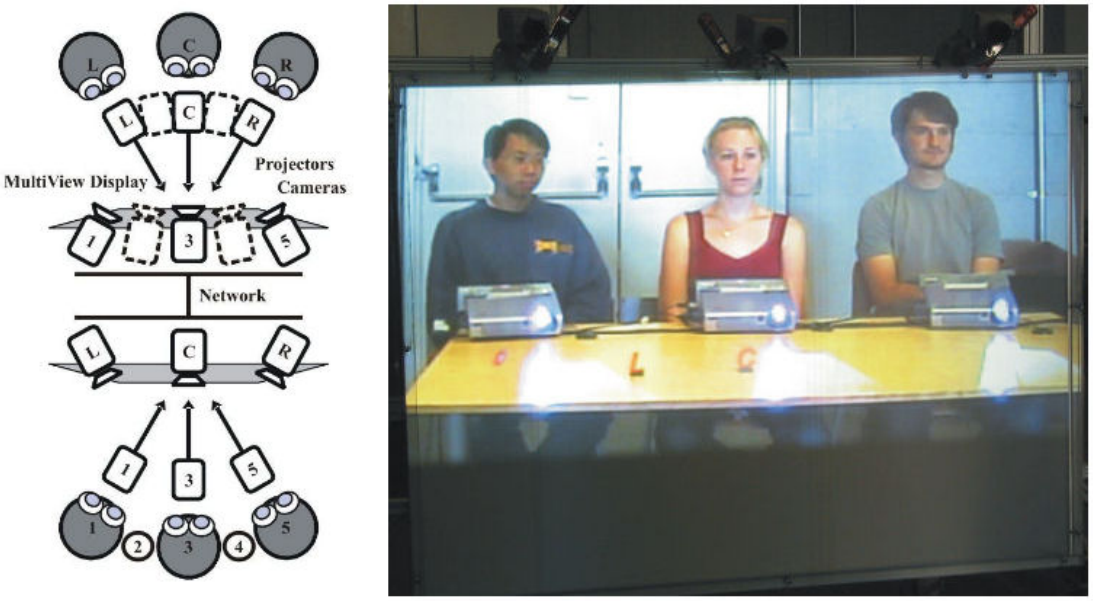


Figure 2.3: Left: The MultiView system's seating arrangement with screen and projector, Right: A view of live MultiView at work. Images taken from [NC05].

monitors used in HERMES. The video cameras are placed next to the monitors approximately at the height of the eyes of the participants and each participant has his/her own monitor to look at. All monitors showed the same video and there was just one sound source. The system showed three kinds of shots in the video: a whole shot, speaker shot, and a non-speaker shot.

Thus, multiple views from multiple cameras facilitate a wider coverage of the remote site. The users could also enjoy the freedom to choose and focus on one of the views provided by the system. However, the camera setup with multiple views has been criticized for consuming higher bandwidth. Sending multiple views across the network requires not just higher bandwidth, but also a high degree of synchronization [Jou02, MS99].

Considering these issues, Dourish *et al.* investigated the potentials of multiple views to increase the sense of awareness in remotely located collaborators without requiring very high bandwidth. They addressed these issues by updating the views once every few minutes. Similarly, in the Argohalls system the users could join one of the virtual halls (represented by an icon box) and their images would

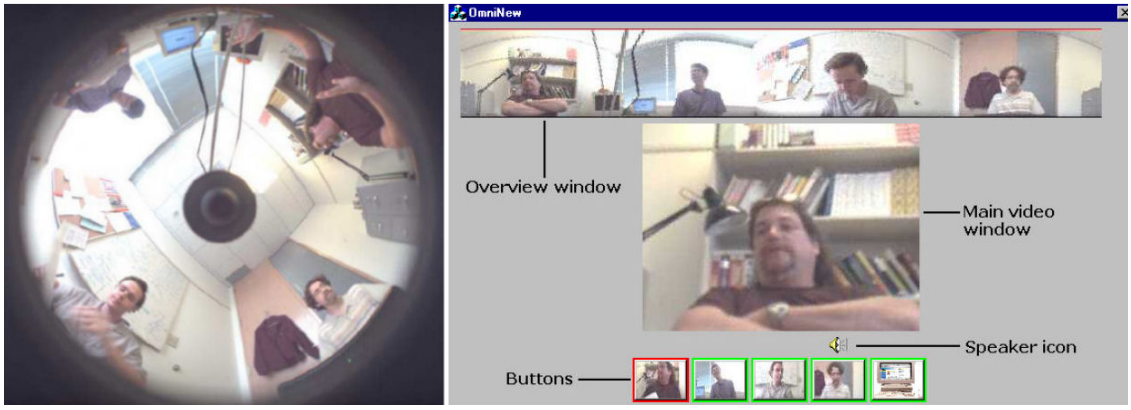


Figure 2.4: Left: A 360° frame captured by an omnidirectional camera, Right: each users view separated in the interface. Images taken from [RGC01].

be shown (and slowly updated) in that room [GMR95].

Studies have shown that these systems increase awareness among remote collaborators despite the low update rates. However, it is not clear how useful multiple views would be in live video communication, because the monitoring of multiple live views itself could potentially be interfering with the actual communication.

Multiple cameras, single view

Recently, some systems have addressed the aforementioned issue of multiple views by selecting one appropriate view out of several captured views and sending only that view across to the remote site. A single view is selected to either preserve gaze direction of the participants [BSS97, VWSC03] or to capture the most interesting event in the remote site [CRG+02, RHGL01, LRGC01, Bia04b, MR02, IOM97].

2.3.2 Event detection, view selection

When multiple cameras are covering the entire scene independently, automatic selection of a single most appropriate view to be sent to the remote site at a par-

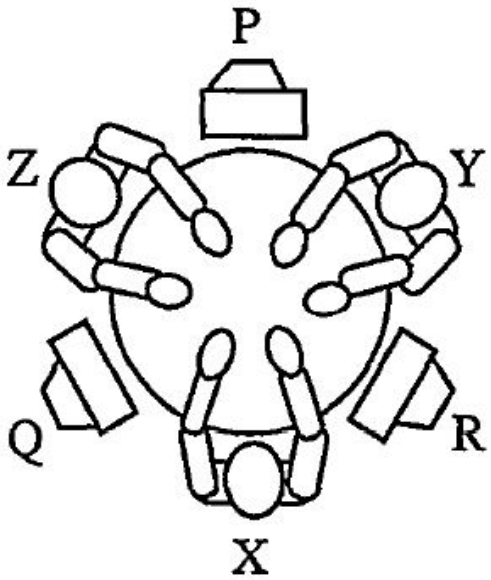


Figure 2.5: Hermes seating and monitor arrangement. Three participants sit in a circle with one monitor between any two participants. Image taken from [IOM97].

ticular time is still an open research problem. One naive strategy would be to periodically cycle through all captured views. However, this strategy will not only fail to capture the right events at the right time, but also result in a boring presentation of an otherwise lively scene.

Previous systems have approached this problem in various ways, but they have one common aspect: automatic detection of interesting events. The definition of interesting event varies from system to system. However, just the detection of events does not solve the problem, since there could be multiple interesting events occurring at the same time or multiple cameras could be capturing the same event. Therefore, selection of an event and an appropriate view is also another problem that a camera control system needs to solve. Here we discuss various strategies used by previous systems to address these problems.

Strictly speaker based

In section 2.2.1 we saw that verbal communication plays an important role in meetings and a speaker is often the one who draws the attention of most of the participants. Therefore, a view of the current speaker could potentially be of utmost importance for remote participants. Based on this conclusion, various systems show a view of the current speaker [BSS97, MR02].

The LiveWire system [BSS97] used voice-activated image switching such that all non-speakers see the current speaker in full screen. The current speaker sees the last speaker and only one person owned the screen at any given time.

Empirical studies with this system revealed some interesting drawbacks of the purely speaker based view selection. Since the participants could see only one person at any given time, they lost a sense of the larger group. They were also not aware of any non-verbal activities at the remote site. Furthermore, if the speaker was changing frequently then the automatic switching was often found to be distracting and confusing.

In order to address some of these issues, the Brady Bunch system incorporated various views of the scene in a separate screen [BSS97]. These view were refreshed every 5 minutes. The remote participants could select one of the views to see its live version in full-screen (see Figure 2.6). The advantage of this solution was that it conveyed some visual information to the remote participants without consuming a large amount of resources. However, the remote participants still had the burden of scanning all the views and looking for interesting events, which could potentially interfere with their active participation.

The AutoAuditorium system employs speaker tracking to automatically shoot videos of lectures [Bia04b, Bia04a]. The system uses two cameras: a static camera to show slides and a PTZ camera to track the speaker. The static camera could also determine if the projection screen is turned on or not, and if the projection

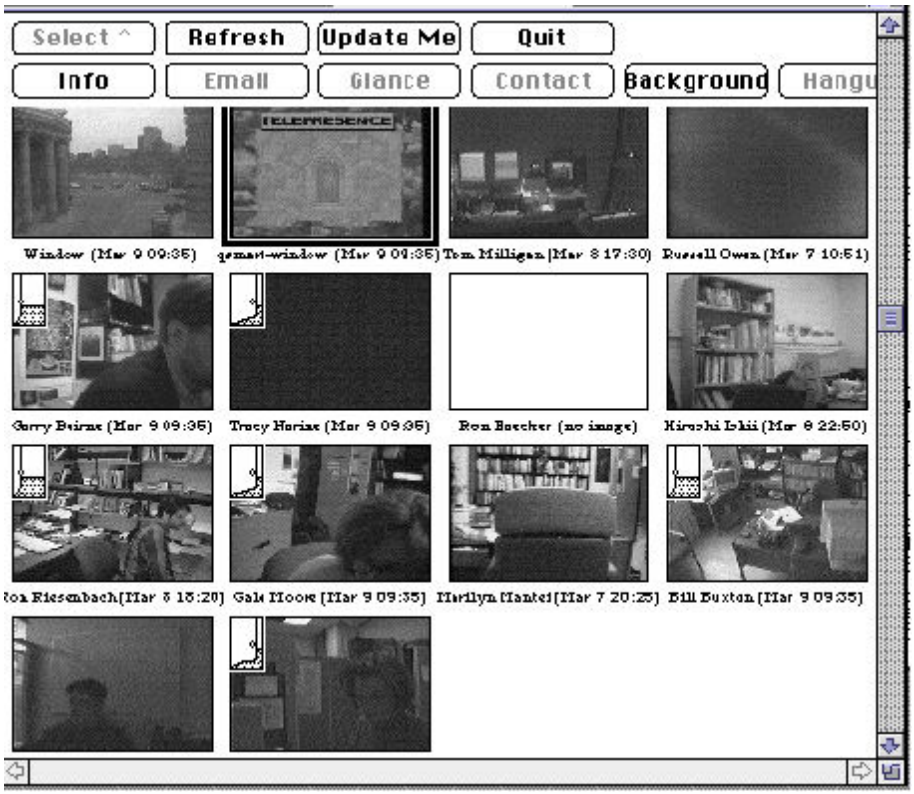


Figure 2.6: Brady Bunch's multiple view display. Each view was refreshed once every 5 minutes. Image taken from [BSS97].

screen has changed. This allows the system to not just follow the speaker, but also detect events from the projection screen.

Thus, speaker based systems succeed in capturing the most important part of a meeting: the speaker. There are, however, some drawbacks. Firstly, in a purely speaker based system, the video may fail to capture contextual information in case of a long monologue. Secondly, a lively discussion could result in a confusing video with quickly changing views. This indicates that camera control for meeting capture requires more information than just the identification of the speaker.

Cinematography based

When a professional camera crew shoots and broadcasts a live talk-show, spectators often find it interesting and engaging, whereas the video produced by most

of the conventional conferencing systems are boring [IOM95]. Clearly, knowledge of cinematography plays a tremendously important role in the former case. This observation has led various researchers to incorporate some rules from cinematography and "film-language" [Ari76, Jon71] to meeting capture systems.

He *et al.* [HFS96] developed an automatic cinematographer that can shoot video of a virtual environment by following some cinematography rules. By determining position, orientation, and role (speaker or listener) of characters, and also the events in the surrounding virtual environment, the system can select a cinematography rule to frame a shot.

Similarly, Drucker at al. [DZ95] designed a camera control method for virtual environments which can be used for shooting virtual scenes based on the rules of cinematography. Although, this method does not automate the capture, it does provide modules that can be used to automatically capture scenes.

Several others have also attempted to include cinematography principles to capture videos from virtual environments [TBN00, Dru94, DGZ92] primarily because virtual scenes allow low-overhead camera placement and free camera movement. Furthermore, detecting activities in virtual environments is tractable, whereas detecting them in real world often involves several open research problems including tracking and recognition.

Inoue *et al.* [IOM95]applied the rules from TV program production to a video-conferencing system in a semi-automatic manner where a human operator identified a speaker among various conferees sitting in a row and the automatic camera controller decided how to shoot the video.

In order to design the controller, they analyzed various TV debate programs and came up with a list of eight different kinds of shots: speaker shot, speaker and neighbor shot, speaker and listener shot, listener shot, listener and neighbor shot, third person shot, third persons shot, and whole view shot. Further, they calculated the probability of switching from one shot to another and distribution

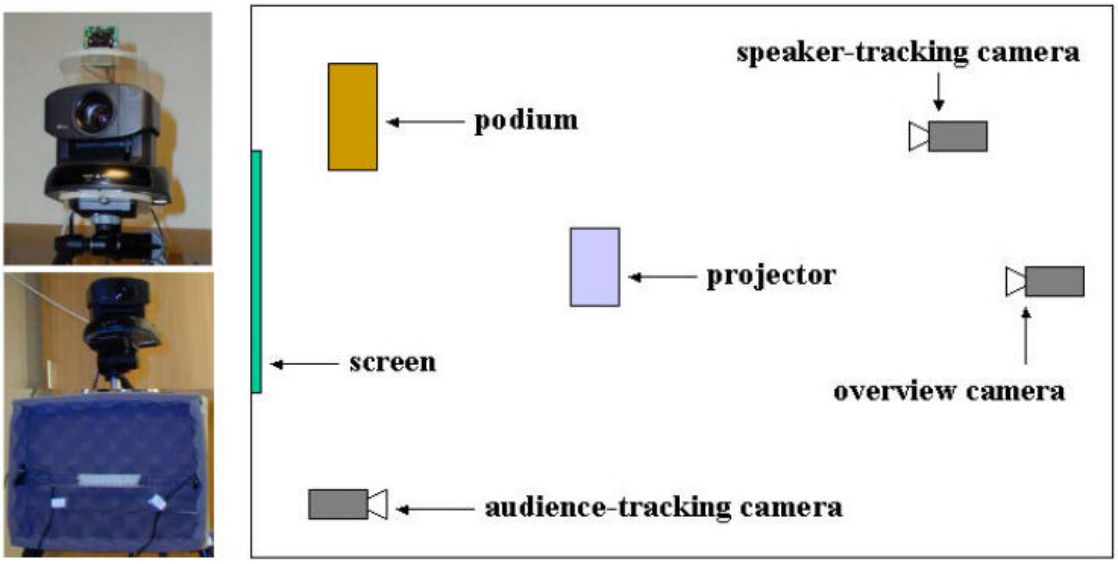


Figure 2.7: Left-top: speaker-tracking camera, Left-bottom: audience-tracking camera with a mic-array, Right: Lecture room layout. Images taken from [LRGC01].

of duration for each shot type by watching several TV shows. The controller was provided with of all this information and during runtime, based on the speaker and listener information provided by the operator, it determined the most probable shot along with the shot duration.

This system demonstrated a novel paradigm to make meeting video visually pleasing and raised various issues in the implementation of this paradigm, such as how to show the listener if he/she is at the remote site, what to show when there is no speaker at all, and how to introduce other roles that conferees play (*e.g.*, chairperson, notetaker *etc.*).

Rules of cinematography have also been applied to control cameras in lecture room environments by Liu *et al.* [LRGC01] and Rui *et al.* [RGG03, RHGL01]. While Inoue *et al.* analyzed TV debate videos to determine the most common shots, Liu *et al.* interviewed five professional video producers and collected rules for various range of activities including camera setup and video editing. Firstly, cameras were set up in a lecture room according to these rules (see Figure 2.7). Further, automatic speaker tracking (using computer vision) and audience track-

ing (using sound-source localization) were incorporated in the system to keep the speaker in the shot while he/she is walking and to focus on an audience member when he/she is asking a question. The system had four types of shots:

- Overview or establishing shot
- Speaker shot
- Shot of the audience member asking question
- Random audience shot

For shot transitions there were the following rules:

- Video should start with an establishing overview shot of the room.
- Two consecutive shots should significantly differ in the number of people and view direction.
- System should not transition to a dark shot.
- Each shot should have a minimum and maximum time length.
- System should transition to an overview shot when all other cameras fail.
- Audience member asking the question should be shown promptly.
- Short random audience shots should be shown occasionally.

The user study reported in this work primarily compared the performance of the automatically controlled camera system with an operator controlled camera. The results showed that the operator controlled camera had significantly better speaker tracking, whereas audience tracking was not significantly different. Furthermore, the system passed the Turing test in that users could not clearly differentiate between automatic and operator conditions.

In an extension to this system Rui *et al.* [RGG03] enhanced various technical aspects of the system and conducted an extensive user study which included both professional videographers and non-expert users. Several lectures were captured by the automated system as well as four professional videographers. In order to learn the rules applied by them during the capture, the videographers were interviewed after each recording and had some of their videos reviewed.

The result of the study showed that automatic speaker tracking needed further improvements and automatically framed shots lacked aesthetics. Further, the authors discussed various camera configurations and shot rules for similar lecture room scenarios with varying number of participants (see Figure 2.8).

Thus, this work explored the application of cinematography in capturing lecture room scenarios. The authors also attempted to extend this idea and briefly suggested camera configurations for various other scenarios, such as a medium size lecture room and a small meeting room. However, from our study of meetings (Section 2.2) we learned that these scenarios significantly differ from each other. Therefore, application of cinematography to capture each of them would also introduce a variety of potential research problems.

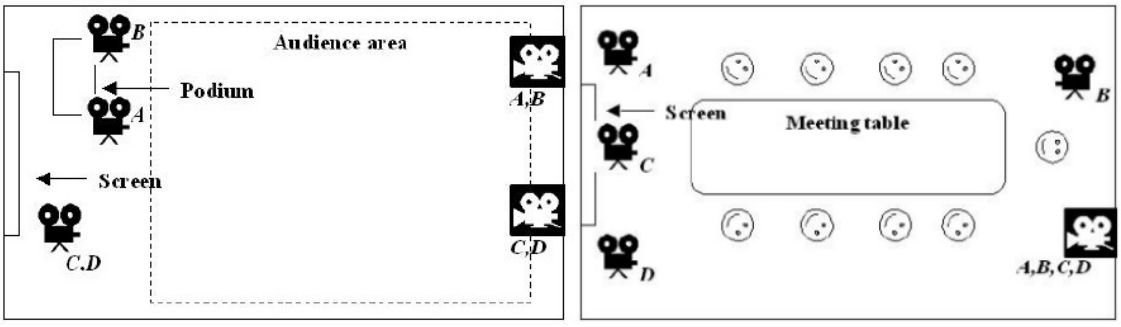


Figure 2.8: Suggested camera configurations for a medium size (50 people) lecture room setting (Left) and a small (10-20 people) meeting room (Right). Images taken from [RGG03].

Rui *et al.* [RGC01] also applied some of these cinematography rules, such as minimum and maximum length of a shot and speaker transitioning rules, to a

smaller meeting capture system using an omni-directional camera. This system was later incorporated into a more elaborate meeting capture and broadcasting system [CRG+02].

Inoue *et al.* [ISOM04] also evaluated the role of two cinematography rules in automatically shooting a face-to-face meeting of five people around a table and their findings further indicate the benefits of applying cinematography to automatic meeting capture.

Gleicher *et al.* [GHW02] also proposed a framework to apply cinematic principles to a lecture room scenario. Their framework was based on various available computer vision techniques (referred to as 'building blocks' in their work). Using these techniques, they suggested tracking chalkboard regions and gestures to capture video automatically.

In short, the introduction of cinematography not only helped enhance the aesthetics of video capture, but also solved some problems of purely speaker based systems. However, most of the systems only framed shots on the basis of the current speaker. Next we discuss approaches to include other cues, such as gestures, postures, and gaze.

Gaze and head-orientation based

Not all interesting and relevant events are related to the speaker. For example, the facial expressions of the listeners also convey rich information about the communication. Most of the speaker based systems cannot capture such information since they cannot exactly determine who the listener is.

Recently, Takemae *et al.* [TOM04, TOM03] proposed that participants' gaze direction could be used to detect speaker-listener pair. In their study they observed that if the majority of participants are looking at a particular participant then that participant is highly likely to be the speaker or the listener. They used this information to create a video which showed the close-up shot of the person most of the

participants were looking at. It was observed that the video created in this manner conveyed the information about speaker and listener significantly better than other videos shot using one of the following three methods: whole view of all the participants, separate closeup view of each participant shown side-by-side, and view of the current speaker only.

Jenkin *et al.* [JMFV05] also used gaze direction to determine the focal person in a video conferencing system. Their eyeView system showed the video of each participant in a separate window, but the participant being looked at by the largest number of other participants received the largest size window. Furthermore, a user could also send a request for a side conversation with another user by looking at that person's window and pressing a button.

Vertegaal *et al.* [VWSC03] have also used gaze direction to design a video conferencing system (Gaze-2) that conveys participants gaze information to the remote site. The Gaze-2 system uses multiple cameras to capture a single participant's view, but only sends the image captured by the camera with the least parallax with the gaze. This allows the system to transmit the eye contact in a parallax-free manner.

Thus, gaze is a good indicator of focus of attention in meetings, but accurate gaze detection is difficult to achieve in most practical meeting settings. Head-orientation could be a possible substitute for gaze direction because it is relatively easier to detect and has been shown to be a good indicator of focus of attention in meetings [SZ02]. Therefore, Takemae *et al.* [TOY05] extended their work to detect speaker-listener using head orientation. They designed a vision based system to determine head orientation and conducted an experiment similar to the one described above. It was observed that video editing based on head-orientation conveyed speaker-listener information as efficiently as the gaze based edited video.

In the Reactive Room system, Cooperstock and others [Coo95, CTB+95, Bux97] used sensor based inputs from different activities of the user to trigger various

Figure 2.9: Reactive Room head-tracking based camera control. Remote user is shown, in picture-in-picture format, in three configurations: moving to the left, to the right, and closer to the screen. Images taken from [Coo95].

system actions. One of the various features of the system was head-tracking based camera control. By applying head tracking algorithm to video signals, the system can determine the position of the remote participant's face in relation to his or her monitor. This position is then used to drive the camera. Thus, if the remote participant wants to see the left/right side of the room, he or she moves the head to the left/right of the screen and the local camera will move accordingly to show the corresponding part of the room (see Figure 2.9). Similarly, when the participant moves the head closer to the screen then the local camera zooms to provide a sensation of moving toward objects in the conference room. Despite the fact that controlling camera by head movements is awkward, the system showed the potentials of gestures/postures in camera control.

Gaver *et al.* [GSO95] used a similar approach to control the remote camera in their Virtual Window system. In this system, when users move their heads, the remote camera moves about a focal point always facing the focal point. This kind of camera motion provides some depth cues because the objects closer to the camera appear to move in the direction opposite to the head movement, whereas the objects further away move in the same direction as the head movement. The authors observed their system in use and found that when the camera movement was smooth and fast, the system helped the users explore the remote site as well as maintain awareness. However, when the movements were not responsive and

smooth (due to various technical reasons) the system was very confusing.

In brief, gaze could be used as a potential cue for determining the focus of attention for camera control purposes, but, in practice, accurate gaze tracking is difficult. A number of systems leveraged head-orientation as an approximation for gaze direction and used this for camera control. However, while a camera control system implicitly controlled by head orientation often produces shaky and confusing videos, when head-orientation is used as an explicit gesture for camera control users often have to make some awkward motions with their head. Some possible approaches to address these problems include a combination of light-weight hand gestures with head-orientation and voice commands.

Hand gesture based

Sherrah *et al.* developed a vision based system, called VIGOUR, to track and recognize activities of multiple people [SG00]. The VIGOUR system can find multiple people, track their head and hand positions, and recognize some simple gestures such as pointing and waving. Further, this system, in combination with an algorithm to interpret group activities, was used to control a PTZ camera for video-conferencing [SGHB00, HB02].

Various activities and behaviors could be divided into two categories: *implicit* or unconsciously made movements that accompany regular communication (*e.g.*, hand movements during conversation), and *explicit* or consciously made movements to highlight regions of interest in the environment (*e.g.*, pointing, waving *etc.*). The algorithm to interpret group activities was a supervised learning algorithm which created a behavior model from group behavior training data. The training data consisted of a feature vector of various gestures of the people in a group and the corresponding camera position.

During runtime, the VIGOUR system detected *explicit* gestures and the algorithm decided on the camera movement based on the model. Despite the fact that

no evaluation of the system was reported, the application of vision based interpretation of group behavior in camera control indicates the potentials of gesture based camera control.

Chen [Che02a] used hand gestures to control frame rate in a multi-party video-conferencing system. The system detected the vertical motion of a user's hand. Whenever the user raised or dropped his/her hand, the frame rate was increased, but at all other times frames were transmitted at a very low rate. A user evaluation showed that this gesture sensitive frame rate control approach was as effective as a continuous high frame rate approach.

Thus, hand gestures have been shown to be of potentially high utility for camera control and camera-view control in video-conferencing. However, technological limitations involved in tracking have resulted in its limited use. As tracking is becoming increasingly robust due to vision and sensing technology [Bux03], the utility of gestures (both implicit and explicit) in camera control should be explored to greater extents.

2.3.3 Summary

In brief, a single static camera can capture some amount of visual information, but it is not enough to create a sense of presence for remote users. Multiple cameras, particularly PTZ cameras, could convey more information but the human effort involved in controlling them is expensive. Therefore, at least partial automation is required if we are to use multiple cameras.

Previous systems have mostly used speaker based automation to make sure that remote users are always aware of the speaker. Some of the camera control systems allowed users to control the camera by gestures or postures. However, all these strategies address the question of what should be captured by cameras. The question of when and how it should be captured requires separate effort.

Some of the more recent systems explored cinematography to find out how and when an event should be captured. Given the popularity of live talk shows, this approach could well be assumed to have the potential to make meeting videos more engaging and informative. However, apart from applying a few elementary rules from cinematography to shoot a speaker in a lecture room scenario or a meeting room scenario, this approach has been largely unexplored.

2.4 Understanding meetings with critical visual information

Meetings often involve various physical artifacts. Capturing the visual information associated with these artifacts form an integral part of the overall meeting capture system. In this section we discuss a more general problem of collaboration on physical tasks. Dealing with meeting artifacts is an instance of this general problem.

Collaboration on physical tasks, such as operation on a patient, repairing of a machine or construction of a building model, differs from a collaborative meeting in that it is much more focused towards a 3D physical task. A typical scenario in such collaborative tasks involves a remote *helper* or *expert* helping or instructing a *worker* who actually performs the physical task.

In such tasks collaborators strive to attain *common ground* through conversation and share their knowledge to complete the task at hand [GKE90]. Here, common ground in communication could be formally defined as a set of mutual knowledge, mutual beliefs, and mutual information [CB91]. The process of establishing common ground, also termed as grounding, changes with the communication channel and there is growing evidence that visual channel supports this process [CB91, VOOF99].

The two most important aspects of the task, communication properties and shared visual space, are influenced by camera control providing the visual channel. In this section we summarize some of the previous studies to better understand this influence.

2.4.1 Communication properties

One of the earliest studies to analyze the effect of visual channel on the properties of verbal communication in video-mediated collaboration on 3D tasks was conducted using a system that allowed the helper to draw gestural instructions on top of the video of the worker's space (similar to the Videodraw system [TM90]). Further, the worker had a head mounted display (HMD) and a small head mounted camera. The head mounted camera captured the workspace of the worker, and the HMD showed the gestural instructions provided by the remote helper (similar to the Sharedview system [Kuz92]).

In the study, verbal communication was analyzed to compare the communication properties in face-to-face and video-mediated conditions, with and without gestures. It was observed that when helpers were not allowed to use gestures in the instructions, they used significantly higher number of verbal expressions, irrespective of the communication being face-to-face or remote. However, when the gestures were allowed, the number of verbal expressions was higher in remote collaboration as compared to face-to-face collaboration.

Kraut *et al.* [KMS96, KFS03] analyzed the effect of video mediation on communication properties using a bicycle repair task. The study compared two camera control conditions. In the first condition, the worker used an HMD to view the instructions and had a tiny head mounted camera to capture video of the work area. This video was viewed by a remote expert who was communicating with the worker over a full duplex audio link. In the second condition, the HMD

camera-view was removed from the helper's site.

It was observed that pairs' performance in collaboration did not vary with communication technology. However, some significant differences in the communication pattern were found. Workers were less explicit in describing their state of work and workspace when they shared a view of the environment, which is in accordance with the findings of Kuzuoka [Kuz92]. Video channel also allowed helpers to give proactive suggestions. This study provides some more evidence in favor of using video for remote collaboration on 3D physical tasks.

As McCarthy *et al.* [MM94] suggested, efficacy of communication can be measured along several dimensions other than task completion time and quality of the outcome. Analyzing communication properties could also be one of the dimensions. Thus, analysis of communication properties in prior work suggests that face-to-face collaboration differs significantly from video-mediated communication along this dimension. What remains to be seen is how the understanding of these differences could be utilized to improve systems in the future.

2.4.2 Shared visual space

A shared visual space allows multiple people to see similar views of objects and environments [GKF04, KGF02]. For example, since images from a camera can be replicated across distances, they could potentially provide a shared visual space for people collaborating remotely. Gergle *et al.* [GKF04] observed that when a shared view of the workspace was available the workers were more likely to use their actions as language (*e.g.*, nodding instead of verbal approval or pointing to an object instead of describing it), which made the communication more efficient. Thus, to some extent, success of video-mediated collaboration on physical tasks relies on how well a system can create a shared visual space.

In face-to-face collaboration, shared visual space consists of the entire workspace,

but in most video-mediated collaborations it is constrained by the view of a static camera. Some prior studies used head-mounted cameras to create a better shared visual space [Kuz92, GKF04], but in a recent study, Fussell *et al.* [FSK03] showed that a head-mounted camera did not improve the performance time in some robot building tasks.

The study compared five media conditions: face-to-face, audio-only, head-mounted camera, scene-camera, and scene plus head-mounted camera. As expected, performance time under the face-to-face condition was significantly faster than any other media condition. In addition, performance with the scene camera was significantly faster than audio-only condition. However, neither the head-mounted camera alone nor the combination of both cameras was significantly better than audio-only. Furthermore, the head-mounted camera was found to perform worse as far as the amount of worker talk was concerned.

The study described above cautions against the use of multiple video feeds to create shared visual space. Similar observations have been made independently about creating shared visual space using multiple fixed camera views [GSHL93, HLS97]. In a study of the Multiple Target Video system (MTV) [GSHL93, HL92], participants were asked to remotely collaborate on a room planning task.

It was observed that even multiple fixed views failed to offer complete access to the remote space. Various issues related to orientation, pointing, and partner monitoring were revealed in this study, but the problem of finding what the remote collaborator is pointing to or focusing on still remained unsolved.

These studies strongly suggest that creation of a shared visual space for efficient collaboration requires much more than just providing a wider coverage of the environment.

Recently Fussell *et al.* [FSP03] and Ou *et al.* [OOF+05, OOYF05] approached this problem from a different perspective by studying the eye gaze pattern to understand focus of attention in collaborative tasks. Fussell *et al.* found that in a

robot construction task, the pieces and worker's hands were glanced at significantly more often than all other targets.

Ou *et al.* [OOYF05] used a similar gaze tracking approach where the task was to collaborate on an online jigsaw puzzle. The shared visual space for the collaborators had two distinct regions: pieces bay and workspace. The study showed that when puzzle pieces were harder to discriminate, the helpers spent more time looking at them in the pieces area. Furthermore, as collaborators repeated the trials, helpers spent less time looking at the pieces when they were easy to discriminate. However, for harder to discriminate pieces trials had no effect. These results showed that visually complex objects required higher attention and establishing a common ground was difficult in such cases.

In short, previous studies have shown that communication properties change as the medium changes and pointed out several objective metrics to compare media, such as amount of verbal communication and types of words used. These metrics could also be used to compare and contrast the efficacy of the systems supporting collaborative physical tasks.

Furthermore, some evidence has been gathered to show that appropriately created shared visual space could successfully support collaborative 3D physical tasks. This motivates the study of camera control strategies for creating shared visual space.

2.5 Camera control for collaboration on physical tasks

In this section we study various existing camera control systems that support collaboration on physical tasks. Similar to Section 2.3, we divide a generic camera control system into three components. We then identify these components in various previous systems and analyze them.

2.5.1 Camera setup

The number and type of camera forms an important part of any video based system to support collaborative physical task. On the one hand, as the number of camera views is increased, the visual access to the remote space is enhanced. On the other hand, this increase also raises the attention divide of the helper. Here we discuss the two most widely used camera setups.

Single camera, single view

The most common form of camera setup is a static scene-camera that sends a single wide view of the workspace to the remote helper [OOYF05]. This camera configuration often suffers from two problems: lack of detail and single constrained view of the world. In order to address these issues various systems have deployed a single movable camera. The camera could be static and mounted on the head of the worker [Kuz92, FSK03] or on a robotic arm/actuator [WHK92, PC98, Jou02], or could be a pan-tilt-zoom camera controlled remotely [RBB06]. Such dynamic cameras give a better coverage of the scene including both wider and closer views. However, depending on their setup they have their own disadvantages.

A head mounted camera constrains the camera view to the direction in which the worker is looking [KKT94]. Furthermore, it has been shown that they are not any better than a fixed scene-cameras for collaborative physical tasks [OOYF05]. A pan-tilt-zoom camera or a camera mounted on a robotic dolly does not have this problem, but they need to be controlled explicitly. The burden to control the camera interferes with the primary actions of the users [WHK92, RBB06].

Multiple cameras, single view or multiple views

Gaver *et al.* [GSHL93] designed and studied a system called Multiple Target Video (MTV) in which multiple cameras pointed to different locations in the workspace and the remote user can select any of those camera views. While this system provided access to most of the areas of interest, selecting the correct view was still a cumbersome task (see Figure 2.10). In a modified version of MTV, called MTV II, multiple views were made available simultaneously. However, separate fixed cameras still failed to offer complete access to the remote space [HLS97].

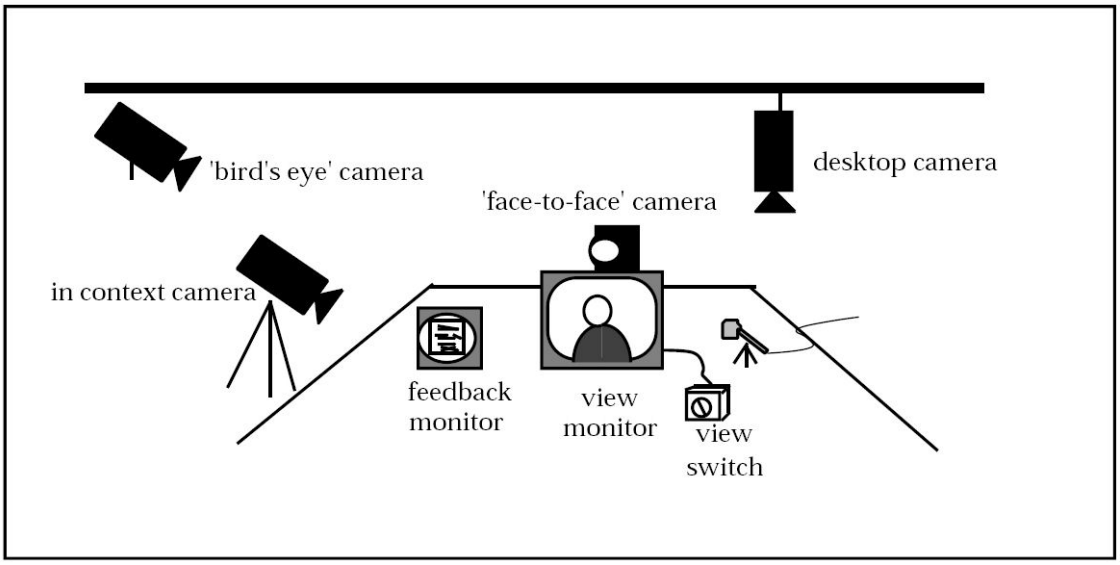


Figure 2.10: A diagrammatic layout of the Multi Target View (MTV) System. Image taken from [GSHL93].

Thus, various setups of the cameras have been shown to have their own advantages and disadvantages. From the discussion on shared visual space (see section 2.4.2) we identify one property of an ideal camera setup to be the ability to show the right amount of information at the right time. However, in most of the setups we discussed, there is a constant trade off between appropriate scene-coverage and burden of control. Therefore, finding an optimal point in this trade-off could be of tremendous importance towards achieving the goal of "capturing

the right amount of information". This leads to the problem of managing camera view control.

2.5.2 Camera view control

In remote collaboration, the helper needs to be aware of the remote environment all the time. There are various ways in which existing systems use camera view control to maintain this awareness. Here we study them in two categories.

Manual

In most of the systems based on a single or multiple static cameras, the users can select a camera and view the region of interest captured by that camera. When cameras are of pan-tilt-zoom type or mounted on an actuator then they are controlled remotely using a software or hardware interface. The GestureCam [KKT94] system employs both these types of interfaces to control a camera on an actuator. Software based control allows the user to select a portion on a touch-sensitive screen and the camera moves to center on that portion. Hardware based control provides the user with an exact replica of the actuator (or the *master*), and the user can directly use the *master* to control the camera (or *slave*) [KKT94].

Some systems allow the remote helper to control the motion of the camera mounted on a movable robotic dolly. In Gestureman [KOY+00] and PRoP (Personal Roving Presence) [PC98], the remote user controls the movement of a tele-robot with a camera mounted on top of it. By moving the robot on wheels the remote user can detect interesting events, view them, and communicate using a two-way video and full-duplex audio channel. Jouppi [Jou02] developed a similar system by using sophisticated robotics technology to move the robot with

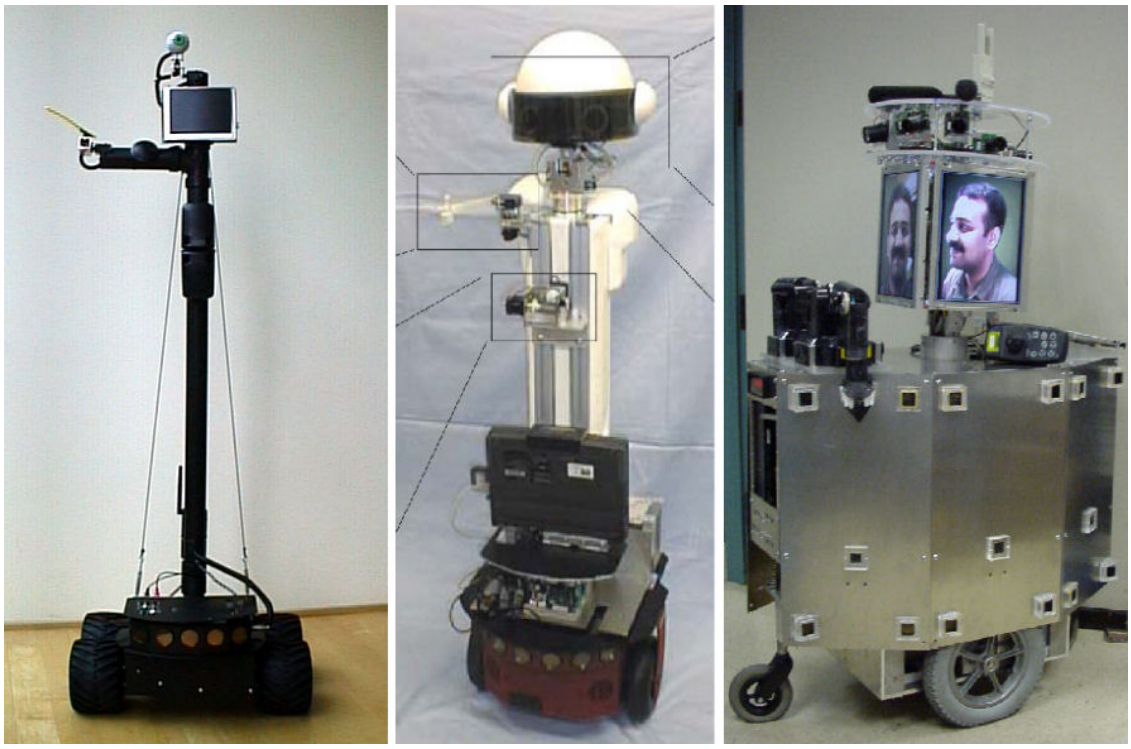


Figure 2.11: Cameras mounted on robots. Left: PRoP system, Center: Gestureman system, Right: Mobile Telepresence system. Images taken from [PC98, KOY+00, Jou02].

several cameras mounted on it. These cameras captured a wide view of the scene at varying resolution. The captured video streams were used to recreate a view with resolution falling off from the center to the edges. Figure 2.11 shows these three systems. While these systems provide much higher control over what the remote person can see, the burden of controlling the robot could overwhelm the remote user.

In the Gestureman-3 system, Kuzuoka *et al.* [KKY+04] used multiple monitors and head tracking to simplify the remote control of the robot. The videos corresponding to the three cameras mounted on the robot's torso are shown to the remote user on three different monitors. The remote user can move his/her head to look at any of these three monitors and the robot's head would also move based on the remote user's head orientation. This interface not only made the robot

control easier, but also reflected the controller's actions. However, the presence of a separate robot itself poses the problem of robot maneuverability, especially when working in small spaces.

Kurata *et al.* [KSK+04] addressed these problems by developing a wearable system, called WACL, with a PTZ camera and a movable laser pointer (see Figure 2.12). The main advantage of this system over an HMC is that the camera can be pointed to any direction, irrespective of the orientation of the user on which it is mounted. The system also supports pointing by laser pointer and is more portable than the Gesturecam systems. In a user study, this system was compared with an HMC system, and the results showed that the users felt more comfortable using the WACL system than using the HMC system. However, no significant difference was found in task completion times.

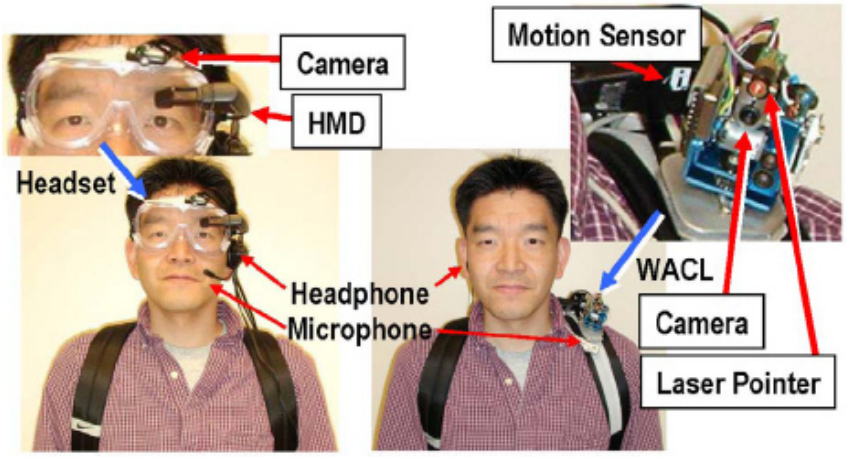


Figure 2.12: Shoulder-worn Active Camera Laser System. Image taken from [KSK+04].

Thus, previous systems implemented manual camera control either using the cameras mounted on top of movable robots or mounted on the worker itself. While such a control allows the user to explore the remote environment, it could potentially increase the cognitive load on the helper. Therefore, automatic control of the camera view would be advantageous.

Head-orientation and gaze based

Various systems used head mounted cameras with a belief that the helper would be mostly interested in what the worker is looking at, and, therefore, it would automate the view control [FSK03, Kuz92]. Recently, this belief was challenged [FSK03] and gaze direction has been proposed as a potential cue to control the camera [OOYF05, FSP03]. Ou et al. [OOF+05] suggested that the helper's gaze can indicate what the helper wants to look at. Therefore, by predicting his/her gaze, the camera view could be controlled automatically. They proposed a conditional Markov model classifier to predict helper gaze for a simple online puzzle task. The model takes task properties (puzzle pieces type), people's actions (mouse movements), and message properties (speech transcription) as input and predicts the gaze of the helper. Despite a low online prediction accuracy (65.4% for solid color puzzle pieces and 74.2% for shaded-color pieces) for a simple on-screen task, the system suggests the possibility of developing an automated camera control system for collaboration on physical tasks (see Figure 2.13).

The gaze based approach assumes that the helper can see the entire workspace. However, for many real world tasks this may not be the case. If a worker cannot see the entire workspace then it is not clear how his/her gaze would be predicted for the areas beyond the view. In such circumstances, the worker's action could be another potential cue to predict the helper's focus of attention.

Worker action based

Wakita *et al.* [WHK92] created an automatic camera control system for monitoring telerobotic tasks. Except for the fact that the worker here is a robot, this setup is very similar to our familiar helper-worker setup. The camera control system used a robot's motion properties to automatically control the zoom level of the

Figure 2.13: Setup for gaze based action prediction and camera control. Image taken from [OOF+05].

camera. They divided the robot's typical pick-and-place task into four motions: Approach motion to represent robot's arm approaching the object, M-t-g motion to represent robot's arm moving close to the object in order to grasp it, Grip motion to represent robot's fingers closing to grasp object, and Lift-up motion to represent robot's hand's upward motion with the object in the hand.

The robot operator (or the helper) had different viewing requirements for each of these motions. For example, during the Approach motion, the operator wanted to see a wide shot of the hand to guide it to the object, whereas during the M-t-g motion, the operator wanted to see a tight shot of the robot's hand to slowly move it closer to the object in grasping position.

Further, based on the requirements, separate camera zoom levels were assigned to the views corresponding to each of these motions. During actual operation, the system automatically decided the zoom level based on the motion type. Thus, this system introduces a worker action based camera control paradigm to

automate camera control when the task involved is very structured. However, the application of this approach to more complex problems may face various research challenges.

In brief, previous studies have explored the communication properties in order to better understand the impact of camera control on collaborative physical 3D tasks. These studies suggested a number of objective measures to compare systems supporting such tasks. Using these measures, the efficacy of shared visual space created by various camera control strategies have been compared.

2.6 Concluding remarks

From the discussion in this chapter, it can be concluded that various types of meetings share some common aspects from the perspective of designing a camera control. They all have some visual focus of attention that must be determined and captured effectively.

In meetings with simple scene and critical visual component, determining the visual focus of attention is relatively simpler, but capturing it could be challenging. In the case of meetings with complex scene, the determination of visual focus of attention itself grows complicated. For example, in a meeting with dynamic discussion involving multiple participants and some use of physical artifacts, the visual focus of attention is not apparent.

Previous systems employed a variety of strategies to address these problems. Out of the various camera control strategies, the static camera control was shown to perform surprisingly well. One possible reason for its relative success is that it captures all the visual information, focal or not, and the viewers of this information pay attention to only the portions which are relevant. By doing so, however, they miss the details of the relevant information.

In contrast to a static camera, cameras with dynamic views could potentially

provide focal information to appropriate details, but their views need to be controlled either manually or automatically. On the one hand, when controlled manually, they put the burden of camera control on the meeting participants. On the other hand, when controlled automatically they are usually ineffective.

When expert humans (non-participants) control cameras to capture collaborative meetings, they perform arguably well. Television talk shows, cooking shows, sports capture could be presented as evidence. This implies that lessons can be learned from these experts to improve the effectiveness of automatic camera control.

In the next chapter, we perform detailed analysis of how television production works and what the key factors are which make it effective in capturing collaborative activities. While discussing the principles of television production, we also note how they can be leveraged in automatically capturing meetings.

Chapter 3

Principles of Television Production

3.1 Introduction

Television production is the process of capturing a staged or spontaneous event as a television program. The process usually involves a trained crew and a studio setup with several types of equipment.

TV production varies significantly depending on various factors including the type of event, location of capture, and budgetary restrictions. However, regardless of these factors, it can be characterized by the following three stages [DS00]: (1) Preproduction, (2) Production, and (3) Postproduction.

In the first phase (preproduction phase) the program is conceptualized (writing script, hiring people, creating sets, *etc*.) and all the necessary elements are organized. In the second phase (production phase), the event actually takes place and the production crew captures the event. In the third phase (post-production phase), the best portions of the captured video are selected and combined to create a coherent and compelling show.

In this chapter, firstly, we provide an overview of some of the major systems and activities involved in these three stages. Secondly, we explore some of these

elements in details focusing on their role in capturing meeting videos. The intent is to help guide the design of camera control systems based on television production. Finally, we identify the principles that we will apply in our camera control designs.

3.2 Overview of studio production system

We interviewed a camera crew to understand details of the studio production system. Next, we invited the crew in our laboratory to set up a studio using the various equipment they suggested during the interview. Finally, we consulted the television production literature to check for any inconsistencies between our observations and the established process as suggested in the literature. Here we present a summary of our observations and literature survey.

3.2.1 System elements

A studio system includes a video system with at least the following components: (1) one or more cameras, (2) a camera control unit or units, (3) preview monitors, (4) a switcher, (5) a line monitor, (6) one or more digital or tape video recorders, and (7) a line-out that transports the video signals to the recorder or the transmitter [Zet05].

There is also an audio system consisting of (1) one or more microphones, (2) an audio console, (3) an audio monitor (speaker), (4) a line-out that transports the sound signal to the recorder or the transmission device.

3.2.2 Production elements

Camera

The camera is the most important production element. There are various components of this element (*e.g.*, lens, CCD, viewfinder, mounting equipment). Single or multiple cameras could be used in the production process.

Lighting

Adequate lighting is required to capture good quality video. In production, lighting has the following purposes [Zet05]: (1) to illuminate the scene so that cameras can capture technically acceptable videos; (2) to provide other information about the environment including the objects in the scene, their relative positions using shadows and reflections, and time of the capture; and (3) to establish the general mood of the event (*e.g.*, bright and cheerful, or dark and gloomy).

Audio

Audio is one of the important production elements. In particular, for the purposes of capturing collaborative meetings, the information contained in audio is of critical value. Depending on the scene settings, different types of microphones are used in TV production.

Video recording

Video recording is used not only in postproduced shows, but also in live shows for commercial breaks, replays, and archiving. They could be recorded either on tape or digitally on a hard drive.

The switcher

The switcher allows the selection of one video source out of several possible sources. This makes it possible to do instantaneous editing for live shows.

Postproduction editing

In postproduction editing, the editor looks at various recorded video streams, picks up the most relevant scenes from these streams, and places them in an order such that they form a coherent story. This step is often not present in its entirety for live shows. *In this dissertation, since we only consider live capture, we skip the postproduction step.*

Special effects

Special effects are often used for placing titles on the video, transitioning from one shot to another or displaying multiple video streams at the same time. They can be used during any of the three stages: preproduction, production, or post-production.

3.3 Camera

TV production is a lot about visuals, and since cameras capture visual information, they form the most important component. The number and type of cameras not only influence the production stage, but also the pre- and post-production stages. For example, in the pre-production stage, number of camerapersons required, placement of cameras, and lighting are determined by the type and number of cameras. Similarly, in the post-production stage, editing techniques depend on this as well.

3.3.1 Camera movements

The type of cameras is characterized by various factors including the focal length of the lens (zoom), viewing angle, bulk and weight *etc*. Regardless of the type, they often perform one of the following types of movements [Zet05]. It should be noted that left and right are defined with respect to the camera's point of view.

- *Pan*. Turning the camera horizontally from right to left or left to right around a fixed vertical axis passing through the camera.
- *Tilt*. Turning the camera vertically up or down around a fixed horizontal axis passing through the camera.
- *Pedestal*. Elevating or lowering the camera on a studio pedestal.
- *Tongue*. Moving the whole camera from left to right or right to left by moving the boom of the crane while still keeping the camera face the same general direction.
- *Crane*. Moving the whole camera up or down by moving the boom of the crane while still keeping the camera face the same general direction. This movement allows camera to move by a distance larger than that allowed by the pedestal movement. Furthermore, the camera also moves in a slight vertical arc (since the boom moves in a circle).
- *Dolly*. Moving the whole camera toward or away from an object approximately along a straight line by moving the camera mount.
- *Truck*. Moving the whole camera from left to right or right to left by moving the camera mount.
- *Arc*. Moving the camera in a slightly curved dolly or truck movement.

- *Cant*. Tilting the camera sideways so that a horizontal line is captured as a slanted line by the camera.

- *Zoom*. Changing the focal length of the camera while the camera remains stationary. By the camera "zooms in" by changing the lens to narrow angle, and "zooms out" by changing the lens to wide angle.

Most camera control systems for capturing meeting video are constrained to use only pan, tilt, and zoom movements. A few systems use complex robotic cameras [Gestureman] that can support greater range of movements. However, since they require a trained operator to control its movements, they are not cost effective for meeting capture.

In this dissertation, we focus on using pan-tilt-zoom cameras. This constrains the camera motion and, hence, the types of shots that the camera could frame.

3.3.2 Shot framing

Shot framing is the art of capturing images in aesthetically pleasing manner with enough visual details so that they convey appropriate meaning. Although camerapersons could use their creativity in framing shots, there are several guiding principles to make a shot technically correct.

Headroom, noseroom and leadroom

When framing a shot, it is advised to leave some space above people's head. This space is called *headroom*. Its purpose is to makes sure that people's heads are visible even if the edges of the frame are lost due to transmission and tape recording errors.

When framing shots of (a) a person facing or pointing, or (b) a person or an object moving in a direction other than straight into the camera view, there should be some space left in the frame in front of the person or the object. This space is

called *noseroom* in case (a) and *leadroom* in case (b). Absence of this space makes the frame aesthetically disconcerting making the viewer feel that the person or the object will fall off the frame edge.

Closure

The principle of closure determines how much information should be displayed in the frame and where should the frame boundaries be placed. The purpose of closure is to help viewers easily construct the context and extend the shape of the object even if it is not entirely captured in the frame.

Shot types

There are various types of shots used on TV. Some of the most commonly used shots are: Extreme long shot, Long shot, Medium shot, Close-up, Extreme close-up, Bust shot, Knee shot, Two-shot, Three-shot, Over-the-shoulder shot (see Figure 3.1 [1]).

Screen motion

In TV production, shots often have moving objects. When the object in the frame is moving along the line of view of the camera, the object should always be kept in focus.

Framing objects moving perpendicular to the line of view of the camera involves more technical issues. The object should be given proper leadroom when the camera is following the motion.

When there are multiple objects or persons moving away from the shot (*e.g.*, a person moving away from a two-shot), only one person should be followed; the camera should not try to keep everyone in the frame.

[1] These shots have been taken from Sidney Lumet's drama *12 Angry Men* [Lum02], 1957

Figure 3.1: Some shot examples used in television production. *First row*: Extreme long shot (left), Long shot (right); *Second row*: Medium shot (left), Over-the-shoulder shot (right); *Third row*: Close-up shot (left), Extreme close-up shot (right), *Fourth row*: Two-person shot (left), Three-person shot (right). Images taken from [Lum02].

In over-the-shoulder shot, if the person closer to the camera frequently blocks the camera view then the camera should be moved using truck or arc movement to frame the correct over-the-shoulder shot.

Shot stability

In close-up shots, every minor movement should not be followed as it could make the viewer seasick. The camera should either be pointed to the major area of action or pulled out to frame a wider shot. In general, when capturing motion on screen, the camera movements should be minimized and motions should be smooth [Zet05].

Shot stability is of particular interest from designing camera control perspective since various previous systems for capturing collaboration on physical tasks suffered from quickly changing views (*e.g.*, camera view from a head mounted camera [FSK03]). This principle strongly warns against using close-up shots following every movement of the objects of interest.

3.3.3 Camera placement

Directors often perform *blocking* before the production [Zet05]. In the blocking process, movement ranges of people and objects are estimated; cameras are placed in such a manner that all the movements are captured with correct shots. During the production stage, camerapersons stick to the blocking and frame shots of the regions they are responsible to cover.

A correct camera placement makes sure that primary characters on the stage are covered appropriately. For example, in a typical 3 person interview shooting (1 host and 2 guests), there should be one camera placed in front of the host to frame a close-up, and one camera should be placed in front of the two guests to frame close-up and two-person shots of the guests.

Since facial expressions of the participants are of utmost importance in television shows, the ability to frame close-ups facing the camera is often deemed high importance while setting up the cameras. Another important constraint in camera placement is that one camera should never see the other camera in the

frame. In live television, the second constraint is often satisfied by preserving the 180° rule.

3.4 Editing

Most television programs are not live [DS00]. They involve some sort of *post-production editing* stage. In this stage, the editor analyzes various recorded videos of the event (possibly shot at different times and locations) and organizes them to create a compelling video. The final video is produced some time after the actual event has finished.

In live television, on the other hand, video is shot and produced without delay. The director decides what should be shown at any point of time while the event is still in progress and orchestrates the switching from one video feed to another. This is also called switching or instantaneous editing [Zet05].

Instantaneous editing is one of the most challenging aspects of live TV. As the director Bryan Russo commented about his experiences in directing the Phil Donahue Show [Ros99]:

"Straight coverage is easy–you just show whoever's talking, and I'm sure if you go back and look at my early shows, that's what I did. But now that's not what's it about for me. What it's about is taking the viewer and putting the viewer in the studio audience. They can't see what's going on in the corner of the audience or the guy shaking his head saying, 'I can't believe what's going on.'"

In this section, we discuss various aspects of instantaneous editing since meetings often require live coverage for real-time collaboration. Furthermore, a live video could also be used for offline activities (*e.g.*, archiving).

3.4.1 Basic transitions

A TV program is essentially a sequence of shots transitioning from one shot to another. There are various ways in which this transition can take place.

The cut

The cut is the simplest and commonest of all transitions. It is an instantaneous change from one shot to another. The cut itself is never visible; only the shots before and after are the cut are visible.

There are various guidelines as to when a cut should be used for transitioning. In general, a cut is performed between two shots with similar time and location [DS00].

The two main purposes of using the cut or any other transition are: clarification and intensification [Zet05]. Cuts performed to clarify the details fall under clarification and to intensify emotion or impact fall under intensification. For example, cutting to the close-up of the speaker to show who is talking is clarification. Whereas, as the speech grows emotional, subsequent cuts to extreme close-up is intensification.

The dissolve

The dissolve is a gradual transition from one shot to another with an intermediate stage when both shots are partly visible. It could be fast or slow in terms of how long it takes to transition from the end of one shot to the beginning of another shot. A very fast dissolve looks similar to a cut (also known as "soft cut").

A dissolve usually indicates a lapse of time from one shot to another [DS00]. Sometimes a dissolve is also used to indicate strong relation between two objects present in two different shots. They may or may not be present at the same time. In meeting capture, a dissolve can be used to show some previous clips from the

meeting. It could also be used to transition to shots of physical artifacts.

The fade

In a fade the shot gradually transitions to a black frame (fade-out) or a black frame gradually transitions to a shot (fade-in). A shot can fade-out to a black frame and another shot can fade-in from the black frame immediately after the fade-out; this is called cross-fade. A fade can be used to indicate a longer time lapse than dissolve, a change in place, or a change in topic.

3.4.2 Editing principles

The goal of editing in TV production is to convey a coherent story to the viewer without making the transitions obtrusive. There are various guidelines to make transitions look natural. These guidelines suggest why a transition should be made, what type of transition should be made, and when a transition should be made. In this subsection, we discuss some basic principles of editing which are followed by most directors regardless of their subjective preferences.

Preserving the 180° rule

One of the most important principles of editing suggests that objects and people in a scene should maintain their relative orientation after transitions. For example, if person A is on the left of person B on the screen in the first shot, then they should have the same orientation after any transition.

This principle is implemented by preserving the 180° line (also known as the Action Axis) [DS00, Ari76]. This is a hypothetical line which usually passes through the main subjects involved in the action (*e.g.*, through the two people involved in a conversation, through a moving person or object in the direction of the movement).

Once the line is established, the cuts should be made only between shots from the cameras on the same side of the line. If a camera has to cross the line, then a shot from the *neutral direction* must be inserted before crossing the line. A neutral direction is right along the 180° line.

Avoiding the jump cut

A jump cut is a cut from a shot of a subject (a person or object) to a similar shot of the same subject. It is usually seen when the middle section of a continuous shot is removed. This type of cut causes the subject to suddenly reposition in the frame, and there appears to be a jump in the scene after the cut.

Jump cuts are usually avoided in TV production. One common approach to avoid this is to insert a different shot between two similar shots (also known as Cutaways). In cinema [God60, Ver02], jump cuts have been used to emphasize particular emotions or themes. This use was introduced by Jean-Luc Godard in *À bout de souffle* [God60].

Including cutaways and cut-ins

A cutaway is cutting to a shot of an object that is not visible in the previous shot. A cutaway introduced between two similar shots of the same subject not only avoids a jump cut, but also makes the viewer aware of the context.

A cut-in is cutting to a shot that shows details of a portion of the previous shot. Cutting from a long shot to a medium shot and from a medium shot to a close-up are examples of cut-ins. Cut-ins usually intensify the emotion and help retain the viewer's attention.

Timing the cut correctly

In order to make transitions unobtrusive it must be timed properly. The timing is specially important when there is some motion involved in the shot. For example, if the current shot is a close-up of a subject sitting on a chair, and she starts to stand up, the next shot should be a medium shot. The timing of this transition can make it either natural or obtrusive.

If the transition is made in anticipation before the subject starts to stand up, it makes the transition seem unnecessary. If the transition is made after the subject already started to stand up, it makes the viewer worried that they will miss the information. The right transition is to cut as soon as the subject starts to stand up.

Cutting for a reason

Cutting for a reason is one of the fundamental rules of editing [DS00]. A cut without a strong reason makes the transition distracting. Most cuts are used to show details, to show context, to show reaction, to maintain 180° rule, or to capture motion. Experienced directors also advise against spurious use of various types of transitions available on current digital switchers [DS00].

3.5 Visual effects

All types of operations that change the appearance or the layout of the frames to intensify or clarify the message are called visual effects. In this section we discuss some of the most commonly used visual effects.

3.5.1 Wipe

In a wipe the current on-air image is gradually replaced (partially or completely) by another image. The new image can replace the old one in various geometric

patterns; the most common patterns are horizontal and vertical. The boundary between the two images could be solid or soft (blurred); the soft boundary wipe is called a *soft wipe*. There are various visual effects that can be created by using the extent and pattern of the wipe.

Split screen

When a horizontal, vertical or diagonal wipe is stopped halfway, it creates a split screen effect. In this effect, portions of the two images are visible at the same time and they occupy different areas of the screen. Their position (side-by-side or one on top of another) on the screen is important in conveying the message.

This effect is often used to show two important and related events occurring at the same time. For example, the split screen effect can be used to show a speaker talking and a listener reacting to the speech. It should be noted that the two images must be appropriately framed to convey the relationship among the people or objects across the images.

Spotlight effect

When a circular soft wipe is stopped midway and the previous frame is allowed to show through (using frame superimposition), it creates a spotlight effect (see Figure).

This effect is often used to highlight a portion of the frame. For example, in a meeting room scenario, if a speaker is referring to one of the multiple objects placed on a desk, the spotlight effect could be used to draw the viewers' attention to that particular object.

3.5.2 Multi-image

The multi-image effect is created by showing multiple full frames at the same time. There are various ways in which multiple frames can be organized on the screen. Here we discuss two popular ways.

Suqeeze-zoomed image

In this effect the first frame of a video is squeezed and placed somewhere in the on-air video. As the squeezed video continues to play, its size is increased to eventually take up the entire screen.

This approach is often used to show a newscaster and the story or a remote reporter on the same screen. The news story frame is proportionately reduced in size and placed over the newscaster's shoulder. As the newscaster talks about the story, the squeezed frame gets bigger to eventually become the main screen video.

In a meeting setting, this effect can be used to show the close up of the current focal person and squeezed image of the next possible focal person or object (*e.g.*, a whiteboard being used, a listener nodding or preparing to speak next). This could provide a temporal context of the events in the next full screen video.

Secondary frame effect

In this effect, several images are shown at the same time with possible geometric transformations. For example, in talk shows, this effect is used to show the host and the remotely located guest side-by-side and with perspective transformation. The transformation conveys the information that the two people are talking to each other.

For showing a discussion involving multiple sites, the screen can be split into multiple part and each site can be shown in one part. Geometric transformations

can also be applied to these frames to convey more information about which sites are involved in the discussion.

3.5.3 Instant replays

An instant replay is reviewing of the video recording of the current event. It is commonly used in sporting events to show some portions of the event (*e.g.*, a goal in Soccer or Hockey) from different views. Replays could be either regular or slow motion.

A regular motion replay can be used in meetings to review a particular part of the meeting or to view the meeting from a different camera angle. Slow motion replays are often used in sports coverage where events occur at a high speed.

3.6 Principles used in this dissertation

In our exploration of camera control design, we applied a specific subset of the principles discussed in this chapter. We wanted to use basic video and audio capture systems that are similar to those commonly found in most conferencing systems. Further, we avoided any complex robotic cameras in our exploration to make the final system more accessible. Based on these constrains, we selected the following principles:

- *Camera movements*: The camera movements used in our camera control design were limited to pan, tilt, and zoom.
- *Shot framing*: We used the concept of headroom and noseroom while framing the shots automatically.
- *Shot types*: The types of shots used in our camera control were: Long shot, Two-shot, and Close-up.

- *Shot stability*: We used the principles of shot stability to capture moving objects and people.

- *Editing principles*: We used only cuts for shot transitions. We also made provisions to preserve the 180° rule and avoid jump cuts.

3.7 Concluding remarks

Television production principles can provide guidance in designing automatic camera control. Depending on the type of collaborative activity, heuristics can be designed to automate various aspects of a camera control including number of cameras required, camera placement, lighting, types of shots, camera movement and shot framing, objects in the shots (visual focus of attention), shot transitions, and special effects to clarify the information.

In this dissertation, we design camera movement heuristics based on these principles. Approaches to deal with screen motion and to maintain shot stability are particularly useful in capturing meetings with critical visual information.

While capturing meetings with complex scenes, we used principles of camera placement, camera movement, shot framing and shot transitions to guide the design. Cues to determine visual focus of attention were also gleaned from television production principles.

In the following chapters we explore various aspects of camera control one step at a time. In so doing, we demonstrate how aforementioned principles can be applied to address numerous issues involved in the process.

Chapter 4

Understanding Desired Visual Information

We start our exploration of automatic camera control by analyzing a simple meeting scenario with critical visual information. In this scenario, a remote expert (the helper) and a worker participate in a meeting. The goal of the meeting is to perform a 3D physical construction task while communicating via an audio-video channel.

The helper sees a video view of the workspace where the physical task is being performed by the worker [FKS00, KFS03]. This shared visual context can then be used to facilitate the negotiation of common ground in the ongoing conversation between the helper and worker [CB91]. They are connected through two-way audio channel.

We chose this meeting scenario since it is simple enough to allow detailed exploration of visual focus of attention in collaborative tasks and complex enough to represent various real world collaborative meetings, such as remote repairing for a complex engines and remote surgery.

4.1 The study

4.1.1 Goals

In the present study, we simulate an automatic camera control by a trained and dedicated camera person operating the camera. Based on the first assumption of this dissertation (see Chapter 1), such a control can be considered the holy grail of automatic camera control. Using this setup, the goal of the study is to:

1. explore the potential usage and value of expertise in camera control by comparing performance between groups with and without a dedicated camera operator, and

2. explore the nature of participant motion and camera motion in carrying out 3-D physical 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 heuristics that will eventually drive camera control.

4.1.2 Design

In this study, we compare performance between pairs of participants performing four construction tasks of varying complexity using Lego plastic pieces. As with the onscreen puzzle task used by Gergle *et al.* [GKF04], this task involves steps common to a range of collaborative tasks: piece identification, piece movement, piece manipulation and placement, and verification of correct placement.

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 workers workspace. The output was displayed on a 13" video monitor in front of the helper (see Figure 4.1). This condition represents the most common configuration for meeting video capture.

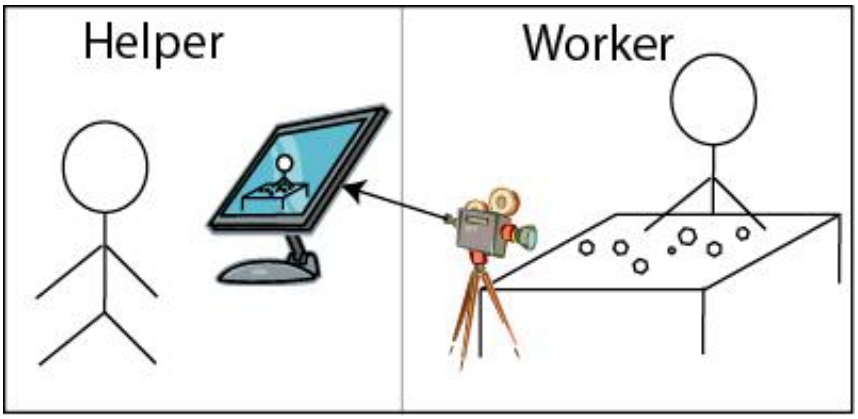


Figure 4.1: A schematic diagram of the Fixed scene camera condition.

- *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 (see Figure 4.2). This condition represents a range of meeting capture systems [PG05, LKF+02, GSHL93, KSK+04] that allow participants to control the camera.

- *Operator-Controlled Camera*: A single pan-tilt-zoom camera, located directly in front of the worker, was controlled by a dedicated operator. The camera output was displayed on a 13" video monitor in front of the helper (see Figure 4.3). This condition represents the gold standard for automatic camera control (*i.e.*, a camera controlled by a trained human operator).

The operator was located in the same room as the worker, but could hear

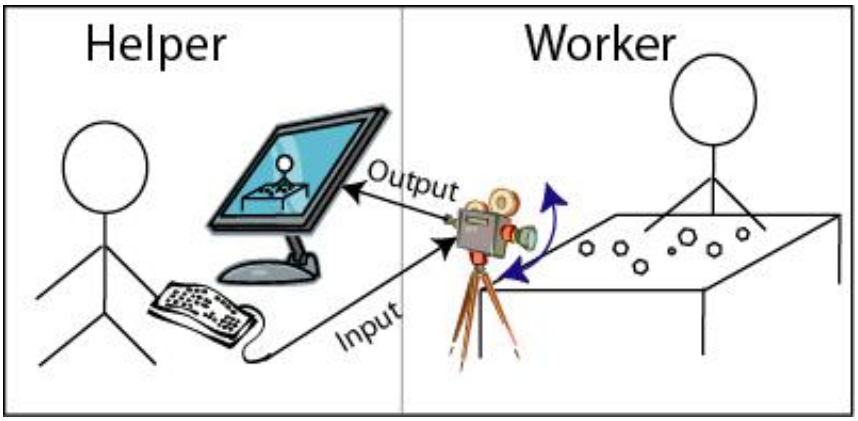


Figure 4.2: A schematic diagram of the Helper-controlled camera condition.

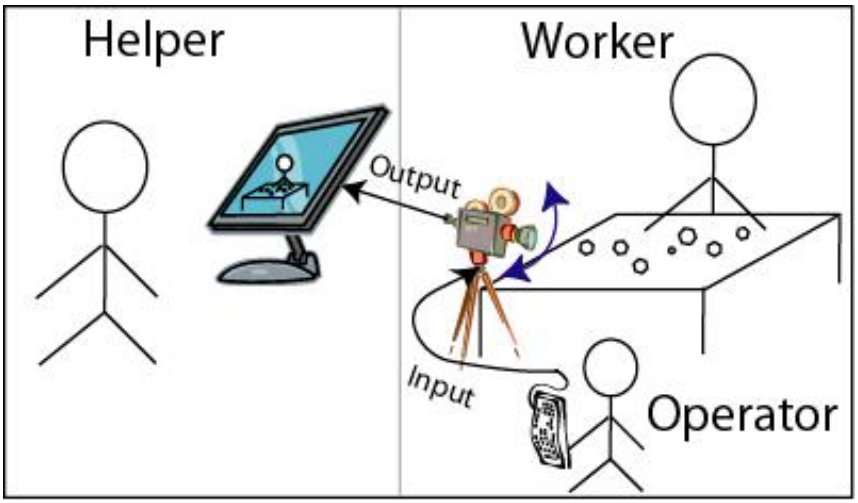


Figure 4.3: A schematic diagram of the Operator-controlled camera condition.

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 workers 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.

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.

4.1.3 Exploratory hypotheses

With regard to the effect of camera control condition on performance, we formulated several exploratory hypotheses. These hypotheses were used to guide the analysis of experiment data.

- *Hypothesis 1*: Adding pan-tilt-zoom functionality to scene cameras would result in improved performance, as measured in terms of performance time, number of errors, and self reported effectiveness.

- *Hypothesis 2*: Since camera operations can have disruptive effects on helper performance in the Helper-Controlled Camera, the benefits of camera control would be strongest in the Operator-Controlled Camera.

We also expected differences across the three camera control conditions in how the workers moved their hands and how these movements were related to camera movements. Based on prior work using two-dimensional mouse tracking, we expected that:

- *Hypothesis 3*: Hand movement in an area will statistically correlate with camera focus on that area.

Since pan-tilt-zoom functionality would allow close capturing of the workspace most of the time, we expected that the worker would be less aware of the position of their hand as they work on the task. We hypothesized that:

- *Hypothesis 4*: Adding pan-tilt-zoom functionality to scene cameras would result in less constrained movements of the workers hands, as measured by comparing the distribution pattern of the workers hand position over the entire workspace during the course of the four tasks.

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:

- *Hypothesis 5*: There would be more hand movement closer to the camera (and away from the workers body) in the Fixed- Scene Camera Condition than in the other two conditions.

Finally, we expected that the nature of the task and progress in the task would impact the amount of camera movement we saw. In particular, since complex tasks had more details, we expected more camera movements. Furthermore, 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 near the end of the task than at the beginning of the task. Specifically, we hypothesized that:

- *Hypothesis 6*: Increased task complexity would result in increased camera movement in the Helper-Controlled Camera condition.

- *Hypothesis 7*: There should be less camera movement in the Helper-Controlled Camera condition during the final third of a task than in the first third.

4.1.4 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 (Ishihara Color Test) and all but one were right-handed. Since the study was exploratory in nature, we regarded this low number of participants acceptable.

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 Toronto, Canada. The study was conducted at the University of Toronto.

4.1.5 Setup and equipment

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

The workers space consisted of a desk at which the worker was seated (see Figure 4.4), that was divided into three distinct areas: the pieces area (25cm wide, to the workers 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 this space:

- *Motion Tracking*: Worker motion was captured utilizing a Vicon motion capture system [Vic08] with five cameras. The workers wore partial-finger gloves and a baseball cap (see Figure 4.4) that had wireless passive reflective markers attached to them. These markers allowed for all motions to be

Figure 4.4: 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.

tracked in three dimensions with sub-millimeter accuracy. Specifically we tracked the position of the workers 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 workers workspace. The camera was connected via analog coaxial cable to the 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 display that showed webcam video of the helpers face.

The helpers 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 helpers 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. [LKF+02]. 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 [CNM00].

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

4.1.6 Materials

Tasks

One set of Lego plastic pieces was used in each task (see Figure 4.5). 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 4.1). Difficulty of description was determined based on our own experience and that of pilot study participants.

Figure 4.5: Lego objects in order (1-4) from left to right.

Instructions

Picture-based, step-by-step directions were printed in color and provided to the helper for each task (see Figure 4.6).

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

Table 4.1: Experiment task summary.

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 the Ishihara Color Blindness Test. The post-test included questions about the collaborative activity.

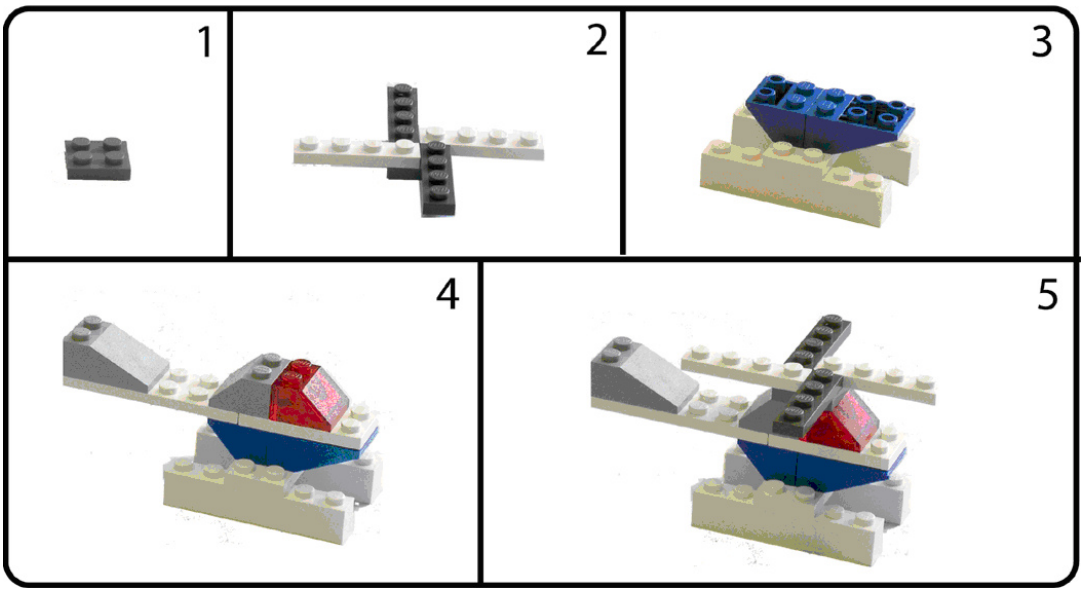


Figure 4.6: Sample instructions from Task 1 (helicopter).

4.1.7 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 prac-

tice controlling the camera for 2-3 minutes. Each participant then put on a headset and was 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.

4.2 Analysis

4.2.1 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 dialog 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.

4.2.2 Motion capture data analysis

The raw Vicon motion capture data consisted of time-stamped 3D coordinates for both of the workers 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 cameras 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 with both hands.

4.2.3 Validating operator consistency

To be sure that our camera operators 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.

4.3 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. We also tracked worker's head movement so that we could estimate the direction in which he/she was looking. We expected that head movements would be correlated with hand movements since the worker would like to look at the area where the hand is performing some task. However, we did not find any statistical correlation or distinct patterns in the head movement data. 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 excluded analyses of head movement.

While reporting the results, all significance levels were tested at $p < 0.05$, and results with $p < 0.1$ are termed as marginally significant. For correlation coefficient, we termed $0.10 \leq r \leq 0.29$ as small, $0.30 < r < 0.49$ as medium, and $0.50 \leq r \leq 0.69$ as large [DV02].

4.3.1 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 (*Hypothesis 1*, and *Hypothesis 2*). As can be seen in Table 4.2, however, the data do not support this hypothesis.

An ANOVA analysis did not show any statistically significant effect of camera condition on task completion time ($F[2, 19] = 0.15, p = 0.86$). We also compared performance time across the three conditions for each task and 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 (with partial $\eta^2 = 0.02$).

Second, we focused on the number of critical errors made by participants in each condition. Recall from Section 4.2.1 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 (*Hypothesis 1*). As can be seen in Table 4.2, 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 workers 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.

	Fixed-Scene: M(SD)	**User-Controlled: M(SD)**	**Operator-Controlled: M(SD)**
Total Time (min)	$48.5(10.31)$	$51.24(13.90)$	$48.38(9.58)$
Critical Errors	$2.57_a(0.53)$	$3.75_a(1.28)$	$2.00_b(1.63)$
Effectiveness	$4.29_a(0.49)$	$4.63_a(0.74)$	$3.57_b(0.79)$

Table 4.2: Mean Values and Standard Deviation By Condition for Performance Time, Critical Errors, and Self-Reported Effectiveness. 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 (*Hypothesis 2*), we expected self-reported effectiveness to be highest in the Operator-Controlled Camera condition.

As Table 4.2 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.* [RGC01] 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.

4.3.2 Hand movement and camera shot

We were interested in the extent to which worker hand movement correlated with camera movement (*Hypothesis 3*). While some prior evidence from screen-based puzzle tasks [OOYF05] 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 cameras 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 large with the right hand $(r = 0.54, p < 0.01)$ and medium with the left hand $(r = 0.39, p < 0.01)$. In the Helper-Controlled Camera condition, we found that this correlation was small $(r_{right} = 0.22, r_{left} = 0.24, p < 0.01)$ 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 workers 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 workers hand was outside of the camera shot.

Of these, the vast majority were cases where the workers 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 workers 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 4.7 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 movement back to the pieces area while the hand was in the work area.

4.3.3 Worker behavior modification

We were also interested in the extent to which worker behavior was different across the three camera control conditions (*Hypothesis 4* and *Hypothesis 5*). On the one hand, support for hypothesized differences in behavior would suggest

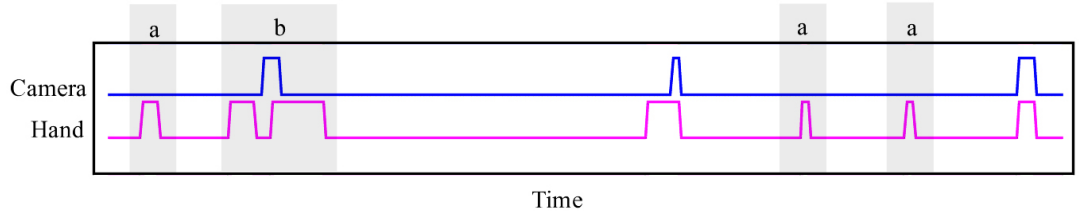


Figure 4.7: A 500 second snapshot of hand and camera movement in the Operator-Controlled Camera condition. A rise in the plot indicates move to the pieces area and a drop indicates move to the work area.

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. We hypothesized that adding pan-tilt-zoom functionality to the camera would result in less constrained hand movements (*Hypothesis 4*). When we looked at hand movements 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 the central regions of 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, when the camera could be moved to track them, workers seemed to utilize a greater range of the available space, including the peripheral region of the work area immediately bordering the pieces area (henceforth called the "intermediate area").

We performed an ANOVA to test our hypothesis that the fraction of time the worker's hand spent in the intermediate area varied significantly across the three conditions. For this analysis, we first performed a $\log$ transformation operation on the fraction data (since the Levene's test for variance homogeneity failed for the raw data). We observed a significant effect ($F[2, 18] = 4.25, p < 0.05$) of cam-

era condition on the fraction of time the workers hand spent in the intermediate area. Mean values were $5\%(SD = 3\%), 8\%(SD = 13\%)$, and $28\%(SD = 29\%)$ of total time for Fixed-Scene, Helper-Controlled and Operator-Controlled Camera conditions, respectively.

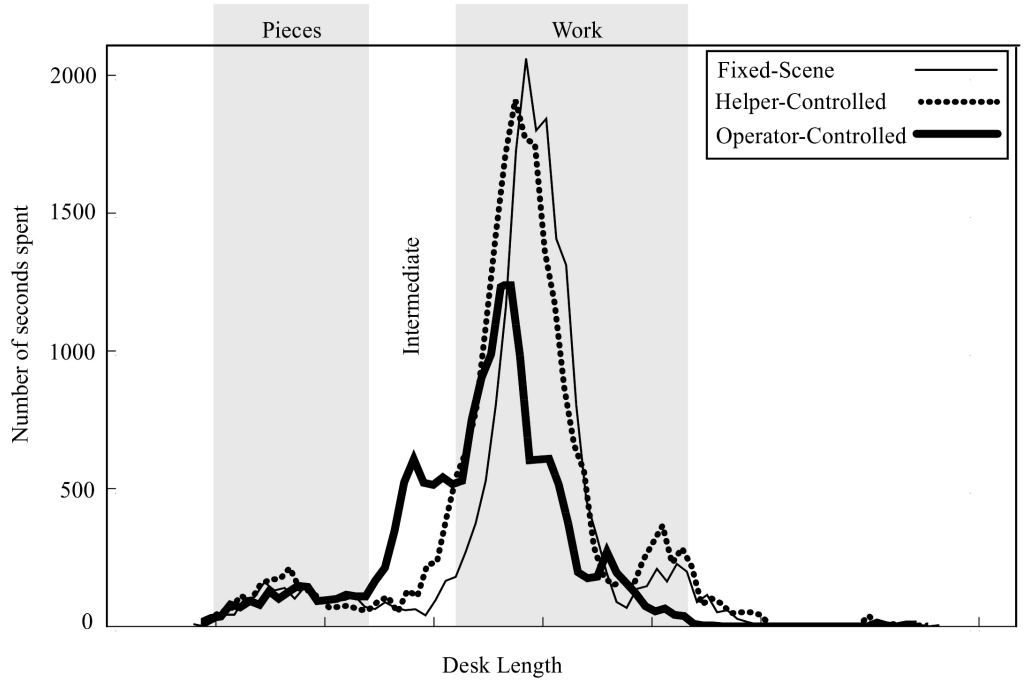


Figure 4.8: A plot of the number of seconds spent by workers' right hands in the workspace, for all three conditions.

Figure 4.8 shows a continuous plot of how long the workers right hand spent in different areas over the entire duration of the experiment for all participants, under different conditions.

We can see that in the Operator-Controlled Camera condition the workers spent more time in the intermediate region than in the other two conditions. This 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 considered worker hand activity close to and further away from the camera (*Hypothesis 5*). We divided the work desk into two halves: towards the camera, and

away from the camera. Figure 4.9 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 workers left hand indicated as black circles. The shaded area in the figure shows the desk half towards the camera.

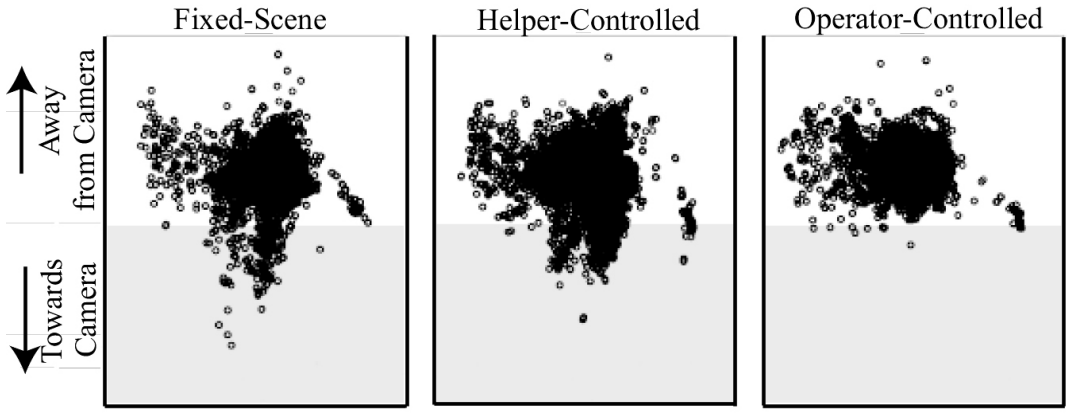


Figure 4.9: Three top views of workers' desk with left hand positions indicated for the entire duration of the experiment, for all participants. The shaded area indicates the desk half closer to the camera. In the plots corresponding to the Fixed-Scene and Helper-Controlled conditions, there are more instances (points in the scatter plot) when the hand was present in the half towards the camera.

It can be seen in the scatter plot (Figure 4.9) that in the Fixed -Scene Camera and the Helper-Controlled Camera conditions the workers' hands were present in 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 workers 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.

This result suggests 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 indicates that adding pan-tilt-zoom functionality to the camera eases the establishment of a joint focus of attention (*Hypothesis 4* and *Hypothesis 5* combined).

The fact that participants made less use of the desk region closer to the camera in the Operator-Controlled Camera condition suggests that there was less need for workers to move objects closer to the camera to distinguish them from the rest of the shot, because the camera was already focused on these objects.

With regard to worker hand movement above the work surface, behavior appeared to be consistent across conditions. It is clear in Figure 4.10 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 small peak emerges in the figure in between 30-40 centimeters above the desk which is just below the physical height of the camera. This peak shows that workers moved their hands above the desk and near the camera to show the pieces whenever needed.

4.3.4 Understanding camera movement

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

Why does the camera move?

We first wanted to characterize the nature of camera moves. As was pointed out earlier and in prior work [FKS00], there are several potential uses for a shared visual space. Of these, camera movement and zooming are particularly well suited

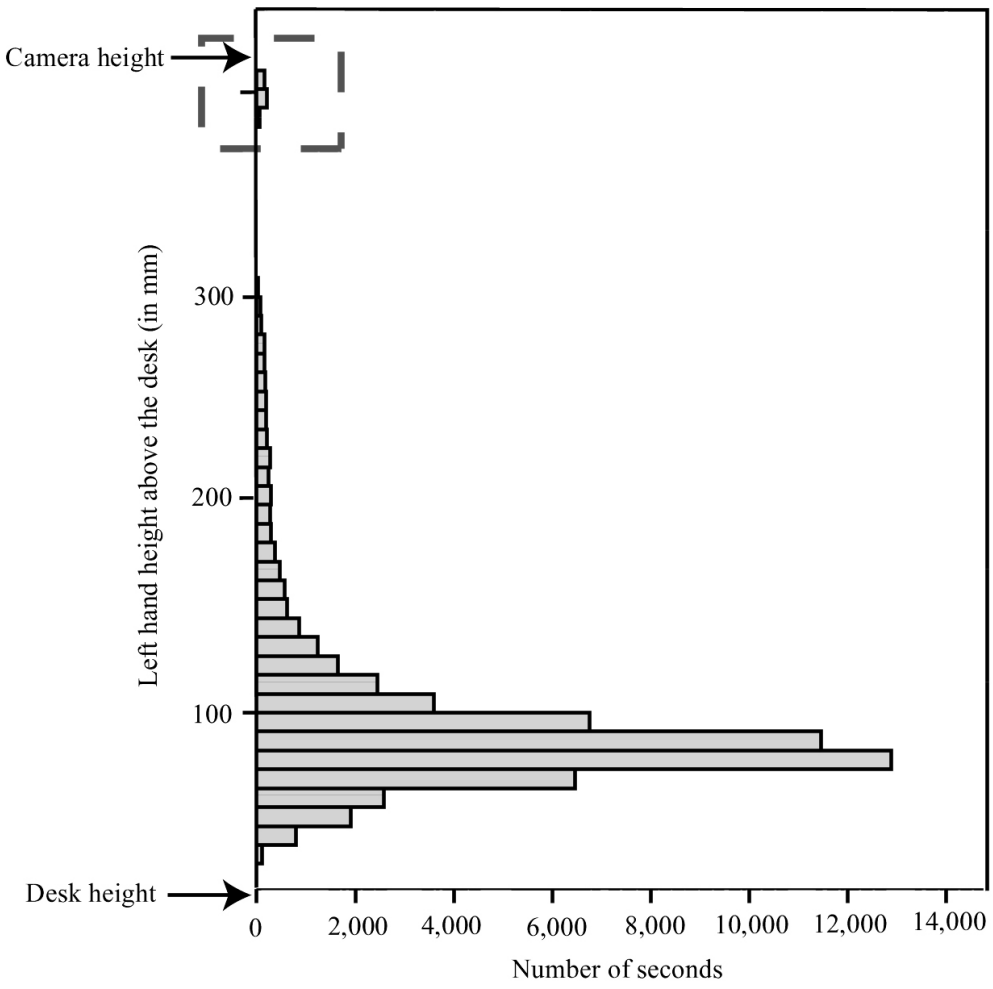


Figure 4.10: A plot of different heights above the desk against the number of seconds workers' left hand was present at a given height. A small peak around 300-400 mm above the desk is shown in dotted rectangle. This rectangle, just below the cameras physical height, shows the region workers used to show the objects to helpers.

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 workers 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 (see Table 4.3). 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 workers face or otherwise monitor comprehension.

	Detailed	**Higher level**
Joint focus of attention	23%	29%
Monitoring	27%	21%

Table 4.3: Distribution of the various purposes of camera moves.

This suggests 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).

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 (*Hypothesis 6*), and in the early part of each task, when ambiguity about piece selection and object form were highest (*Hypothesis 7*). 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 ($F[3, 28] = .60, p = .62$). Thus Hypothesis 6 could not be verified. 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 noticed a statistically significant main effect in an ANOVA analysis performed on the $\log$ transform of the data ($F[1, 39] = 4.7, p < 0.05$). 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 0.6 per minute ($SD = 1.2$, Median = 0.0) in the final third. This presents evidence in support of Hypothesis 7. 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.

4.4 Discussion and conclusions

4.4.1 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 visual 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 that there may be some psychological value in providing participants with a manual override in an automated setting that could boost perceived control and effective-

ness.

We also hypothesized that camera movement would correlate with hand movement. While there was a medium 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 seemed 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 that moment. This adaptation to technology is consistent with a broad range of field observations [OO01], and the mutual adaptation of users, technology and the environment is reminiscent of design scenarios described by Furnas [Fur00].

This modification could also be related to Benford et. al.'s [BBFG94] idea of

Aura, Focus and Nimbus. Workers were trying to bring objects in the Focus of helpers. In this sense, the visual space created by the camera causes the collision of the Auras of the participants.

4.4.2 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 workers 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 the camera was moved 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.

Another question raised by the utility of worker motion driven pan-tilt-zoom camera movement is why such a system would be useful when prior head-mounted cameras (that also, to some extent, track worker motion) have failed to show performance benefits [FSK03]. This question merits empirical exploration, but one

possible explanation is that a fixed-position camera such as the one used here need not move every time the worker moves, and can be moved independently of worker motion when this is desirable. The head-mounted camera also forces the helper to guess the worker's head position in making sense of a view, rather than having a fixed point of view (as in this case).

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 [NS03].

4.4.3 Limitations

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 workers reference frame (*e.g.*, "on your right").

4.5 Concluding remarks

One of the findings of the study was that when the camera view was controlled (either by the helper or by the operator) the movements were minimal and made only when necessary. This relates to the fundamentals of television production with regard to screen motion and shot stability.

While several previous efforts to provide dynamic visual space by using head mounted cameras failed to show significant benefits over a static camera, a pan-tilt-zoom camera controlled by humans in our experiment showed potential benefits, even though the view did not tightly follow the hand. This result is surprising from a camera control perspective, but rather predictable from a television production perspective.

In the next chapter we build on these observations and design heuristics that handle screen motion in accordance with television production principles.

Chapter 5

Camera Control for Simple Scene with Critical Visual Information

In this chapter, we build upon the previous chapter by exploring the basic premise that the worker's 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 run a study to evaluate it. Since most current meeting capture systems use a single overview camera to capture meetings, we set the goal of our evaluation to examine if the automatic camera control performs any better than the state of the art.

5.1 The study

In this study, we compare the effectiveness of automatic camera control with that of a static overview camera. Through this study, we also want to 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. Furthermore, we are interested in developing insights for the design of automatic

systems that have roots in prediction, but that exploit adaptations in user behavior.

5.1.1 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 [Ger06], 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.

5.1.2 Hypotheses

With regard to the effect of camera configuration on task performance, we proposed the following hypotheses:

- *Hypothesis 1*: Participants would complete all tasks faster with the automatic camera than with the static camera.
- *Hypothesis 2*: Participants would make fewer errors in the automatic camera configuration than in the static configuration.

- *Hypothesis 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:

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

Based on our previous observations of behavior changes due to camera movements (see Subsection 4.3.3), we also expected differences in worker behavior:

- *Hypothesis 6*: Hand movements towards the camera will be less in the automatic camera configuration than in the other two configurations.
- *Hypothesis 7*: The use of the dominant and non-dominant hand will differ significantly across camera conditions, *i.e.*, participants would adapt their behavior depending on the type of camera control provided.

5.1.3 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.

The number of participants for the study was selected arbitrarily based on the numbers suggested in previous similar studies by other researchers. And since our statistical tests showed moderate effect size (see Section 5.3), we decided to report the results without repeating the study with more participants.

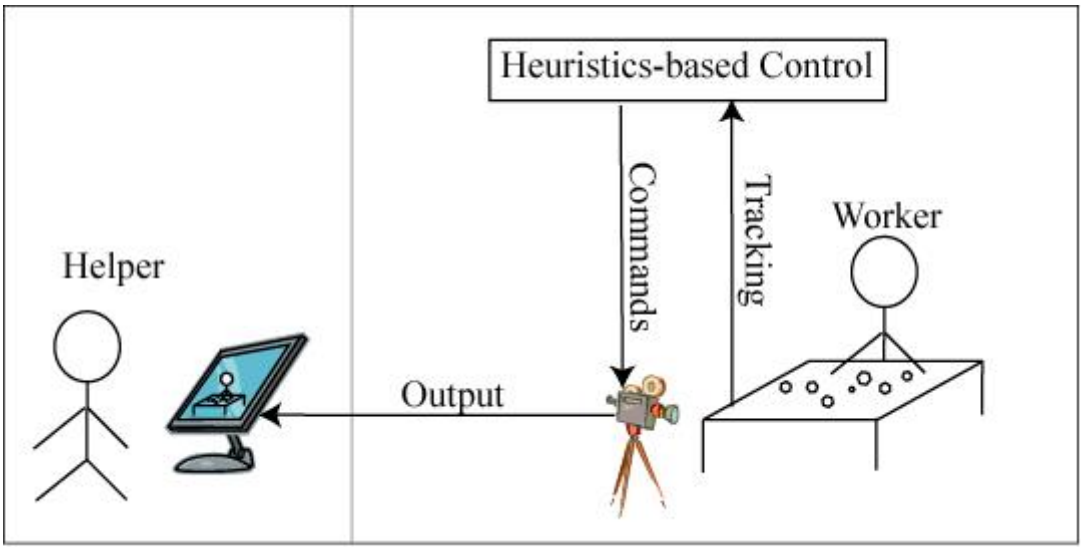


Figure 5.1: A schematic diagram of the system setup.

5.1.4 Setup and equipment

The helper and worker were located in the same room, so they could hear each other, but separated by a 5-feet-high partition wall. The worker was seated at a desk (Figure 5.2) 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.

Motion Tracking

The workers wore partial-finger gloves (see Figure 5.2) that had wireless, passive reflective markers attached to them. We tracked the location of these markers with sub-mm precision [Vic08]. 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

Figure 5.2: Workers space showing position of the camera, the monitor and workspace on the desk.

Camera

A Sony SNC-RZ30 pan-tilt-zoom camera was positioned on a tripod 30 cm behind the workers space, and above the workers 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 workers desk. It displayed the camera output so that the worker was aware of what the helper could see. The helpers 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 workers space, and recorded all sessions for later analysis.

5.1.5 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 Figure 5.3). 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.

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.

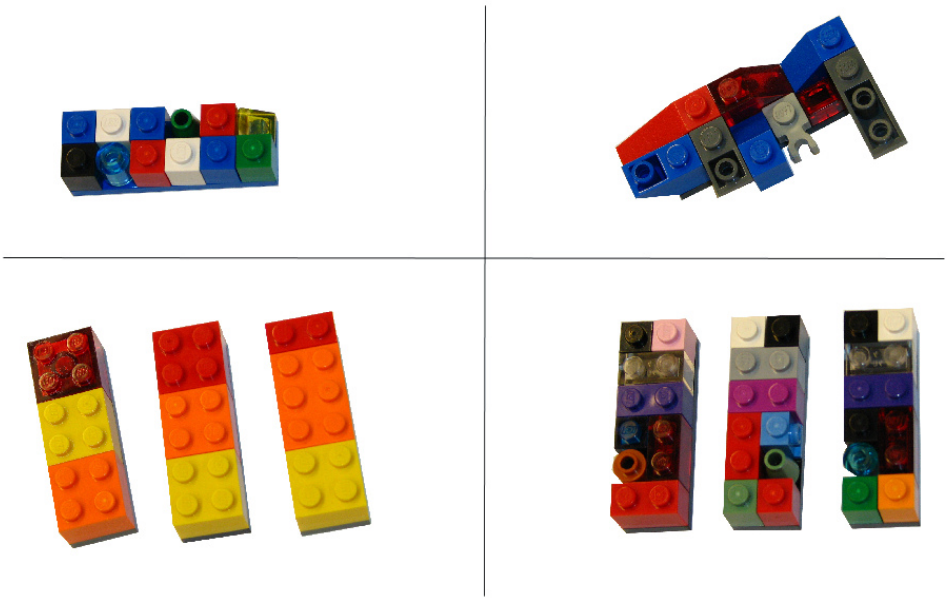


Figure 5.3: Top-left: Simple construction, Top-right: Complex construction, Bottom-left: Simple identification, Bottom-right: Complex identification.

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.

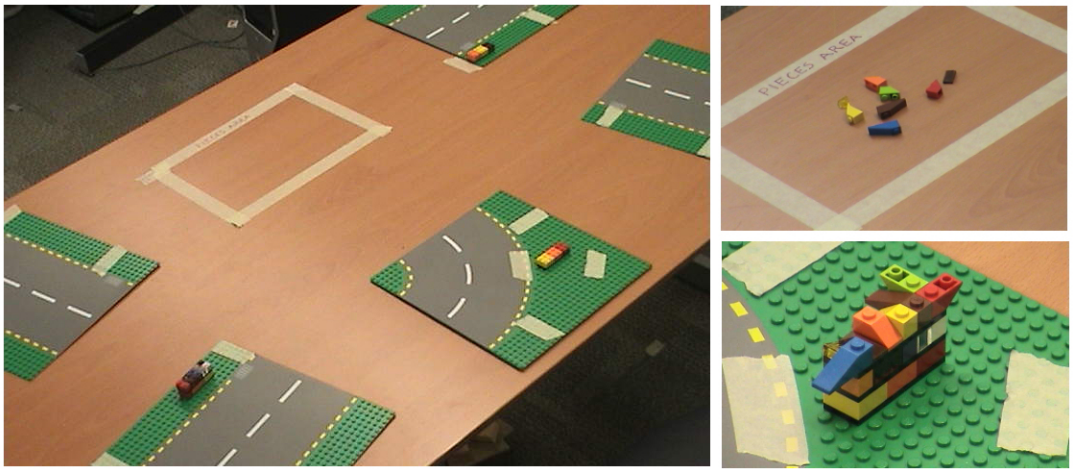


Figure 5.4: Left: Wide shot of the workspace, Right: Example close-up shots (Top: pieces region, Bottom: work region).

5.1.6 Camera control system

The automatic camera control system was based on data from the study described in the previous chapter (Chapter 4) 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 5.4). 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.

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 (see Table 5.1).

	Movement	**System action**	**Rationale**
1	The dominant hand enters a current work region that is different from the previous work region	Show the overview shot.	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	The dominant hand stays in the current work region for at least 3.5 seconds after Movement 1	Show close-up of current work region.	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	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)	Immediately move to close-up of the current work region.	Moving from the pieces area to a work area typically indicated that detailed work was about to occur.
4	The dominant hand moves to the pieces region and stays there for at least 2 seconds	Show close-up shot of the pieces region	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 the previous study (see Chapter 4).

Table 5.1: System actions for different types of hand movements.

Figure 5.5 shows a state diagram of the automatic camera control. The states

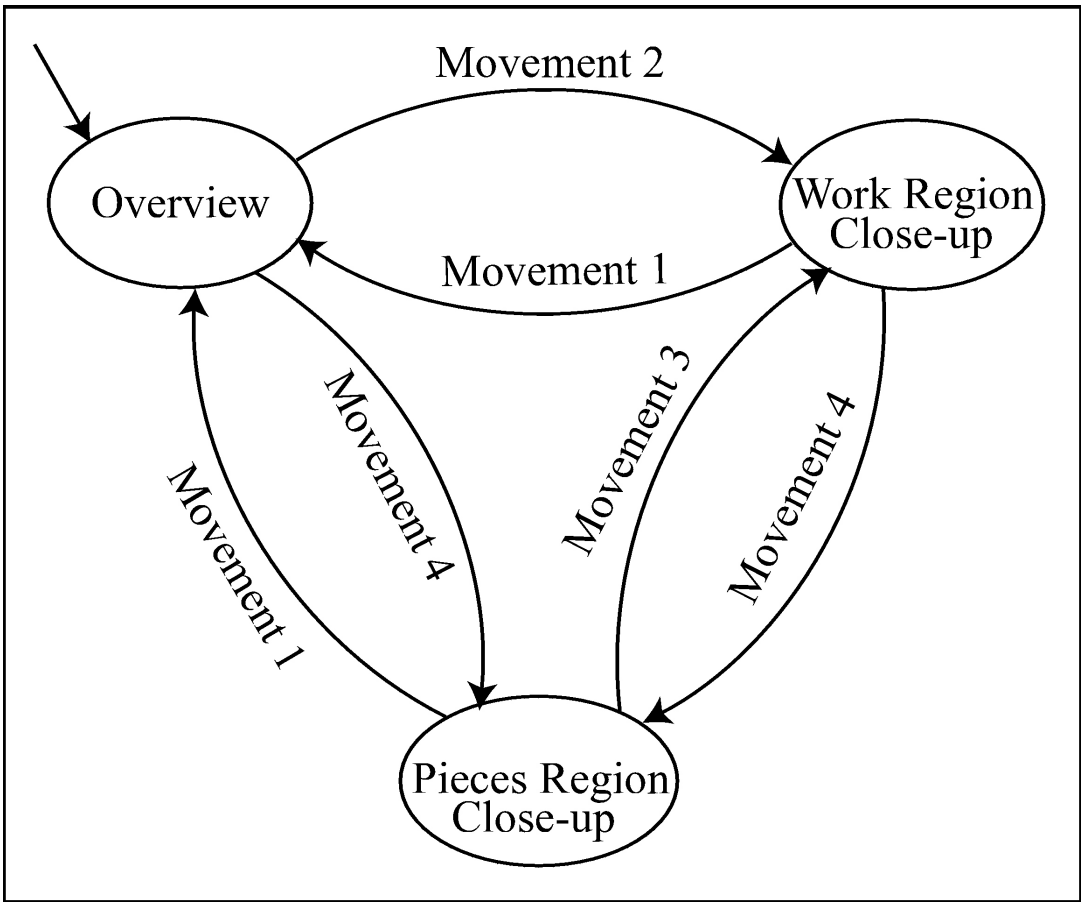


Figure 5.5: State diagram of the camera control algorithm showing the three camera shots as states and various movements as transitions from one camera shot to another.

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. This provision was motivated by the principles of handling screen motion and shot stability in television production (see Chapter 3).

5.1.7 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.

5.2 Analysis

5.2.1 Completion time and error analysis

Videos of each session were 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.

5.2.2 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.

5.2.3 Questionnaire data analysis

Reliability of the questionnaire items was assessed using Cronbachs α, which is a measure of the extent to which a set of scale items can be said to measure the same latent variable [DeV03]. 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 [Nun78]. The one remaining scale had an α value of 0.62,

which is acceptable for exploratory work. Using the Principle Component Analysis we also extracted the number of components for all the questions measuring the same construct. The result of the analysis showed a single component for all the questions measuring the same construct. The component selection criteria was set as eigenvalues > 1.

5.3 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.

5.3.1 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 ($F(1, 11) = 15.41, p < 0.01,$ partial $\eta^2 = 0.6$). No significant asymmetric transfer was observed between the two camera conditions.

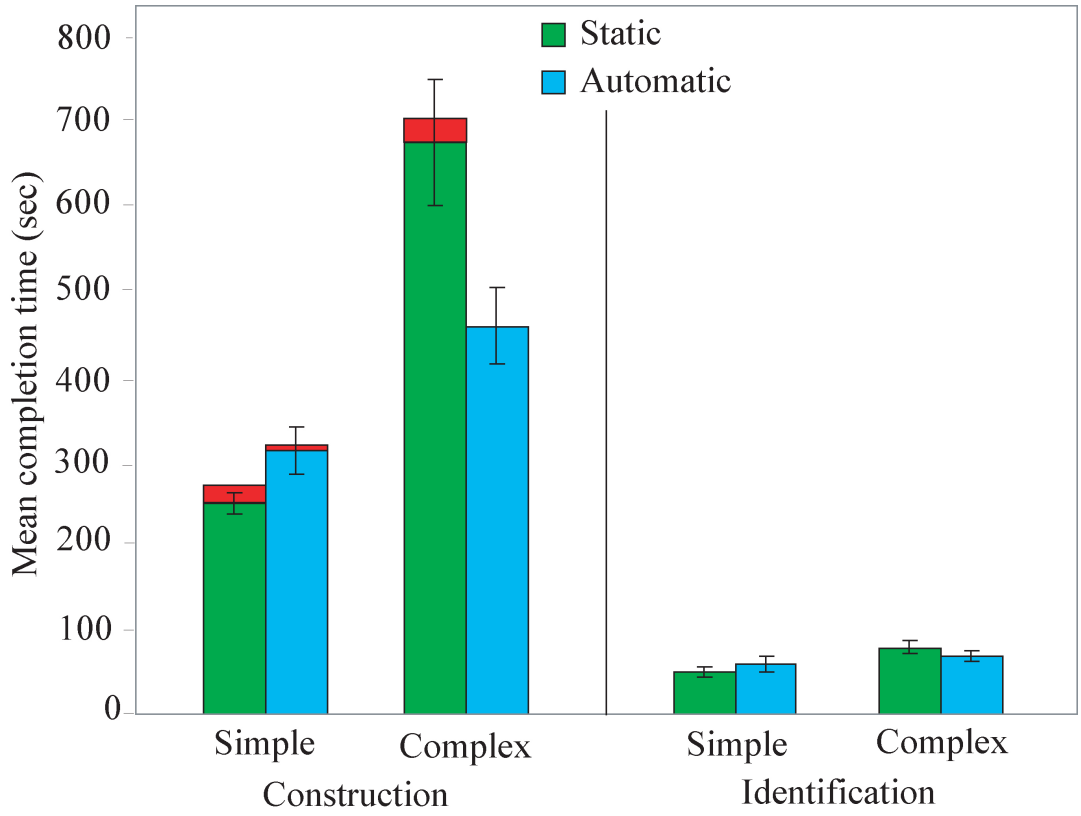


Figure 5.6: Mean completion time by camera condition for both task types. Error correction times are shown in red.

Paired sample t-tests for log transforms of 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 $(M = 680.6s, SD = 258.6)$ $(t(11) = 3.09, p < .05)$. For the simple tasks, the static camera condition $(M = 250.3s, SD = 45.6)$ was significantly faster than the automatic camera condition $(M = 313.9s, SD = 95.4)$ $(t(11) = -2.48, p < 0.05)$. The log transformation operation was performed to reduce the skewness in the data. 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 5.6 shows mean completion times under various conditions for the construction task. The error correction times are shown on top of the bars.

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 (see the right half of Figure 5.6). It should be noted that identification task completion times are substantially shorter than construction because the task involved fewer discrete steps.

5.3.2 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 [Wei99]. 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. Due to a small number of total errors, we did not perform statistical tests on this data. Instead, we report some descriptive analysis here. All the errors were found 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.

5.3.3 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 5.2, 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 ($F(1, 20) = 3.66, p < .1$).

5.3.4 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 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.

	Static Camera: Mean (SD)	**Automatic Camera: Mean (SD)**
Pair performance ($*$)	$5.8(0.6)$	$6.0(0.6)$
Individual performance ($**$)	$5.4(1.0)$	$5.7(0.7)$
Ability to see details ($**$)	$3.1(1.4)$	$5.9(1.4)$
Utility of video view ($**$)	$2.9(1.3)$	$5.2(1.2)$
Awareness of partner location	$5.5(1.3)$	$5.7(0.9)$
Difficulty of learning	$5.6(1.2)$	$6.0(0.7)$

Table 5.2: Mean ratings and their SD for performance, effectiveness of visual space, and learning under the two camera conditions. Asterisks indicate statistically significant mean differences as follows: ($*p < 0.1$); ($**p < 0.05$). All items used 7-point Likert scales.

5.3.5 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.

5.3.6 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 worker's 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.

5.3.7 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 5.3).

	Simple	**Complex**
Identification	60.7	70.3
Construction	79.3	83.4

Table 5.3: Percentage of time the dominant hand was in the camera shot for different tasks.

As can be seen in Figure 5.7, 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

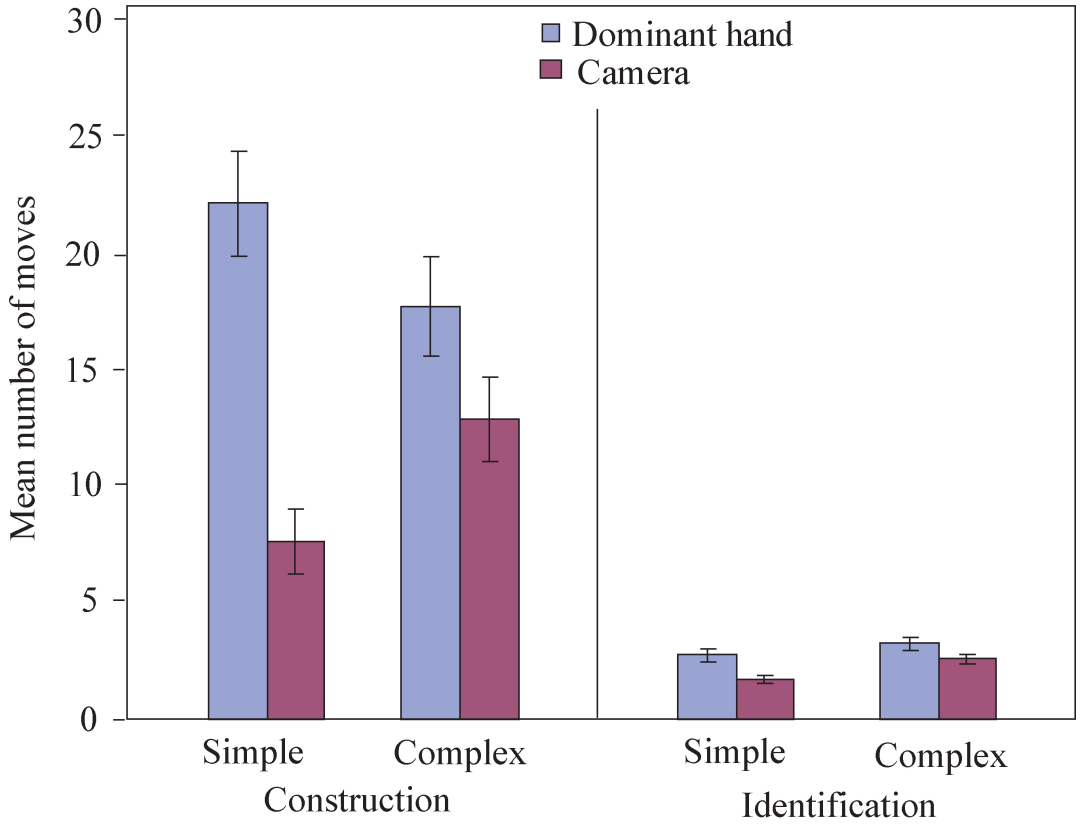


Figure 5.7: Mean number of moves with standard error of mean for the dominant hand and the camera by task complexity for both task types.

intents of our initial system design.

5.4 Discussion

5.4.1 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. The reason participants finished the simple task faster under the static camera control condition was their reliance on verbal communication under this condition. Since the static camera control did not provide enough visual details, participants relied on verbal communication for both simple and complex tasks. For simple tasks, this mode was highly effective since a precise description (*e.g.*, "blue piece", "green piece") of a piece was easy and enough to form a common ground.

Under the automatic camera condition, they performed significantly worse because of the time taken to communicate unnecessary confirmations on the pieces. Whenever visual information was available, the participants tried to use that information for grounding in order to avoid any possible error in identifying a piece. For example, consider this conversation (taken from a pair performing the simple task under automatic camera control condition):

```
Helper: pick the blue piece
Worker: this one?
Helper: yeah
Helper: now, put that under the Yellow piece
```

Such a detailed communication could have been avoided in the verbal only communication by saying "pick the blue piece and put it under the yellow piece". An analysis of the videos of simple tasks for two groups (selected randomly) showed prevalence of such precise and quick instructions. This significantly lowered the completion time. However, when performing complex tasks, such quick descriptions resulted in incorrect piece identification and led to corrective moves. Corrective moves are expensive and resulted in significantly longer time to com-

plete the task. A comparison of the number of moves under different conditions confirms our explanation. This explanation is also similar to the one proposed by Clark [CB91] when discussing Cohen's remote repair study.

This partly reinforces Gergles [Ger06] 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. Our findings suggest that providing more visual details could induce more discussion, regardless of whether the discussion is required or not. Indeed, our questionnaire data suggest that the real value of the automatic camera system lies 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 [FKS00] or helper selection between multiple shots [FSP03], 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 [Sim96] 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.

5.4.2 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 workers 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 [GSHL93] room layout task, tracking the workers 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 [FKS00]. 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 workers 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.

5.4.3 Limitations

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.

5.5 Concluding remarks

At the end of the previous chapter we set out to utilize simple shot stabilization in automatic camera control. The results of the study in this chapter indicate that dynamic visual space created in this manner was useful for the viewer (the helper). However, this study explored only a simple serial task with critical visual information. As scene complexity grows (with multiple simultaneous focii of attention), determination and capturing of the most important focus of attention gets increasingly difficult.

As the next step in this dissertation, therefore, we consider a more complex scenario of a small meeting room. Although the scene complexity for a meeting room is higher than the Lego construction task, the fundamentals of automating camera control remain the same. Instead of a hand moving the Lego pieces back and forth between the piece and work regions, in a meeting room, the role of speaker is passed around from one participant to another. However, what makes this scene complex is that there could be multiple speakers at the same time, or a non-speaker person could be more important for the viewer than the speaker.

Chapter 6

Camera Control for Complex Scene

6.1 Introduction

The previous two studies involved a scenario with a single camera and a single local participant. The results of the study indicated that the single camera automatic system performed significantly better than the single static camera configuration. Since we already achieved an improved performance with a single camera, in the next step, instead of perfecting the single camera system to reach the expert human operator level, we focus on the issue of capturing complex meetings.

6.1.1 From simple to complex

According to our definition a complex scene has one focus of attention at any given point of time. Whereas, in a complex scene multiple simultaneous foci of attention could be present. We describe such a complex scene as a scene in which multiple simple scenes exist independently and interact with one another. This structure is based on Poltrock *et al.*'s [PE97] analysis of meeting activities. For example, in a simple scene only one person could be speaking, but in a complex

scene, other meeting participants could interact with the speaker either subtly (through facial expressions and nods) or explicitly (through hand gestures and verbal interruptions).

A camera control system capturing such complex scenes must take into account these interactions. We propose that a camera control for complex scenes can be designed by combining multiple simple camera control systems and making appropriate provisions for interactions among them. In television production, a director achieves this interaction by using a wide variety of shots and shot transitions. Our design of a complex camera control system also combines multiple simple camera control systems by cutting from a shot captured by one system to another. However, there are various criteria that must be met by the visual space created by a complex system, and these criteria provide guidelines as to how multiple simples systems should interact with one another.

6.1.2 Criteria for effective meeting video

We propose three essential criteria that an effective meeting video must meet.

1. It must capture enough visual information to allow viewers to understand what took place. Capturing the desired visual information can be challenging in that meetings may involve rapid dialogs, physical artifacts, presentation media, whiteboards, *etc*. [PE97, Sel92]. This requires either a single camera shot that can include everything [Pol08], or the capacity for multiple shots via a movable camera or multiple cameras [GSHL93, IOM95, LKF+02].

2. It must be compelling to watch. People's expectations for, and ability to engage with, video recordings they view are shaped by their prior experience in viewing video recordings [Rub02]. The problem with the fixed wide-shots used by many existing capture systems is that the video itself (apart from content) is monotonous when compared with professionally produced

video [IOM95]. In this regard, it could be useful to understand the techniques [Ari76, Zet05] that make television more compelling for viewers.

3. It must not require substantial human effort. Meeting participants are primarily there to attend a meeting, and typically do not operate cameras reliably when user controls are provided [GSHL93, RBB06]. Similarly, professional crews are only affordable for some events [Bia04a, RGG03].

If we assume that the third criterion requires an automated solution for everyday use, our problem then becomes one of automatically creating a video that captures necessary visual information and is compelling to watch. Capturing and recording a meeting is fundamentally comprised of three tasks, executed repeatedly: 1) determining what is or is likely to soon be the most important piece of visual information in the setting (*e.g.*, the face of the person talking), 2) getting an appropriately framed shot of that bit of information, and 3) cutting to that shot. We now turn to the problems and prior work in achieving these goals.

6.1.3 Finding the most important part of the scene

The first task in a complex environment is to determine what the viewer will want to see. In a meeting setting, this is typically the person who is talking, and prior efforts reflect this. Several systems [IOM95, LRGC01, RGC01], for example, use speaker detection algorithms to determine who is talking and select a camera known to have a shot of that person.

While effective in determining the speaker, this approach can lack the variety of shots that provide viewers with contextual information about other attendees. To address this issue, Inoue *et al.* [IOM95] augmented a speaker detection system and cut between multiple camera views using an algorithm based on shot content and transition probabilities gleaned from professionally produced television

shows. This approach adds shot variety, but their implementation, like the system cited above, does not account for human movement in transitioning between shots.

Human television crews are able to overcome these issues because they are able to see and anticipate peoples movements [DS00, Zet05]. The ability to make these predictions comes partly from experience, but also from the ability to recognize subtle cues (*e.g.*, gaze, gestures) that people are getting ready to talk or move. Reflecting this approach, Takemae *et al.* [TOM03] used gaze direction as a cue in editing video recordings of conversation. They proposed that in a meeting, the focus of attention can be predicted by finding the participant who is being gazed at by the maximum number of participants.

6.1.4 Getting the shot

After determining what the viewer is likely to want to see, the next step is ensuring that a shot is available. This involves locating the object in space, determining which camera is best suited to get a shot of it, and framing that shot properly.

While locating the object is typically easy for human directors for reasons discussed above, it is difficult or impossible for systems without some sort of motion tracking component. Previous systems [Bia98, LRGC01] coarsely tracked a single individual, such as a speaker at the front of an auditorium, using vision techniques, but most systems to date have not leveraged the potential of seeing objects or people in the 3D space.

Once objects can be located precisely, determining the camera to get the shot can be simplified by employing camera placement heuristics used in television studios. In a typical 3-camera studio (Figure 6.2), on which our prototype system is modeled, one camera is placed in the center and the other two are placed to the sides. Each of the side cameras is then responsible for shots of the partici-

pants opposite them, and the center camera typically provides wide shots as well [Ari76, Zet05]. Depending on which camera is live at any given moment, there may be some variation in how cameras are actually used to get required shots.

Finally, framing the shot also requires the ability to locate objects in space. Assuming this capability is present, television production heuristics can again assist with this process. In particular, the notion of headroom suggests that some space be left above people's heads in framing close-up shots. And the notion of noseroom and leadroom suggests that, when people are not looking or moving directly toward the camera, some extra space be left on the side of the screen toward which they are looking or walking. This serves to both make the shot look more pleasing, and to anticipate future movement by allowing room for it to occur (see Chapter 3 for a detailed discussion of these principles). While Liu et al. [LRGC01] noted the importance of these principles, they could not implement them due to inadequate technology.

6.1.5 Cutting to the shot

The final step in the process is cutting to the shot. While this may seem obvious, this step is actually subtle and nuanced. Television directors are trained to avoid certain types of cuts (*e.g.*, jump cuts where a person appears to jump on the screen) and to pay attention to visual signals, such as gaze or physical movements (*e.g.*, cutting from a close-up to a wide shot while somebody stands up rather than after the head has already left the shot [DS00]).

Cutting is the basic step in creating a single coherent visual space by combining multiple visual spaces created by individual cameras. Directors use this as a way to cause interaction among independent camerapersons views.

Liu *et al.* [LRGC01] draw on these heuristics to automate camera control in an auditorium setting where only a single speaker is typically of interest. However,

our setting is that of small meetings which are inherently dynamic and complex, with several participants of interest. Inoue *et al.* [IOM95] tackle a similar setting using probabilistic shot transitions. However, their system was limited to organized meetings where people strictly took turns to talk one by one [IOM96].

6.2 Our iterative system design process

In this section, we describe the design process we used in developing our prototype system. Throughout this process, we worked from the principles described above and sought guidance from two people with professional television directing training and experience. One of them currently is a professor in the television arts program at a local university and has over 30 years of experience in the television industry as a director and camera operator. The other is a member of our research team, who spent 8 years training and working in the television industry (see Appendix D for their brief biographies).

6.2.1 Initial prototype design

Meeting scenario and room layout

In our prototype system, we considered a small informal meeting scenario with three collocated participants. Such a meeting is common in many settings and provides us with a basis for design that is realistic, but not so complicated as to render prototyping and testing intractable.

The seating arrangement and the room layout are shown in Figure 6.2. The meeting room had a rectangular table in the center, and the three participants were seated around it. Since whiteboards are often used as a medium to present ideas in small group meetings [PE97], we placed one near a corner of the table, visible to participants and the cameras.

Number of cameras and their placement

Once we decided on the meeting scenario, the next issue was to decide the number of cameras and their placement. We consulted with the experts regarding the possibility of using a one PTZ or two PTZ cameras. The experts suggested that camera crew sometimes do use a single camera, but the camera is held by a cameraperson who can move around the meeting area to frame shots appropriately. This would not be feasible for an automatic camera control with currently available technology.

The experts suggested using three cameras instead of two cameras because of the versatility of a three camera setup. A three camera setup is often used to capture a wide range of talk shows with two or more participants and different configurations including one host, one guest; one host, two guests; two hosts, one guest; one host, 3 guests, *etc*. Based on these suggestions, we decided to use three cameras in our setup. The camera placement was based on typical studio designs [Ari76, Zet05] and suggestions of our two expert directors.

Equipment

We used three Sony SNC-RZ30 PTZ cameras (640x480 pixels resolution, IP enabled) to capture video and three Shure SLX wireless clip-on lavaliere microphones to capture audio. The wireless microphone system allowed participants to move in the meeting space without losing the audio input.

To allow the system to locate people in the meeting space, we tracked participants location and motion using a Vicon motion tracking system [Vic08]. Each participant wore a headband with passive markers. These markers were visible to an array of infrared cameras in our lab space and allowed us to track participant head position and orientation in realtime.

While these headbands with markers were required for our prototype system

Figure 6.1: Close-up shot (left) and overview shot (right) used in the initial prototype.

at this stage, in the next stage of this dissertation we develop unobtrusive vision techniques to perform tracking (see Chapter 7).

Tracked events

In order to use as many cues as possible to determine the most important visual information, the system tracked the following events using the microphones and the motion tracking system:

- *Speaker change*: Each microphone was constantly polled to read audio signals from each participant, and change in sound energy level was used to differentiate speech from silence.

- *Posture change (sitting, standing, or moving)*: The height of the participants head was calibrated to differentiate between sitting and standing positions, and head movement range was calibrated to detect if the participant was moving.

- *Head orientation*: Head orientation has been shown to be a good approximation for gaze [TOY05]. We tracked head orientation in 3D space by applying methods used by Birnholtz *et al.* [BRB07].

Shot transition: when to cut

The system used the aforementioned cues to determine what the viewer might want to see and frame a shot of it. In particular, whenever there was a speaker change detected, the system showed a close-up shot of the new speaker. When multiple speakers started to speak at the same time or took turns quickly, the system cut to an overview or wide shot that showed all three participants (see Figure 6.1). Furthermore, whenever a participants posture changed from sitting to standing or walking, the system showed the overview shot to convey the posture change to viewers.

One consequence of these shot transition rules was that, since people in meetings frequently speak at the same time or in rapid succession, the system cut to the overview shot more often than we would have liked. This issue is further addressed below.

In television production there is a notion of screen duration which refers to the duration for which a shot stays on-air. In order to avoid extremely short or extremely long shots, screen duration often has a lower and upper limit. Rui *et al.* [RGG03] also used this notion in their system. In our system, every shot had a minimum length of 3 seconds and a maximum length of 15 seconds. These bounds were decided after consulting our two expert directors and performing iterative adjustments.

Getting the shot

If, based on the shot transition rules described above, the needed shot was not immediately available, the system then had to allocate a camera for this task. Even though there were three cameras available, this was sometimes nontrivial as one of the cameras was always on-air and could not be moved quickly (as that would be jarring to the viewer). Thus, at any given moment we actually

only have two available cameras for getting a new shot. Given the amount of interpersonal interaction taking place, this sometimes meant that the system had to cut to an intermediate transition shot to free up a camera to get the shot that was actually needed.

Professional directors often approach this problem by cutting to a reaction shot from another participant or a back-up shot such as a wide shot for a short duration and then using the previously live camera to frame the new shot. In our initial prototype, an overview shot was used as the intermediate shot whenever the live camera needed to switch shots. We reserved one camera for an overview shot at all times and used the other two cameras to frame closeup shots of the three participants. However, as we will discuss in a later section, this choice resulted in several issues related to predictability and lack of variety.

Shot framing

Once a camera was allocated to get a particular shot, the next step was to frame that shot appropriately. Our system draws on the heuristics described earlier, which are implemented as follows.

First, we make use of participant head position and orientation data from the motion capture system. Headroom was created by locating the topmost point of the persons forehead and leaving $250mm$ space above this point when framing the shot along the vertical dimension.

Similarly, using the motion tracker system we located a point approximately 100 millimeters in-front of the foremost headband marker. This point was used as an approximation for the nose position and the center of the frame along the horizontal axis. This resulted in appropriate noseroom and leadroom under different view angles.

6.2.2 Expert feedback on the initial prototype

We captured a 23-minute long meeting using our initial prototype. The meeting involved three participants discussing the Arctic Survival Task [HS08]. This task was selected to ensure a substantive discussion and active participation. We gathered feedback on the video from our two expert directors by having them watch the video, comment via email, and then meet with the system developers. Their comments fit into four major categories.

Monotonous and predictable

As noted above, our initial prototype used an overview shot as a back-up shot when cameras were not immediately ready with the next needed shot. Since the discussion in the meeting we recorded was rich with multiple people talking at the same time and people taking turns quickly, the cameras often were not ready to show the new speaker. This resulted in the system defaulting to the overview shot, which led the experts to comment that the system was monotonous and highly predictable. One of the experts commented as follows: *"There is too much of the wide shot, in my directorial view, so the overall feeling of the video is somewhat repetitive... Television (and conversation on television) is about people and their faces; we want to see them talk as they converse."*

Unexpected cuts

Since participants were often talking over each other, the system could not always determine the focal person based on the available information. This resulted in some awkward cuts. For example, in one case a participant was talking and the system was showing a close-up shot of that person, but suddenly another attendee started talking over the speaker. The system switched the focus to the new speaker and the old speaker could not be seen in the shot at all, even though

they were taking rapid turns back and forth.

One of the experts suggested that the speaker should not be moved out of the shot halfway through a sentence, and emphasized the following mantra: *"There is a rhythm as to when to cut, and when not to"*.

This issue indicates that a system based only on speaker detection may not be effective for capturing meetings with rich discussion since there could be multiple speakers at the same time, and finding the appropriate focus of attention is a difficult problem in these cases.

Slow Reaction time

In television production, prediction plays an important role in shot framing and cuts. Camerapersons often predict and anticipate how people will move and frame their shots accordingly [DS00, Kun90]. Similarly, directors often try to predict the most likely next speaker and try to have a shot of this person ready to show as soon as they begin to talk. In our initial prototype, we did not have any notion of prediction.

The system waited until someone spoke; it framed a shot (if not already framed) as soon as the person spoke, and cut to the pre-assigned minimum screen duration. These steps made the systems response time noticeably long. One of the experts commented: *"The reaction time has to be quicker on the cuts: somebody starts speaking, camera repositions (if necessary) and then cut right away. That's the way a highspeed director works and keeps the audience much more engaged."*

Lack of variety

The initial prototype showed two types of shots: close-up shots of attendees and a fixed overview shot. The experts suggested including various overview shots using different cameras and shots with props and artifacts. One of them commented: *"Consider other shots for example, when they are talking about the list make it*

possible to show the list, even if a human being is not standing next to it."

Deciding when to frame a shot of artifacts is a difficult problem since recognizing an artifact as the focus of attention (such as a list in the above comment) requires understanding the role of the artifact in the context of the discussion in real-time. However, some non-verbal cues (*e.g.*, gaze, posture) could also be used to estimate the role of such artifacts. In the revised prototype, as we will discuss in a later section, we used this information in combination with the notion of noseroom to make some of these types of shots possible.

Based on the feedback from the experts, we revised our prototype design and ran an evaluation on the revised version.

6.3 Revised prototype design

The revised prototype was designed for a similar scenario: three participants informally meeting around a table and using a whiteboard. The number of cameras and other hardware were also the same; however, the camera placement and algorithm to select and drive the camera movement were significantly modified.

6.3.1 Modifications in camera placement

Following the principle of camera blocking from television production, two of the cameras were moved further apart (see Figure 6.2). This improved the composition of close-up shots (compare Figure 6.1 (left) with Figure 6.3 (left)). A person's close-up shot was framed only by the camera directly opposite to him or her. This also provided more depth in the overview shots (see Figure 6.4).

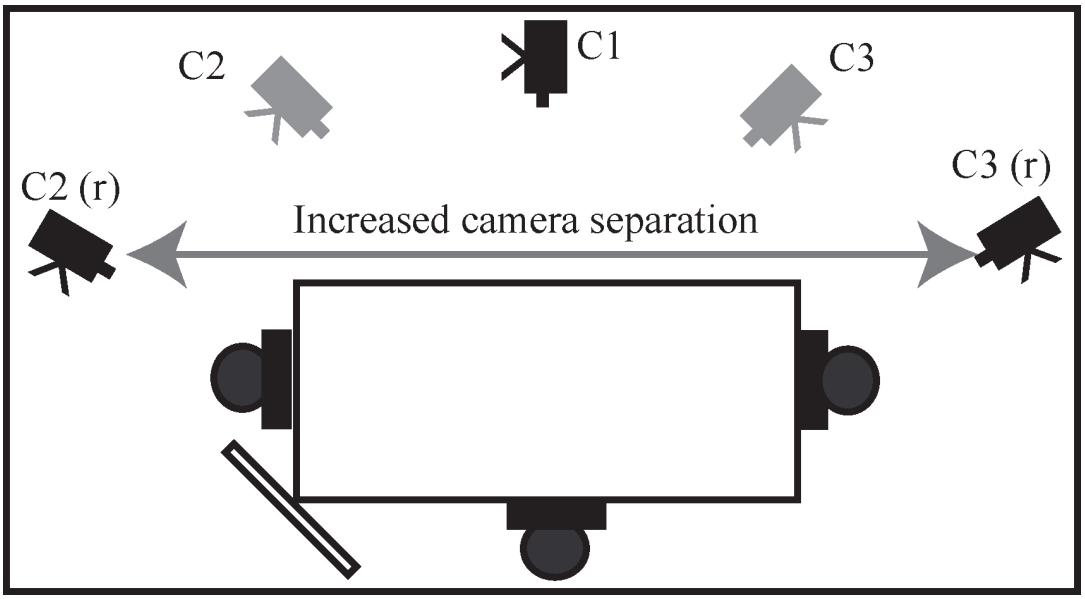


Figure 6.2: Room layout: C1, C2, C3 represent camera positions in the initial prototype; C1, C2(r), C3(r) represent camera positions in the revised prototype.

6.3.2 Use of gaze and speaker history for prediction

In television production, professionals often anticipate the next speaker by determining focus of attention of the participants. In meetings, gaze direction has been shown to indicate people's attention [TOM03, JMFV05, VWSC03].

In the revised prototype, we used head orientation as an approximation for gaze direction and used it to resolve the focus of people's attention when multiple participants were speaking at the same time. The system tracked the head orientation and estimated the person who was the most popular gaze target. The system then framed a close-up shot of the target and cut to it.

A purely gaze-based prediction and transition, however, could result in a sequence of quickly changing close-up shots if the participants engage in a heated discussion. Therefore, we decided to use this approach only when the current shot on-air was an overview shot and multiple participants started talking.

For cases in which the current shot on-air was a closeup shot, and multiple participants started talking, we use another prediction strategy that leverages

speaker history. This strategy was motivated by the observation that when two people quickly take turns it is possible to predict the next speaker. In our revised prototype, whenever two speakers took turns quickly, the system switched to a two person shot of last two speakers (see Table 6.1). This increased the probability of keeping the speaker in the shot when a new person starts speaking. This approach also addressed the issue of unexpected cuts in that when the camera shows the two person shot, the previous speaker still remains on screen along with the new speaker.

6.3.3 Variety in shots

Based on feedback and suggestions from the experts, we included a wider variety of shots in the revised prototype. These shots are commonly used in television production studios to shoot talk shows [Ari76, Zet05, DS00]. Various shots used in the final prototype are shown below.

- *Close-up shot* (Figure 6.3 (left)): Often the speaker was shown using this shot. This shot was used in the initial prototype, but the modifications in camera positions now made it possible to frame it more accurately. This shot was also used as a reaction shot we describe later.

- *Two-person shot* (Figure 6.3 (right)): Two participants talking at the same time or taking turns quickly.

- *Overview shot* (Figure 6.4). Depending on the camera that framed the shot, one of the participants was typically in full facial view while the others were viewed from the side.

- *Shot of artifacts* (Figure 6.5): We did not make provisions for explicit shots of artifacts. However, the use of noseroom and view direction allowed a close up of the whiteboard in the vicinity of participants.

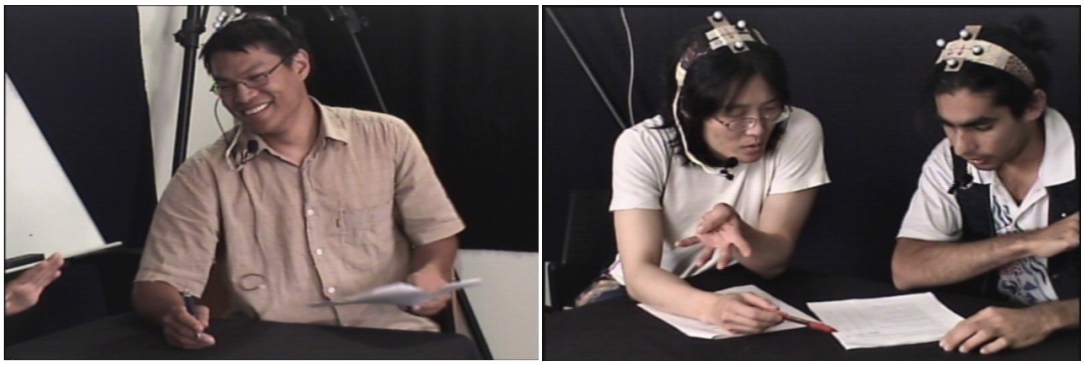


Figure 6.3: (left) Close-up shot, (right) Two-person shot.

Figure 6.4: Samples of overview shots.

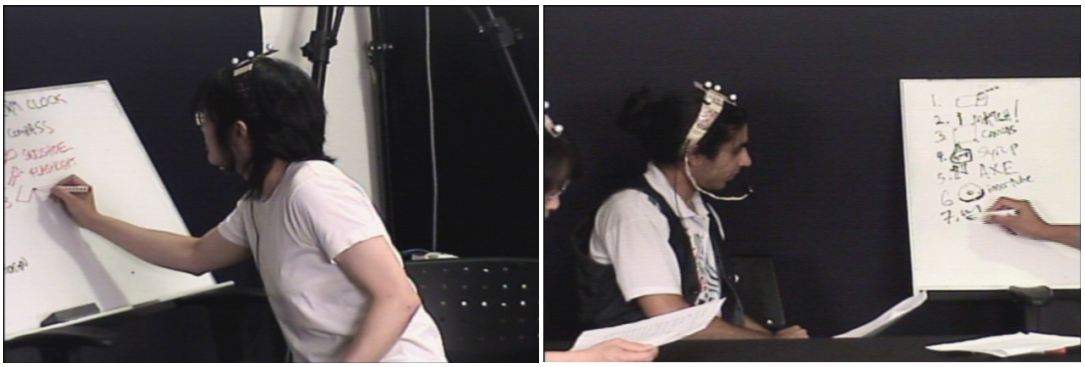


Figure 6.5: Samples of artifacts shots.

6.3.4 Modifications in camera control and shot transition

The experts commented that our initial prototype defaulted to the wide shot too often. To address this issue, we modified the camera control algorithm. Whenever a camera was not on-air, it framed a close-up shot of a participant directly opposite to it. A constraint was placed so that two cameras did not frame the

same person. This configuration had two advantages: (1) if one of the two already framed persons spoke, a close-up shot would be immediately available to cut to, and (2) a close-up reaction shot, instead of a monotonous overview shot, could be used as a transition shot.

In the revised prototype, since there were more shot types and multiple cues, the shot transition rules were more complex (see Table 6.1). One of the most important differences was the introduction of two-person shot. Although Inoue *et al.* [IOM95] also used two person shots in their system, the transition to this shot was purely probabilistic. In our system, most of the transitions were based on verbal or non-verbal cues, since that is how professionals usually decide on shot transitions [DS00].

Current shot	**Action/Event**	**Next shot**
Close-up	One person speaks	Close-up
	Two people speak	Two-person
	More people speak	Overview
	Silence	Close-up/Overview (50% probability)
	Maximum screen duration exceeded	Reaction shot of the current speaker's gaze target
Two-person	One person speaks	Close-up
	More people speak	Two-person shot
Overview	One person speaks	Close-up
	Two people speak	Two-person
	More people speak	Reaction shot of the most popular gaze target

Table 6.1: Shot transition table: the system switches from Current shot to Next shot when the corresponding Action/event happens. A close-up shot or a two-person shot always shows the most recent speaker or the two most recent speakers, respectively.

Whenever the system detected that two people were talking over each other, it framed a two-person shot using the camera which was offline and was opposite to one of the two speakers, and cut to that camera.

A cut to an overview shot was made when: there was silence, everyone was talking at the same time, or someone was standing or moving. The camera to

frame the overview shot was selected based on the most recent speaker. This selection added variety and depth to overview shots and made the recent speaker the focus of the shot. The experts emphasized the role of reaction shots in keeping the video interesting. In order to incorporate this in the revised prototype, whenever a speaker was on-screen for more than the maximum screen duration, the system showed a reaction shot of the speakers most recent gaze target.

6.4 System evaluation

In our approach to the problem of automating camera control, we assumed that human experts can capture meetings more effectively than existing computational camera control systems. The system described in this chapter attempts to automate this expertise. Therefore, we set the goal of the evaluation to estimate how well the system automates the expertise. This led us to design a study that compares the performance of the automatic system in capturing a meeting against the performance of a trained human camera crew. Our intent was not for the automatic system to surpass the performance of the professional crew, but rather to see how it measured up and if we could gain insights from the comparison.

The metric we used for the comparison was user (both expert and non-expert) feedback about the quality of the videos captured. In particular, we examined if the video captured by the automated system met the three conditions stated earlier: 1) informative enough for viewers to understand what took place; 2) compelling to watch; and 3) cost-effective in terms of human production effort. Furthermore, instead of objectively comparing the technicalities of video capture (*e.g.*, shot framing, transition accuracy, transition frequency), we relied on comparing the subjective preferences of the viewers.

Preparation of the evaluation videos

Two videos were shot for the comparative evaluation: one by the automatic camera control system and another by a trained camera crew. Both videos were about 40 minutes long and involved 3 people under the Arctic Survival scenario used in our initial prototyping phase. Different sets of three people were used in the two recordings to ensure that the participants in the subsequent comparison phase of the study did not get bored watching two videos with roughly the same content.

To ensure a valid comparison, however, both videos were recorded in the same space in our laboratory and using the same cameras. Though the details of the discussion differed across the two videos, they were largely similar in terms of their overall patterns of interaction and artifact usage.

The professional crew were instructed to replicate a professional television studio as closely as possible. A control room was set up in an adjacent space using nine video monitors (3 for camerapersons, 3 for showing camera feeds to the director, 1 for preview, 1 for program, and 1 for transitions), an audio mixer and a video switcher. A director selected and requested shots from the camera operators who controlled the PTZ cameras with a mouse-based interface.

They practiced using this interface for about 20 minutes before the recording began. We decided to use the same PTZ cameras for both videos to ensure that the two videos were as similar as possible, and to see how a professional crew made use of them.

6.4.1 Comparative user evaluation

We selected an approximately 15-minute clip from each video. These clips were selected such that they included frequent interaction with the whiteboard. Since whiteboards are common artifacts in most meetings, this allowed us to compare how well the system handled it as compared to the crew.

Participants and procedure

11 participants (4 female, 7 male, $M_{age} = 26$) were recruited at our university and asked to carefully watch these two videos (without rewind or forward) in our laboratory. They were instructed to pay attention to both the content and the quality of the recording, but they were not told that they were evaluating a camera control system. They provided feedback in the following two ways.

The number of participants for the study was selected arbitrarily based on the numbers suggested in previous similar studies by other researchers. And since our statistical tests showed moderate effect size, we decided to report the results without running the study again with more participants.

First, they were provided with a physical slider at the beginning of the experiment. By moving the slider head, they were able to continuously express their satisfaction (at the integral scale of -3 to $+3$) with what they were seeing. The center of the slider represented the neutral rating (or 0). Similar techniques have previously been used in focus groups and for measuring emotional responses [Lot07]. There was a small window on the screen showing the value corresponding to the slider head position. These values were recorded by the system once per second. The participants were instructed to use the slider as often as necessary so that it always reflected their satisfaction level with the video coverage (and not the content).

Second, questionnaires were administered at the halfway and end point of each video. They consisted of Likert scale and free response items that asked participants about the video contents and the quality of the coverage. The content questions were asked to ensure and validate that participants were paying attention to the video. The order in which the two videos were presented was balanced across participants.

Results: how did the videos compare?

Our first question concerned participants overall satisfaction with the two recordings. To make this comparison, we calculated the mean slider value for each participant under the two conditions by taking the sum of all the slider values and dividing it by the duration in seconds. A dependent sample Wilcoxon Signed Ranks test on the mean values ($M_{crew} = 1.1, SD = 0.8; M_{automatic} = 0.6, SD = 0.6$) indicated that participants were, on the whole, more satisfied with the crew video than with the automatically shot video by a statistically significant margin ($Z = -2.8, p < 0.05,$ Effect size $= 0.6$).). The video presentation order did not result in any quantifiable transfer effect.

While we were slightly disappointed that the system did not perform as well as the crew, we were pleased that the average satisfaction level for the automated system was positive, and that the difference between the recordings was not that great ($< 1SD$).

To understand the details of these scores, we analyzed the frequency of each satisfaction level in the two videos. We aggregated the time spent by all users under different satisfaction levels and calculated the frequencies. Since the slider values were logged every second, the percentage frequency of a particular satisfaction level (or the corresponding slider value) indicates the percentage of total time the participants felt that particular level of satisfaction while watching the corresponding video (see Figure 6.6).

The frequency data suggest that participants while watching the crew video spent 85% of the playback time in neutral or positive satisfaction level, with approximately equal amount of time in each positive satisfaction level. Whereas, for the automatically shot video, they spent 78% of the playback time in neutral or positive satisfaction level, with 10% of the playback time in the high satisfaction level. This analysis indicates that the crew video had more instances where

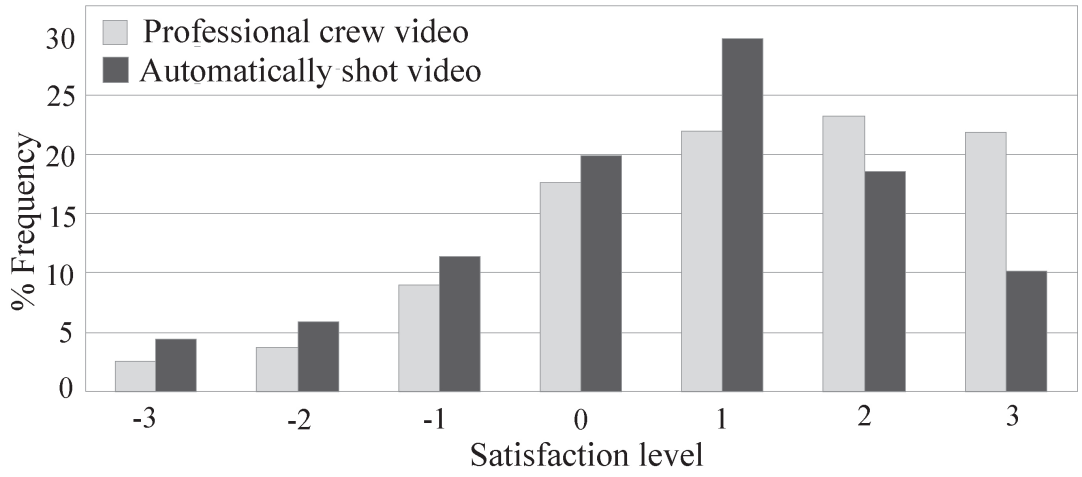


Figure 6.6: Percentage of time (for all participants) spent under different satisfaction levels for different videos.

participants were highly satisfied, whereas, in the automatically shot video participants were more likely to be neutral or moderately satisfied.

To better understand why participants seemed to be more satisfied with the crew video, we turned to our questionnaire data. One question asked participants how often they wished they could have seen something that was taking place, but could not. Figure 6.7 (left) shows that most of the participants indicated that this happened sometimes when watching the automatically shot video. However, three participants indicated that this happened often. When we analyzed the free-response comments and midpoint questionnaire results for these three participants, we observed that their responses were at the sometimes level after watching the first half of the video, but towards the end they changed it to often. The reason for this change was our systems inability to properly capture the whiteboard.

We were also interested in how often participants saw something, but wondered why they were seeing it or wished they could see something else. As it can be seen in Figure 6.7 (right), there were few times when participants could not figure out why they saw a particular shot, though this did occur somewhat more

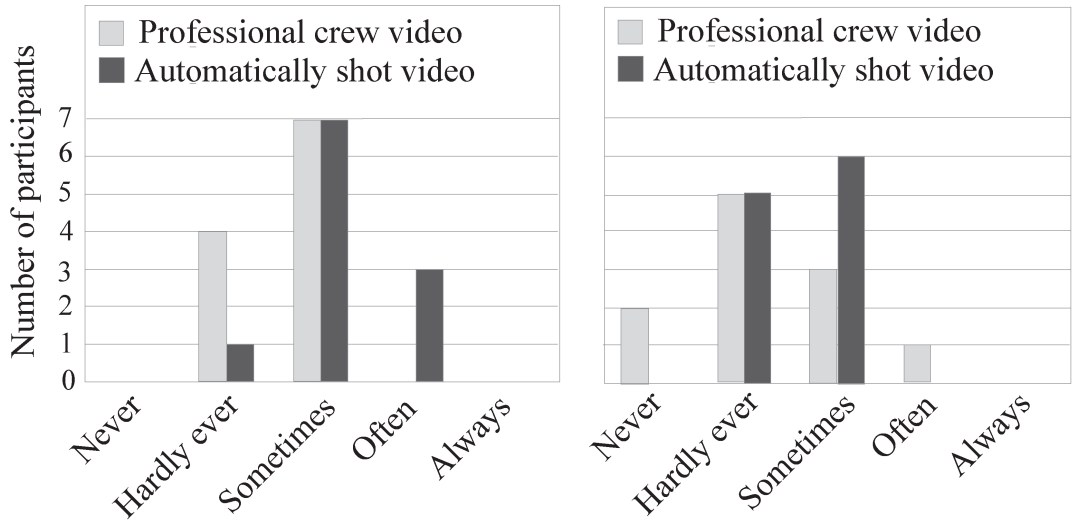


Figure 6.7: (Left) Response distribution for how often participants wished to see desired information but could not. (Right) Response distribution for difficulty in figuring out why a particular shot was chosen.

frequently in the automatically shot video. This indicates that, for both systems, most of the participants were able to figure out most of the time why they saw a particular shot.

Since we significantly modified the shot transition algorithm in the revised prototype, we were interested in estimating the effectiveness of shot transitions. While we relied on the professionals for detailed feedback about this, we asked study participants whether they felt the system was making too many, too few, or about the right number of cuts. As Figure 6.8 shows, most of the participants were split between 'About right' and 'A bit too frequently' options.

We further analyzed the data to estimate how much influence the whiteboard coverage had on the participants responses. The analysis showed that 3 out of 11 participants changed their response from 'About right' to 'A bit too frequently' after watching the second half of the video. One of them explicitly commented that frequent camera switches away from the whiteboard towards the end were tiring. Furthermore, when we performed Wilcoxon Signed Ranks test on slider values for two halves of the videos separately, we observed that the difference in

the rating was significant only in the second half ($Z = -2.2, p = 0.03$), but not in the first half ($Z = -1.9, p = 0.06$).

In both the halfway and end point questionnaires, we asked participants if they enjoyed watching the videos. While 9 out of 11 participants had generally positive responses, one was neutral and the other had a negative response. The participant who did not enjoy the videos also rated the shot change frequency as 'A bit too frequently' for both videos, and generally preferred wide shots.

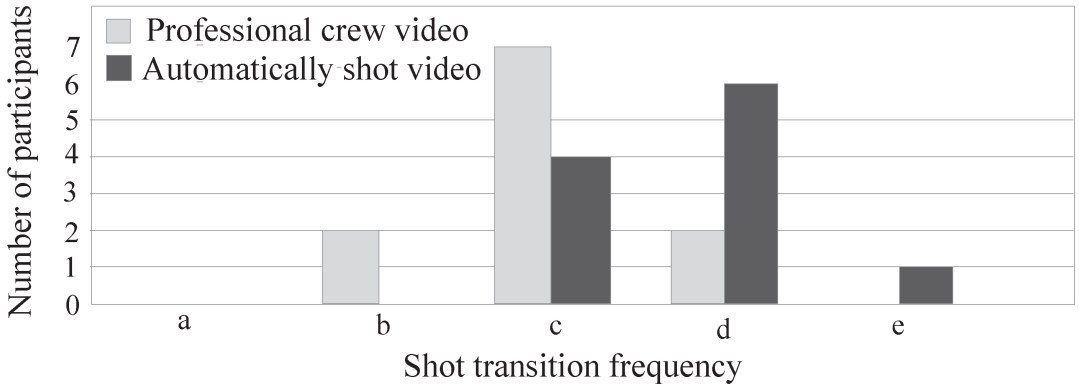


Figure 6.8: Response distribution for perceived shot transition frequency for the two videos (a) Way too infrequently, (b) A bit too infrequently, (c) About right, (d) A bit too frequently, (e) Way too frequently).

6.4.2 Expert feedback

We showed the videos to an independent expert (initially unaware of our research) who has professional experience in television studio production and is currently an editor in a television studio (this is a different person from the two experts who advised in our design phase). In order to get an unbiased opinion, we showed him both the videos without mentioning their sources. He was also unaware that the two videos were shot live without any post-production step. When asked to compare the two videos, he commented about the automatically shot video: *"Overall the video was pretty good, because the editing engaged me a little more than the first [camera crew] video. Even though it was somewhat lacking in close*

ups the multiple angles made it somewhat more interesting."

He also commented that the shot transition frequency was about right in his editorial view. However, he mentioned that the correct shot transition frequency is highly subjective. As discussed previously, this subjective nature is also evident from the distribution obtained in our comparative user study. When we later told him that one of the videos was shot by an automatic system and recorded live, he was surprised. When asked about the effects of the aforementioned issues in the video on the audience, he said that people often look over problems in live settings that would be simply unacceptable if they occurred in movies.

6.5 Discussion

We started our iterative design process with the goal of meeting the three conditions outlined earlier. In this section, we assess if we met these three conditions.

6.5.1 Does it capture enough visual information?

Assessing a system's ability to capture visual information is non-trivial since there is no standard metric for it. In our comparative evaluation, we assessed it based on the user's response to if they could see what they wanted to see. The results indicate that the system succeeded most of the time in providing enough visual information.

Sometimes when users could not see the desired visual information, it was due to the system's inability to capture various artifacts (whiteboard, papers *etc.*) in the meeting. Previous systems [LRGC01, Bia04b] approached this problem in specialized auditorium settings by showing the electronic whiteboards or slides in a separate window. In our more general setting, we attempted to address this general problem by including the artifacts (e.g non-electronic artifacts such as papers, books, coffee mugs etc) in the shots with people using noseroom. While

this approach successfully conveyed to the viewer that some activities were being performed on the whiteboard or on the list, it could not effectively capture the details of that activity.

6.5.2 Is it compelling to watch?

The analysis of slider data indicates that participants were highly satisfied ($+3$ level) for approximately 10%, mostly satisfied ($+1$ to $+2$ level) for approximately 50%, and neutral (0 level) for approximately 20% of the total playback length of the automatically captured video. The questionnaire data further support this in that participants mostly enjoyed watching the video. The expert's comments were also encouraging in that he found the editing sufficiently engaging. However, a common issue raised by some participants in their comments was that the shot transitions were a bit too frequent.

As far as shot framing was concerned, one participant specifically liked two person shots: *"It did help when two people were shown in frame, to see who was talking. The 3rd voice that was heard often said short phrases, and it could be easily extrapolated that he/she was talking even when not visible."* Two participants pointed out a few framing issues in the video where a person's forehead went out of the frame, or a part of their body was not included in the frame which should have been included.

6.5.3 Is it cost effective?

Although the slider and questionnaire data analysis show that the video produced by our automatic system was not at par with the crew video, the differences were not large. The mean slider value mean difference indicates an overall difference of 0.5 on a 7-point scale, which is less than one standard deviation. Furthermore, the percentage of time for which participants were dissatisfied with the

automatically shot video was 22% and that for the crew video was 15%, which is also a relatively small difference.

While the crew video did surpass the automatically shot video in inducing a very high level of satisfaction (22% *vs.* 10% of the time), this quality came at the cost of three professional camerapersons and a director working for approximately 2 hours (including the set up and planning time) to shoot an approximately 40 minute long meeting (see Figure 6.9).

Figure 6.9: An example television production crew setup.

Our automatic system required approximately 10 minutes preparation time and no human intervention during the shooting. To be sure, it did require a substantial investment in motion capture and sound detection equipment. It is, however, feasible to design a similar system using inexpensive vision and audio tracking. We describe the design of such a system in Chapter 7.

6.5.4 Implications for practice

Our design process demonstrated that lessons can be learned from the experts in television production to make meeting capture videos compelling. The use of audio signal level and some non-verbal cues (gaze, posture) in the design have realized a performance approaching that of a professional television production crew. Furthermore, as noted by one of the experts who advised on the design, the prototype has an interesting property that most human television production crews do not have: it does not require the content of the conversation to operate. This makes this prototype essentially language independent. Our evaluation of the prototype also suggests the importance of capturing the usage of non-electronic artifacts in meetings.

The real-time nature of the system means that our results and techniques apply not only to those interested in meeting capture, but also to those developing real-time applications such as video conferencing or webcasting.

6.6 Concluding remarks

Visual information captured from meetings is well-known to be monotonous to watch, whereas the information when captured by professionals is often compelling. Motivated by this observation, we designed and implemented an automatic meeting capture system that uses audio detection and motion tracking to apply various television production principles for capturing meetings.

While prior systems have applied some of these principles to capture lectures in auditorium settings, we extensively explored them to capture dynamic environment of small meeting rooms. A user evaluation of the system indicated that despite its limitations the videos were compelling to watch, and comparable to those shot by professionals.

Chapter 7

A Practical Camera Control System Based on Computer Vision and Audio Tracking

7.1 Introduction

The user evaluation of the system described in the previous chapter showed that the videos captured were comparable to those produced by professionals and the system generally seemed to be showing useful visual information. But the system itself was cumbersome. It relied heavily on expensive and sophisticated motion tracking equipment that required participants in meetings to wear passive reflective markers. Participants also had to wear microphones to enable the system to determine who was speaking.

This sophisticated and obtrusive technology was used in order to remove the confounding factor of tracking error from our studies designed specifically to explore the utility of automatic camera control. Having established that automatic camera control is useful with precise tracking, in this chapter we address the is-

sues involved in using light-weight tracking for automatic camera control.

In this chapter, we describe a meeting capture system that builds on the positive aspects of the previous system, but is far more accessible. We use simple vision- and audio-based tracking to capture meetings in real-time in a compelling and unobtrusive way. Participants need not wear reflective markers or microphones, and all tracking is done using basic, off-the-shelf equipment.

In our presentation of this system, we highlight two aspects of it that we regard as useful to others seeking to address this problem. First, we present heuristics for tracking individuals in meetings, including identifying potential error scenarios. Second, we propose a camera control algorithm for capturing small meetings that relies on television production principles to remain robust in the face of possible (and likely) tracking failures.

7.1.1 Tracking technologies

The various techniques for camera control described above use a variety of tracking technologies, which can roughly be divided into those that use visual and those that use sound information.

Visual tracking involves dynamically updating information about the location of specific objects in the environment, using sensing technologies that may be active or passive in nature [YKA02, Vic08]. For example, in tracking the speaker at the front of their auditorium, Bianchi used vision-based technology to track this information [Bia98].

In our two previous systems, finer grained tracking was achieved through the use of a very high-resolution motion capture system using infrared cameras and reflective markers. These systems have the advantage of very detailed tracking, but all objects to be tracked (including the participants themselves) had to be equipped with reflective markers configured in unique ways [BRB07].

Howell and Buxton [HB02] proposed an alternative vision technique to recognize gestures people could use during meetings to get camera focus. However, their approach required specific gesture based camera control, *e.g.*, raising or waving hand to attract camera focus.

Other systems have used sound-based tracking, in which inputs from arrays of microphones [BW01] are used to isolate the source of a sound in the physical environment, and a camera can then be aimed at this region of the room [LRGC01].

Regardless of the type of tracking used, however, it should be pointed out that tracking involves inherently imperfect techniques and technologies [Com01, YKA02]. As a result of this, Birnholtz *et al.* [BRB08] argue that basing camera control exclusively on tracking technologies can result in erroneous camera movements that are distracting and potentially misleading. They refer to such systems as exhibiting coupling, the degree to which physical movements by participants result directly in camera movements that are too tight.

Rather than couple camera movement directly to tracking information, an alternative approach is to use 'moderate' coupling. Here, changes in tracking information are processed according to a set of heuristics that determine when a camera shot change should take place [RBB07]. In this way, tracking information is used more judiciously - it is assumed to be imperfect and some 'intelligence' goes into determining when a shot change should take place. The key question then becomes one of isolating a set of heuristics that work in different scenarios.

7.1.2 The role of television production principles

One potential source of heuristics to guide camera control systems is television production, as we have considered before. Directors of live television programs work in a constantly-changing environment, deal with human camera operators

who are inherently imperfect, and do not have the luxury of post-production/editing to fix mistakes [Kun90, Ros99]. They must constantly make do with what they have, do the best they can to anticipate the next shot they will need, and avoid the appearance of errors in the live program [Ros99].

Television production provides a number of principles [Zet05, Ari76], that can be used in making meeting capture videos more compelling. In this chapter we focus on the role of television production in helping us to develop heuristics to overcome imperfect tracking and create more compelling videos of dynamic meetings.

7.1.3 Design goals

Beyond lecture rooms, where many others have experimented with meeting capture technologies, there are a range of meeting settings that could benefit from the technologies we describe. An effective meeting capture system could potentially make for a richer videoconferencing experience possibly even motivate participants to attend e-meetings; it could make a meeting archive less cumbersome to watch; or make a webcast less monotonous.

To date, however, systems have been plagued by problems with cumbersome and obtrusive technology, errors in tracking, and systems that are so rigid as to excessively constrain user behavior. Considering these issues, we identified the following design properties that our system must have.

- *Unobtrusive*: The system should not require meeting participants to wear sensors or get tethered in any way. This will make the system more readily usable for in-formal meetings that might often benefit the most from effective archiving.

- *Robust*: Most current unobtrusive tracking sensors provide noisy tracking data. The system should be able to handle this by making provisions for

graceful degradation and recovery. A robust capture system should not fail when its tracking component provides erroneous data.

- *Low overhead*: The setup cost of the capture system should be low, both in terms of time and money. It should not require substantial human effort to set up and operate. Furthermore, the components of the system should be cost effective.

- *Reconfigurable* Although we are considering only small group meetings, multiple variations could be found even in small meetings. The architecture of the system should allow for small variations in the setup without substantially influencing the performance.

7.1.4 System overview

Our camera control system can be described in terms of its hardware and software components. The hardware involved:

1. Cameras to track participants and capture the video, and
2. Microphones to detect speakers and capture their speech

The software components consisted of three modules:

1. Vision-based detection and tracking module that used the cameras,
2. Audio intensity based speaker detection module that used the microphones, and
3. A camera control module that used the inputs from previous two modules and, based on that input, framed the camera shots and switched between them.

In the following sections, we describe how we designed both the hardware and software components to meet the above design goals.

7.2 Hardware design and layout

7.2.1 Using camera-cameraperson metaphor

Cameras are at the heart of a meeting video capture system. If the system is required to be reconfigurable and have low setup overhead, the cameras must be selected and arranged accordingly.

We were intrigued by the versatility of TV production crews in using a relatively small number of cameras (3-4 in a typical studio setting) to capture a wide range of events and behaviors. We therefore turned to TV production professionals for ideas. In TV production, each cameraperson is assigned a camera and multiple camera-cameraperson units operate independently of one another. While they are under the overall supervision of a director, their framing decisions are made individually [DS00, Ros99]. We aimed to design our cameras in a similar fashion so that they can be operable independently of one another.

In our system we used cameras for two purposes: capturing video and tracking the location of participants. Cameras for these two purposes were physically attached to each other, and we refer to each pair as a *camera set*.

Each camera set consisted of a relatively inexpensive pan-tilt-zoom (PTZ) camera often used in off-the-shelf conferencing systems (SONY SNC-RZ30) and a basic webcam (Logitech Quick-cam pro5000). The webcam was attached on top of the PTZ camera, and their views were calibrated with respect to one another (see Figure 7.1). Each PTZ camera was connected to the controlling computer through an Ethernet connection. Webcams were connected to the USB ports of the computer. The PTZ camera frames were captured using a Matrox Morphis Quad video capture card.

Each camera set was controlled by a software module running on the controlling computer. This software module had two responsibilities:

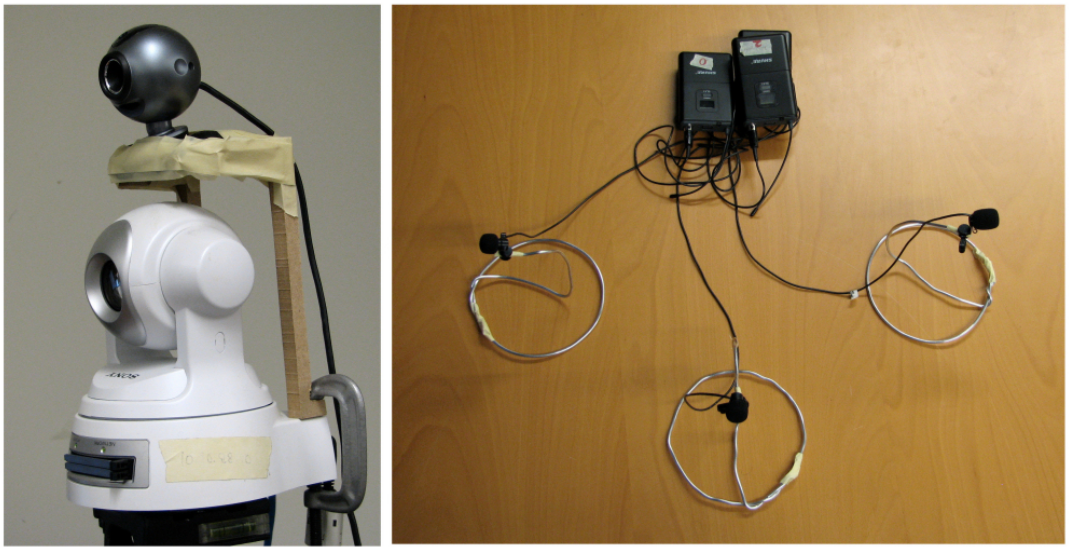


Figure 7.1: Left: Camera set with a webcam on top of a PTZ camera, Right: Microphone fan with three microphones.

- processing the webcam frames to detect faces and motion using vision techniques that we describe later in this chapter, and
- adjusting the PTZ camera to frame shots of people visible in the webcam frame.

Here the webcam and the vision based processing module acts as a cameraperson who controls the PTZ camera assigned to it. This setup allowed any given camera set to be placed anywhere in the room and still be able to frame shots of faces detected in the webcam. Multiple camera sets can be placed in the room appropriately to cover the entire scene.

In our setup, we placed three camera sets so that there was a camera set facing each portion of the scene (see Figure 7.2). The total cost of each camera set was approximately the cost of a common PTZ IP camera and webcam (1550 USD). As long as each camera set is not broken apart after a one-time calibration upon initial assembly, no additional external calibration is required to set up the system.

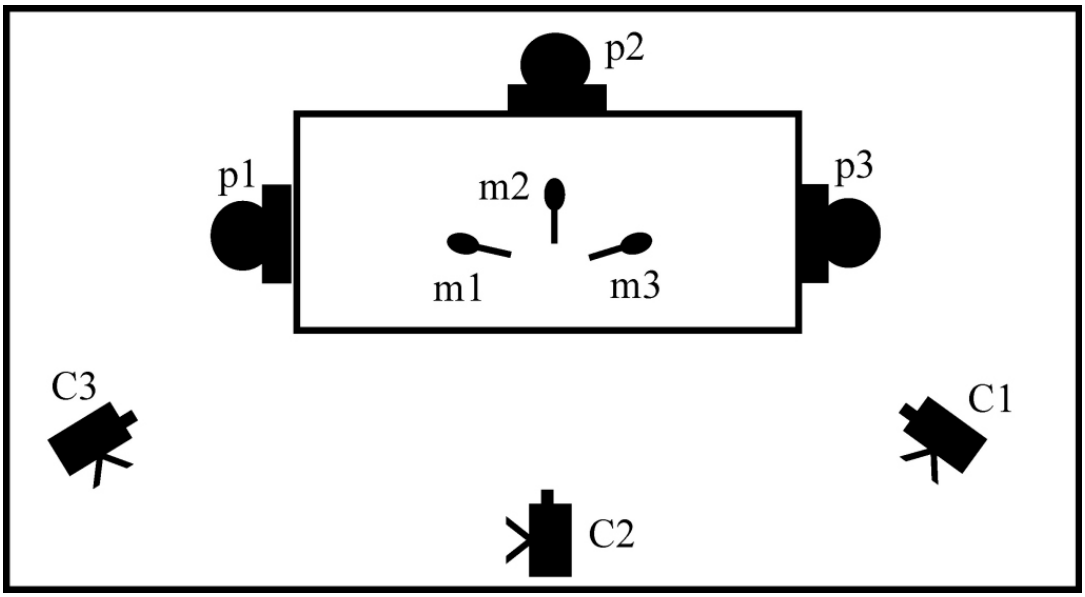


Figure 7.2: Room layout for the prototype system. There are three participants (p1, p2, p3), three camera sets (C1, C2, C3), and a microphone fan with three microphones (m1, m2, m3).

7.2.2 Microphone fan design

The other important component in any meeting capture system is sound. We used audio data to determine which participant was speaking at any given time. In our system a set of microphones was used to estimate single or multiple speakers. We used three Shure SLX wireless hyper-cardioid microphones to make a fan (see Figure 7.1).

Although microphone fans are known to have low directionality resolution [RHGL01], we used this for the following reasons:

- it is reconfigurable and simple to set up;
- our camera control system used vision tracking for framing shots, and audio tracking was used only for coarse level speaker detection;
- it allowed us to coarsely detect not only a single speaker, but also multiple speakers.

While Rui *et al.* [RHGL01] used a microphone array to coarsely track audi-

ence members and directly control a PTZ camera based on the tracking information, we combine audio and visual tracking to select cameras and precisely frame shots. The use of wireless hyper-cardioid microphones does increase the cost of the system, but a similar level of coarse tracking can readily be achieved by cheaper microphones [LZHC07].

7.2.3 Mapping camera and microphone inputs

In order for the camera sets and microphones to work together we added some constraints on the number of microphones required. Our microphone-fan-based speaker estimation algorithm requires as many microphones as there are participants in the meeting. Since the target scenario for this system is a small group meeting (3 to 6 people) and microphones are relatively inexpensive, the cost of the system is not adversely influenced by this constraint.

First, the system assigns each microphone in the fan a unique number (microphone-ID: *e.g.*, $m1$, $m2$, and $m3$ in Figure 7.2). Since the number of microphones is the same as the number of speakers, the system finds a unique mapping from microphone-ID to participant. This mapping is determined based on the room layout and camera placement.

We derived this mapping from the well known TV production principle referred to as the "$180°$" rule (see Chapter 3). This principle is intended to ensure that spatial notions of 'left' and 'right' are consistent between multiple video images of the same space, so as not to confuse viewers.

This is achieved by placing all cameras on the same side of an imaginary $180°$ line that can be drawn across the set. Interestingly, the goal of not confusing TV viewers also has the effect of simplifying our tracking problem.

Due to this configuration each camera set sees all the participants in the same left-to-right order (*e.g.*, $C1$, $C2$ and $C3$ see the participants in the order $p1$, $p2$, $p3$).

This allows the cameras to order the participants and assign a unique number to every participant corresponding to his/her position in the order. Since microphones in a fan are also ordered, the system maps a microphone to a participant (*e.g.*, $m1$ to $p1$, $m2$ to $p2$, $m3$ to $p3$ in Figure 7.2).

It should be noted that this framework can be extended to other configurations, as long as they have an 'open side'. Furthermore, it is also a common meeting room layout since it allows for a whiteboard to be placed on the 'open side' or a view of the remote site in case of video-conferencing. Similar configurations were covered in the automatic camera control system proposed by Inoue *et al.* [IOM95].

7.3 Detection and tracking algorithm design

Our system used two modalities to track activities in the meeting room: vision based and audio based. In this section we describe the algorithms used for these two types of detection and tracking.

7.3.1 Vision detection and tracking

In our system, the face detection and tracking was based on the popular Viola-Jones face detection algorithm [VJ04]. We modified the OpenCV implementation of the algorithm [BKP05] so that it searched for faces within a certain size range. By constraining the search space of the face, the algorithm could detect multiple faces and still be usable for real-time camera control.

The tracking system required initialization of the faces. In the current implementation we required participants to look at the camera for one second. In this time the tracker detected their faces, initialized the face positions, and assigned an ID to each face. Once the system starts, the detection algorithm updated face

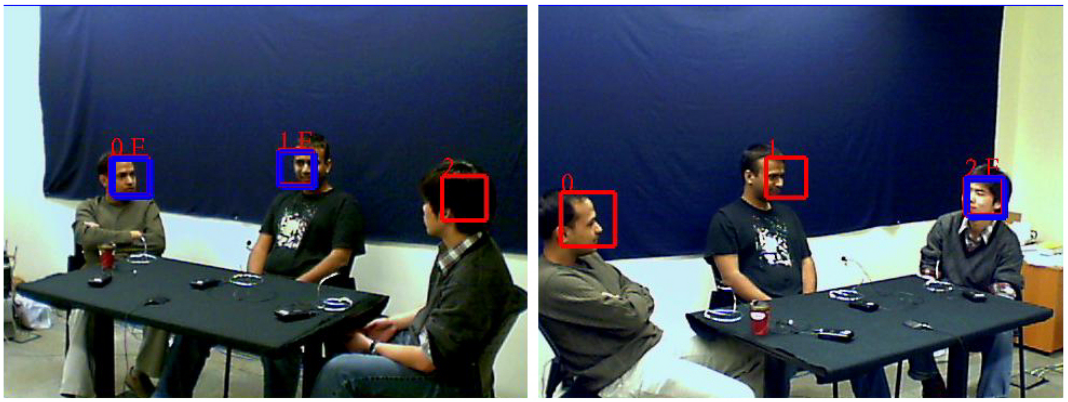


Figure 7.3: Same scene as seen by two cameras. Blue rectangle: face detected, Red rectangle: face not detected. Position of Red rectangle is the position where face was last detected.

positions every 2 seconds. We follow a simple spatial proximity based approach to find correspondence between faces found in two consecutive frames.

In our meeting settings, even though people were sitting on chairs, their facial poses were not constant. Furthermore, variations in lighting, occlusion, and facial expressions sometimes made vision based face detection erroneous. In Figure 7.3 we show two views of the same scene as captured by two camera sets ($C1$, $C3$ in Figure 7.2). One view only has two faces detected (shown as blue rectangles), and the other has only one. The red rectangles show the last position where the face was correctly detected.

There has been some interest in designing algorithm for face tracking and recognition in meeting rooms [GYW00], but the accuracy of these approaches is far from perfect when pose variations and occlusions are taken into account. While previous systems do not propose how to handle these tracking issues, our design goals motivated us to include features to facilitate graceful degradation followed by recovery in case of errors.

7.3.2 Flagging errors

Any system aiming to recover from errors must be able to detect errors, either preemptively or after it has occurred. Here we describe how we detect the two most common types of vision tracking errors [YKA02].

False positive errors

This error occurs when the tracker detects a face where no actual face is present. A system directly following the tracking results without handling these errors would 'think' it was capturing participant faces, but actually capture irrelevant objects in the meeting room, which would make the video confusing.

These errors can be flagged by considering the confidence of the tracker on the detection [IPKK06]; a low confidence value could be flagged as a potential false positive. Similarly, external knowledge of the scene could be applied to flag these errors. We derived some constraints from our knowledge of the setup. If a detected face did not satisfy any of these conditions, it was flagged as an incorrectly detected face:

- *Overlap*: Since the cameras followed the 180° rule from TV production and the room layout had an 'open side' (see Figure 7.2), faces could not overlap when participants were sitting on their chairs.
- *Face size*: Plausible face sizes were determined based on the distance of the camera sets from the participants.
- *Face location*: If the camera view is centered at the face when the participant is sitting, the face could not possibly be at the bottom or top of the frame at any point.
- *Face movement*: Face location information is not permitted to vary by a distance more than a predefined threshold in two consecutive frames. This

threshold was defined assuming that participants are moving smoothly while sitting on their chairs. When participants did make a sudden movement, *e.g.*, standing up from a sitting down position, we handled it differently as we will discuss later.

Using these constraints, the system was able to catch if a particular detected face was likely to be a false positive. It should be noted that flagging of false positives was done either during or after the run of the detection algorithm.

False negative errors

This type of error occurs when the tracker fails to detect a face where an actual face is present. These errors are more severe for meeting capture systems because they can lead to a loss of valuable information. For example if a speaker's face is present and the detector fails to detect it, a system that depends entirely on tracking would fail to capture an image of the speaker.

In order to detect false negatives, we first detected large motion (person standing from sitting posture, or walking from standing) in the scene. Since large motion could potentially result in occlusion and face posture change, it could provide preemptive warning to the camera control system that an error is likely, allowing it to respond appropriately (see Figure 7.4). Note that this detection is separate from face detection and is done at a much coarser level.

Large motions were detected by applying background subtraction on camera frames sub-sampled to one-third the original size. Sub-sampling was performed to reduce the processing time associated with the background subtraction. The system updated the background frame every two seconds and subtracted it from the current frame. This allowed the system to detect any large motion in the last two seconds. As soon as the tracker detects a large motion, it signals the possibility of false negative errors.

Figure 7.4: Shot transition sequence due to the detection of large movements in the scene: Close-up on the left, close-up with movement in the center, overview shot on the right.

Other than the aforementioned approach, a false negative was also caught when number of faces detected was less than the number of participants in the meeting (see Figure 7.3). However, this information alone is not enough since error handling strategy might also need the person-ID of the person whose face is missing.

The person-IDs for the faces detected in the current frame were determined by finding the person-IDs of the faces in the previous frame that are closest to those detected in the current frame. The person-ID that could not be assigned in the current frame was reported to be the missing face.

We now formally describe our algorithm for detecting this. Let $p_{t,i}$ represent the face position vector of the ith person at time t. Let there be three participants in a meeting, and in a frame at time t the three face positions with the person-IDs assigned are $p_{t,1}, p_{t,2}, p_{t,3}$. Suppose at time $t+1$, the detector detects only two faces: f_1 and f_2.

The system assigns person-ID k to f_i if it satisfies the following condition:

$$distance(f_i, p_{t,k}) = \min_{j \in 1,2,3} distance(f_i, p_{t,j}) \tag{7.1}$$

This procedure assumes that f_1 and f_2 are not false positives. Thus, if a person-ID p could not be assigned to any f_i then that person-ID face is declared to be missing. When the number of faces in the current frame and the last frame

was equal to the number of participants, the tracking was assumed to be correct, and every f_i gets a person-ID assigned to it.

7.3.3 Microphone fan based speaker detection

As we described earlier, our system captures audio input using a microphone fan. The number of microphones in the fan was equal to the number of participants in the meeting. The detection of speaker from these microphones was based on signal intensity: first we detected single or multiple active microphones and then determined the person-IDs corresponding to those microphones.

Intensity based speaker detection

In this particular application of meeting capture, we did not want the speaker detection system to consider every single utterance (including minor sounds and acknowledgments like 'Um' and 'Uh huh') as an occasion for a change of camera shot. Therefore, we used a temporal signal averaging filter to smooth out intensity generated by short utterances. The microphone with the highest average intensity level was selected as the active microphone (*i.e.*, a microphone with a corresponding active speaker).

Here we describe our formulation. Let $I_{m,t}$ be the signal intensity level at any given time t for a given microphone with microphone-ID m. If the length of the temporal averaging filter is T, then the microphone with microphone-ID p is detected as an active microphone corresponding to the speaker microphone if

$$(I_{p,t} > I_{noise}) \wedge (I_{p,t} = \max_{m} I_{m,x}), \tag{7.2}$$

where I_{noise} is the ambient noise intensity level for the microphones and is set during the system initialization. A higher length of the averaging filter T lowered the noise sensitivity, but also increased the delay in the system response to audio

activities.

In our prototype, we decided on a particular value of T by iteratively shooting trial videos and viewing them. However, the value of T can be manipulated to have different styles of capture video; the details are discussed later in the chapter.

Estimating multiple speakers

In an informal meeting scenario, participants often talk over one another or quickly take turns. When this happens, the output of the algorithm described above will keep switching from one speaker to another. However, we were interested in detecting all the speakers involved in the discussion so that cameras can be controlled appropriately to capture relevant information.

The main idea behind our approach for detecting multiple speakers is to detect all microphones which have similar intensity levels and which are all above I_{noise}. In order to detect multiple speakers, we first detect the intensity I of the primarily active microphone using the algorithm explained above. Next, all microphones with intensity level $I_{m,t}$ such that $(I - I_{m,t}) < K$ and $I_{m,t} > I_{noise}$ are detected as active and corresponded to speakers.

The constant K determines how tolerant the system will be in detecting multiple speakers. As the value of K gets higher, the probability of detecting multiple speakers goes down. We adjusted this parameter by recording videos at various settings and reviewing the quality of the output video. However, as we will discuss later, similar to the parameter T, this parameter can be tweaked to generate different styles of videos.

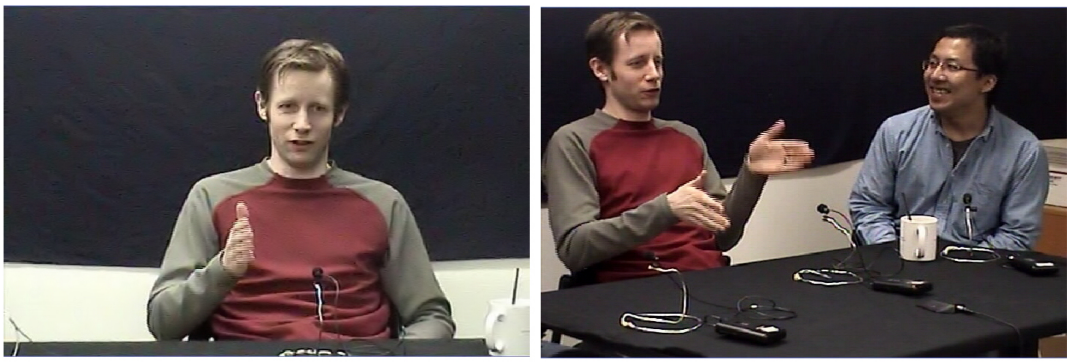


Figure 7.5: Left: A sample close-up shot, Right: A sample two-person shot.

7.4 Camera control algorithm

The camera selection and shot switching forms the final part of the capture system. In our system, the algorithm takes the following two inputs from the audio tracking algorithms described in the previous section: number of people talking and person-IDs of people talking. From the vision based tracking corresponding to each camera set controller module, it takes the following inputs: the person-IDs of the people whose faces were detected accurately for each camera set and presence of significant motion (large body movement, standing/walking) in any camera views. Based on these inputs, the algorithm decides the next shot, selects a camera for framing that shot, and cuts to that shot. In what follows, we describe each of these steps separately.

7.4.1 Deciding the shot and the camera

There were three types of shots used in the system (similar to the system described in the previous chapter):

1. *Close-up shot*: This shot is used to show a close-up of the speaker or reaction of one of the participants (see Figure 7.5).

2. *Two person shot (multiple person shot)*: This shot is used when multiple people are talking at the same time or quickly taking turns. In our prototype, there

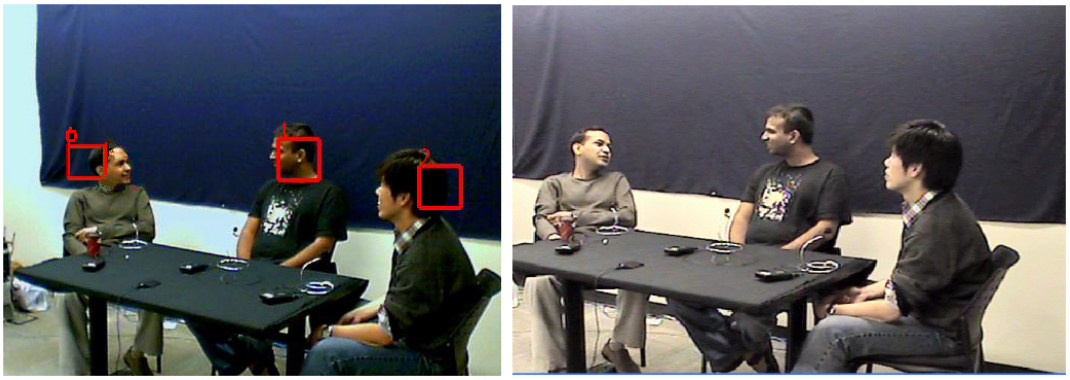


Figure 7.6: Left: Speaker's face (the leftmost person) could not be detected in the webcam frame. Right: Overview shot framed by the PTZ camera opposite to the speaker.

were three meeting participants, so this shot is a two person shot (see Figure 7.5). However, this shot can be extended to include more persons when the number of meeting participants is more than three.

3. *Overview shot*: This shot captures the overview of the entire setting, including the orientation and position of the participants, and other artifacts in the scene.

Based on the inputs from the tracker, the camera control algorithm decided the next shot using some simple heuristics described in the previous chapter. However, since the system described in the previous chapter used a more precise motion tracking system, it did not address the issue of handling error in tracking. In this system, we significantly modified both shot decision and camera selection algorithms to be more robust in the face of errors that the system might face in real world scenarios.

When the audio and video trackers do not report errors, then the system decides the shots based on some simple principles:

- When a single speaker is detected, the next shot should be a close-up shot of the speaker

	One audio source detected (without error)	**Multiple audio sources detected (with or without error)**
Face detected	Close-up	Multiple person shot
Face not detected	Overview from the opposite direction of the source	Overview

Table 7.1: Possible detector outputs and resulting system behavior.

- When two speakers are detected, the next shot should be a two-person shot
- When more then two participants are talking, the next shot should show the overview.

It should be noted that the two trackers used in the system report the results and errors independently of one another. Here we describe two of the possible scenarios involving erroneous tracking:

- The first scenario is when the speaker detector correctly detects a single microphone as active and returns the corresponding person-ID, but the vision based detector fails to detect the face of the person. The system reacts to this problem by showing an overview shot using the camera covering the portion of the scene where the microphone is located. By using this shot, the system does capture the speaker, though the shot is not a close-up and therefore lacks detail (see Figure 7.6).
- The second scenario is when there is a single speaker, but the speaker detector detects multiple microphones as active, and the vision detector is able to track all the faces. In this scenario, the system shows a multiple person shot including all the potential speakers detected by the tracker.

These two fixes make sure that the visual information about the speaker is still captured in case of errors. In Table 7.1 we summarize the different possible tracking result combinations and the corresponding shot the algorithm uses.

7.4.2 Managing camera sets for shot framing and cuts

Since our proposed design has only three camera sets and more than three shots, managing camera sets for framing a new shot becomes a non-trivial task. Once the control algorithm decides the type of shot, it passes through all the camera sets searching for the most appropriate one to frame the shot. We cast this problem here as a search task with several constraints:

- The camera set should have already detected the face of the person to be framed.
- The camera set should have the best possible view of the person to be framed.
- The camera set should not be currently on-air.

It should be noted that the first requirement makes sure that the vision tracking errors are appropriately handled. If a camera is found that satisfies only the first two requirements, then the algorithm momentarily cuts to another camera while the required camera frames the required shot. Only when the shot is ready does the algorithm cut to that camera. If no camera set could see the person's face then it is handled as a vision tracker error (see previous sub-section).

An important aspect of the algorithm is to make sure that none of the camera sets is framing something irrelevant (*e.g.*, empty space, or an empty chair). This problem occurs when a camera set frames a person and that person moves out of the frame, but vision tracking fails to track the person going out of the frame.

In the previous system, we did not take this problem into account since in an ideal tracking scenario (using motion tracking) this will never happen. A perfect tracking system will always give the position of the person and the camera will always be updating the shot to frame the person. However, when we consider

error-prone tracking, this problem could severely influence the performance of the system.

In order to address this issue, our camera control algorithm examined all of the offline (*i.e.*, not "on-air") camera set views at regular intervals (once every 2 seconds). If a camera is set to frame a person who could no longer be tracked by the vision tracker, then that camera set is changed to a wide shot; a wide shot can always be used as a "safety net" shot.

7.5 Discussion

7.5.1 Applying the framework to other scenarios

Although our prototype system consists of three camera sets and can capture three participants or less, the algorithms and the system framework can be extended to other scenarios:

More people

Our framework requires as many microphones as the number of meeting participants (see Figure 7.7). The framing strategy and the camera control algorithm will automatically include multiple person shots (*e.g.*, two-person and three-person shots if there are four participants) depending on the inputs from the speaker detector and vision tracker.

More camera sets

Our framework can easily be extended to include more camera sets as long as their placement satisfies the TV-production principle of a 180° line, *i.e.*, all the cameras should see all the participants in the same order from one side (see Figure 7.7). Since each camera set operates and reports tracking and error information

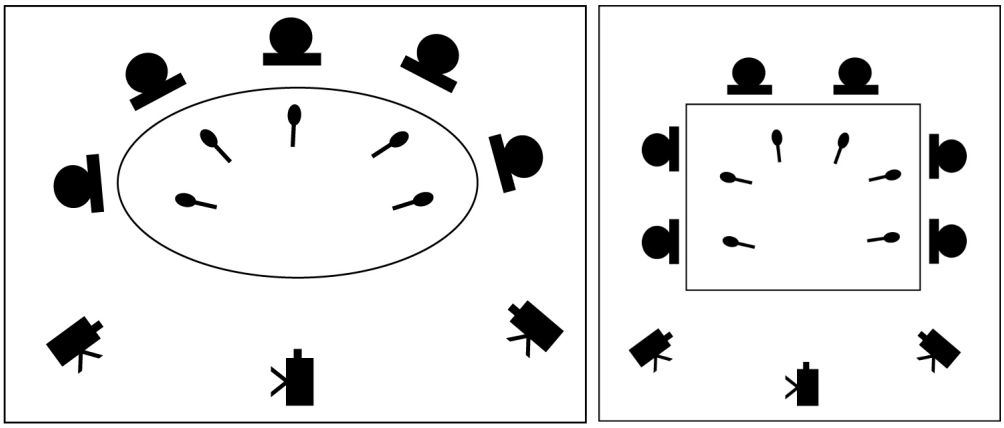


Figure 7.7: Two different meeting layouts with one "open side", which can be captured by the system.

independently to the camera control algorithm, inclusion of a new camera in the existing setup does not require any change in the previous camera sets.

We predict that with more cameras, our algorithm will find more close-up shots available when needed. Furthermore, the probability of a person's face being visible in at least one of the cameras also increases since more cameras can cover a greater range of viewing angles. However, this will come at the cost of computational complexity. Since it's a vision based system, computational cost is one of the limitations of the system.

Different room layouts

Our framework makes one important assumption about the way participants are located in the room: they are all sitting around a desk with one edge of the desk open (see Figure 7.7). Various common meeting room layouts follow this constraint [IOM95]. While Rui *et al.* [RGG03] asked videographers how they would arrange cameras for different types of lecture room scenarios, we aim to incorporate part of the knowledge of professionals in our framework itself. This general framework can then readily be applied to different meeting room layouts.

Varying capture styles

In our speaker detection algorithm we mentioned two parameters T and K that could be adjusted to achieve different styles of capture. The parameter T determines how fast the camera control should respond to audio activities in the room. A lower value of T makes the system more sensitive to audio activities. As a result, a slight utterance or quick response will be picked up by the camera control algorithm as the audio activity significant enough to capture. This could be compared to a fast paced TV director who prefers lots of cuts.

Similarly, the parameter K controls the sensitivity to multiple simultaneous speakers. A higher value of K favors the detection of only one major speaker, whereas a lower value tends to detect multiple speakers. By adjusting K, the system can be tuned to have a balance of close-up shots and multiple-person shots.

While adjusting these two parameters, we also observed potential trade-offs with regard to the perceived system error. If response time to sound is really quick, for example, there'll be more cuts, but there'll also be more shots of the person who just said something very brief.

7.5.2 Limitations of the system

Computational cost

When we ran our system on an Intel Pentium 4 processor (3.00 GHz) computer with 2GB of RAM, we recorded the CPU usage to be approximately 90%. When we analyzed different modules of the program, we observed that vision processing was the most expensive part of the computation.

A majority of vision-based tracking algorithms are known to be computationally expensive for real-time applications [YKA02], and this also becomes a bottleneck for our system. We use a modified version of the Viola-Jones face tracker

and dynamic background subtraction to detect faces and large motion. Despite our modifications to make these algorithms run faster, extending the system to include several cameras and more participants would lower the response time of the system.

One possible solution to this problem would be to perform the vision processing on a separate computer and using the output of the tracker to drive the camera control. In this direction, we recently tested our system on a dual core Intel Xeon Processor, and we recorded significant improvement in the CPU usage (50% *vs.* 90% recorded on the P4 CPU)

Inability to capture a wide variety of room layouts

The framework used in our system allows it to capture only certain types of meeting room layouts (see Figure 7.7). There are several other styles of meeting rooms (such as cabaret style meeting layout, or deep U style meeting layout) which cannot be captured by the current system. One possible approach to capture such meetings would be to divide the room into units consisting of multiple participants. Once divided into units, by treating each unit as a 'person' in our current framework such meetings can be captured.

Another possibility would be to have multiple axes. In other words, have 3 camera sets on one side of the room to capture one half of a "board room" style table, and then 3 on the other to capture the other side. This would make viewing possibly confusing, but would satisfy the system constraints and would capture everybody (and there is precedent for breaking the axis in scenarios where you just can't get the shot otherwise - *e.g.*, in sports replays).

7.6 Concluding remarks

In this chapter, we have presented a system that improves our capacity to automatically produce videos of meetings using off-the-shelf equipment, common tracking techniques, and basic television production principles. Webcams are attached to PTZ video cameras and are used in conjunction with a set of microphones to track the location of meeting participants in real time. Based on this tracking information, the system gauges the likelihood of error and relies on a set of heuristics to select and cut to a shot that is likely to provide visual information that is both useful and compelling. When a useful and compelling shot is not available, the system determines what shot is necessary and gets this shot using one of the available cameras. The framework proposed in the chapter can easily be extended to capture different meeting room layouts.

Chapter 8

Conclusions and Future Work

8.1 Summary

In this dissertation, we presented evidence in support of the utility and feasibility of automatic change in camera views while capturing visual information from collaborative meetings. We started by experimentally establishing the utility of having automatic camera control for simple meetings with critical visual information. In this study we used a trained human operator as a proxy for automatic camera control (Chapter 4).

The results of the study motivated the next step: finding cues to automatically control cameras. We identified hand position and movement as potential cues. The design and evaluation of a prototype based on these cues demonstrated that effective automatic camera control can be designed by deriving heuristics from the behavior of skilled human operators (Chapter 5). These initial steps not only established the utility of automating camera control, but also, suggested its feasibility. However, feasibility of designing such a control for complex scenes required further evidence.

In the next step, we considered a complex scene of informal meeting which

can have multiple simultaneous focii of attention (in the form of multiple speakers and listeners). Similar to our initial approach, we developed heuristics to control cameras based on the behavior of skilled human operators. In this case, skilled human operators refer to television production crew. We consulted with professional television directors and television production literature to design heuristics (Chapter 3 and 6). The design and evaluation of the system provided evidence in support of the feasibility of designing an automatic camera control for complex scene. However, this design used cumbersome and expensive technology.

In order to provide further support to the feasibility argument of the thesis, we developed an automatic camera control using audio and vision based tracking which significantly reduced the cost and cumbersomeness of the technology (Chapter 7). The system demonstrated various techniques to detect tracking errors and recover from those errors or gracefully degrade to an acceptable view.

8.2 Contributions

The series of experiments and design explorations discussed in this dissertation resulted in a number of significant findings. Overall, these findings contribute to three main areas: the role of automation, the detection of cues for automation, and the utility of TV production in camera control.

8.2.1 The role of automation

Previous studies have explored various different types of strategies to either automatically change camera views or present multiple views to provide a better coverage of collaborative meetings. However, they could not demonstrate the significant advantage of using automation over a fixed overview camera.

Through our system designs and experiments we demonstrated that camera

control can be automated in such a way that it either improves task performance (Chapter 5) or engages the viewer's attention (Chapter 6). These findings encourage future research in the field of automating camera control for capturing meetings.

8.2.2 Detection of cues

An approach to determine cues

We proposed that cues to control cameras can be learned from trained human camera operators. We demonstrated this approach in two ways.

- In a controlled experiment setting, we analyzed the behavior of the helper and a dedicated camera operator in the context of a simple collaborative task with critical visual information (Chapter 4).
- We consulted with experienced television directors and analyzed the television production literature in the context of a complex collaborative task (Chapter 3 and 6).

Some cues for a range of collaborative tasks

Our experiments showed that hand position and movement can be used as effective cues to determine visual focus of attention in collaboration on 3-D physical tasks (Chapter 4). For complex meeting scenario, we demonstrated that not only speakers, but also non-speakers can be used as cues to determine visual focus of attention and control the cameras (Chapter 6 and 7).

In general, the person or the object performing the task is an effective and practical cue for controlling the camera. Examples of such cues from our explorations and other real world tasks include a person speaking, a hand performing the construction task, a pen writing on a whiteboard, and a person moving an

object.

Some of these cues can be detected in real world settings using computer vision and audio tracking. The systems discussed in this dissertation (Chapter 5, 6, 7) demonstrated various ways to detect hand movements, detect speaker change and track speaker movements.

8.2.3 Utility of television production in camera control

Capture and presentation

We proposed how television production principles of camera placement, camera movement, shot framing, and shot switching can be used to capture meetings (Chapter 3). We applied these principles to capture simple meetings with critical visual information (Chapter 5) and complex meetings (Chapter 6 and 7).

For meetings with critical visual information, we demonstrated the importance of shot stability and handling screen motion in shot framing (Chapter 5). These findings led Birnholtz *et al.* [BRB08] to develop the notion of decoupling between the camera view and action.

We extensively used the principles of shot framing and shot switching to capture meetings with complex scenes. We experimentally proved that the performance of such a capture system can approach the performance of an experienced human camera crew (Chapter 6).

Error handling

Automatic camera control depends on object tracking which is error prone in real world settings. The basic principles of keeping shots stable during shot framing and shot switching is useful in handling these tracking errors. We demonstrated how these principles can be used to design a robust camera control based on cheap tracking technology (Chapter 7).

8.3 Limitations

In this section we discuss some of the main limitations of this research.

8.3.1 Theoretical limitations

Our studies and system designs had some theoretical limitations. These limitations can be attributed to the type of tasks and participation that were explored in this research.

Task and role limitations

The complex meetings considered in our studies involved primarily verbal discussion. There was only limited use of whiteboard and other artefacts. This limits the implications of our design. For example, a scenario involving heavy use of presentation slides may not be effectively covered by our system.

Furthermore, our designs were independent of the roles participants play in meetings. For example, in a classroom scenario, the lecturer or the instructor plays an authoritative role, but our system will consider the most frequent speaker as the primary focus and fail to assign importance to the instructor. These issues, however, can be addressed by assigning (importance based) weights to the various sources of information.

Meeting participation limitations

In this research, remote participants are assumed to be mostly passive. The implications of the studies, therefore, are limited to scenarios in which visual information from the remote site is not important.

Data analysis limitations

The focus of data analysis in this research was task performance and user satisfaction. This analysis was motivated by our focus on mostly passive remote participation. Therefore, any analysis of effects of camera control on verbal communication was missing. However, the system provides an effective testbed for future work on communication analysis.

8.3.2 Practical limitations

There are some practical limitations of this research primarily due to the tracking technologies and room layouts considered in the study and system designs.

Tracking limitations

Both sophisticated motion tracking and inexpensive vision tracking are limited in the number of people they can effectively track. Even though we propose techniques for graceful degradation in the event of tracking failure, as the number of meeting participants increases, the failures get more severe. This in turn will result in a fixed overview shot. Using other tracking modalities could be explored to address this issue.

Furthermore, affect information plays important role in communication. Our tracking sensors could not detect any such information. Kiesler *et al.* [KZMG85] mentioned three types of affect distinguished in communication: (1) physiological arousal, (2) subjective emotions or affective feelings, (3) expressive behavior. They also used specific sensors and measures to track this information. Vision based sensors can also be to track this information.

Limited exploration of audio technology

The cues to detect focus of attention explored in this research are primarily visual. One of the main cues used in studio production is speech content which was not explored in this research. The reason we did not use speech content was the ineffectiveness of the technology for real-time speech content analysis [GTHT+06, G.99]. As the technology to support automatic speech content analysis improves, this could be used as another cue to detect focus of attention.

8.4 Future Work

8.4.1 Other collaborative meetings

In Section 1.3 we identified three dimensions of collaborative meetings which influence the camera control design: role of visual information, complexity of the scene, and type of remote participation. In this dissertation, we explored camera control for meetings with: (a) simple scene and critical visual information, and (b) complex scene and non-critical visual information. One natural extension of this research is to explore meetings with complex scene and critical visual information (the fourth quadrant of Table 1.1). Any future work in this direction will involve exploring the domain of tasks which represent such meetings and cues useful for determining desired visual information.

Exploring tasks

One possible task representing such meetings could be a modified version of the room layout task used by Gaver *et al.* [GSHL93] The complexity of the task can be increased by introducing more participants. Such a task will closely represent various real world collaborative activities including informal brainstorming sessions with laptops, whiteboards and other props; remote surgery tasks with

several doctors and nurses; discussions on the design of a building or machine involving multiple designers and architects, *etc*.

Exploring cues

In a meeting with complex scene and critical visual information, the set of cues will need expansion. While in this work we used audio signal, head direction and hand/body movements as cues to determine desired visual information, various other cues can be explored, such as speech content, hand gesture, and location of the participant in the room.

In a recent study, Gergle *et al.* [GRK07] proposed that a computation model that integrates visual cues with linguistic cues represents communication more effectively than either visual or linguistic cues alone. This finding poses the challenge of integrating speech and visual cues (*e.g.*, hand position, gesture) for controlling cameras.

8.4.2 Multiple site remote participation

The third dimension of collaborative meetings as identified in Section 1.3 is the "Type of remote participation". While this dissertation deals with a single remote site with a passive viewer, the design of camera control can be explored in other complex site settings.

The next level of complexity will involve two sites with active participants. Various design issues will need to be addressed in this case, including: (a) how the conversation between remote and collocated participants should be captured, (b) where the screen showing the remote participant should be placed in the meeting room, (c) what the participants should see on the screen. Special effects discussed in Chapter 3 can be useful in such scenarios.

The inclusion of more than two sites in the collaborative activity intensifies

various issues including camera placement, screen placement, determination of effective cues, display of multiple site information on screens, *etc*.

8.4.3 Advanced television production principles

In the present work, we explored only few of the various TV production principles discussed in Chapter 3. Some of the other principles could lead to interesting camera control designs.

Replay and other special effects

Replays can be used to review recent points in the meeting by attendees present there or attendees who join late. This can enhance the meeting participation experience for both remote and local attendees.

Different types of wipes and multi-image techniques can be used to present multiple focii of attention simultaneously. Similarly, other special effects can be explored as suggested in Chapter 3.

Wide variety of shots and cuts

While pan-tilt-zoom cameras mounted on static tripods limited the types of shots that could be used in our explorations, cameras mounted on dollies [KOY+00, Jou02] can significantly increase the number of possible shots. Placing the cameras on movable mounts will also allow better coverage in case of occlusion. Different types of transitions can also be included to make the capture more engaging.

8.4.4 Studying the effects of changing camera view

In this dissertation, we demonstrated that automating camera control is useful and feasible. This opens various research questions regarding its effect on the

viewer of the captured video and the participant of the meeting.

Study of individual system aspects

In Chapter 7, we defined parameters T and K which decide how the system reacts to different speaker conditions. While this system provided tools to capture meetings in different styles, how viewers react to these different styles of capture could be one potential future research direction.

Furthermore, we quantitatively demonstrated the importance of shot stability through our experiments. This motivates future studies to understand the effect of different shots and switching on task performance and viewer experience. This might involve measuring engagement and presence, which is an area of active research in media studies and communication [LRG+00].

Holistic field study

Apart from studying individual aspects of automatic camera control, exploring the long term impact of such an installation could be another research direction. Issues related to user acceptance of such a technology can only be studied through long term holistic studies. Study designs based on Bellcore's VideoWindow study [FKC90] could be specially useful.

8.4.5 Modeling complex automatic camera control

In this dissertation, the control heuristics were based on some cues such as hand position and speaker tracking. As the number of cues increases, the complexity of heuristics to control camera will also increase. Future research could address this issue by mathematically modeling the problem of automatic camera control and using machine learning techniques to design complex controls.

One possible approach is to use reinforcement learning [SB98]. Since the out-

come of a camera control (the video) can be rewarded (or punished) based on how well it captures visual focus of attention, the reinforcement learning algorithms can search for the optimum control strategy to maximize the reward. Reinforcement can be provided if the system satisfies certain constraints, such as minimizing the number of people in shot while still capturing the speaker or the focal person. It should be noted that various television production principles can be formulated to represent several other constraints on the search space of the solution.

Bibliography

[ALM90] L. C. Austin, J. K. Liker, and P. L. McLeod. Determinants and patterns of control over technology in a computerized meeting room. In *ACM CSCW*, pages 39–51, 1990.

[Ari76] D. Arijon. *Grammar of the Film Language*. Communication Arts Books, Hastings House Publishers, New York, 1976.

[BBFG94] S. Benford, J. Bowers, L. E. Fahlén, and C. Greenhalgh. Managing mutual awareness in collaborative virtual environments. In *VRST '94: Proceedings of the conference on Virtual reality software and technology*, pages 223–236, 1994.

[Bia98] M. Bianchi. Autoauditorium: a fully automatic, multi-camera system to televise auditorium presentations. In *Joint DARPA/NIST Smart Spaces Technology Workshop*, 1998.

[Bia04a] M. Bianchi. Automatic video production of lectures using an intelligent and aware environment. In *MUM '04: Proceedings of the 3rd in-*

ternational conference on Mobile and ubiquitous multimedia, pages 117–123. ACM Press, 2004.

[Bia04b] M. Bianchi. http://www.autoauditorium.com. 2004.

[BKP05] G. Bradski, A. Kaehler, and V. Pisarevsky. Learning-based computer vision with intel's open source computer vision library. *Intel Technology Journal*, 9, 2005.

[BRB07] J. P. Birnholtz, A. Ranjan, and R. Balakrishnan. Using motion tracking data to augment video recordings in experimental social science research. In *E-Social Science*, Michigan, 2007.

[BRB08] J. P. Birnholtz, A. Ranjan, and R. Balakrishnan. Error and coupling: Extending common ground to improve the provision of visual information for collaborative tasks. In *Conference of the International Communication Association*, 2008.

[BSS97] W. A. S. Buxton, A.J. Sellen, and M.C. Sheasby. Interfaces for multiparty videoconferences. *Video-Mediated Communication*, pages 385–400, 1997.

[Bux92] W. A. S. Buxton. Telepresence: integrating shared task and person spaces. In *Graphics Interface*, pages 123–129, 1992.

[Bux95] W. A. S. Buxton. Integrating the periphery and content: A new model of telematics. In *Graphics Interface*, pages 239–246, 1995.

[Bux97] W. A. S. Buxton. Living in augmented reality: Ubiquitous media and reactive environments. *Video-Mediated Communication*, pages 363–384, 1997.

[Bux03] H. Buxton. Learning and understanding dynamic scene activity: a review. *Image Vision Computing*, 21(1):125–136, 2003.

[BW01] M. Brandstein and D. Ward. *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Verlag, 2001.

[CB91] H. H. Clark and S. E. Brennan. Grounding in communication. *L. B. Resnick, R. M. Levine, and S. D. Teasley (Eds.). Perspectives on socially shared cognition*, pages 127–149, 1991.

[CBC+99] J. Coutaz, F. Bérard, E. Carraux, W. Astier, and J. L. Crowley. Comedi: using computer vision to support awareness and privacy in mediaspaces. In *CHI '99: CHI '99 extended abstracts on Human factors in computing systems*, pages 13–14, 1999.

[CFKL92] C. Cool, R. S. Fish, R. E. Kraut, and C. Lowery. Iterative design of video communication systems. In *ACM CSCW*, pages 25–32, 1992.

[Che01] M. Chen. Design of a virtual auditorium. In *ACM Multimedia*, pages 19–28, 2001.

[Che02a] M. Chen. Achieving effective floor control with a low-bandwidth gesture-sensitive videoconferencing system. In *ACM Multimedia*, pages 476–483, 2002.

[Che02b] M. Chen. Leveraging the asymmetric sensitivity of eye contact for videoconference. In *ACM CHI*, pages 49–56, 2002.

[CNM00] S. K. Card, A. Newell, and T. P. Moran. *The Psychology of Human-Computer Interaction*. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, USA, 2000.

[Com01] D.V. Compernolle. Future direction in microphone array processing. In M.S. Brandstein and D. Ward, editors, *Microphone Arrays*, pages 389–394. Springer, 2001.

[Coo95] J. R. Cooperstock. Making the user interface disappear: the reactive room. In *CASCON '95: Proceedings of the 1995 conference of the Centre for Advanced Studies on Collaborative research*, page 15. IBM Press, 1995.

[CRG+02] R. Cutler, Y. Rui, A. Gupta, J. J. Cadiz, I Tashev, L. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg. Distributed meetings: a meeting capture and broadcasting system. In *ACM Multimedia*, pages 503–512, 2002.

[CTB+95] J. R. Cooperstock, K. Tanikoshi, G. Beirne, T. Narine, and W. A. S. Buxton. Evolution of a reactive environment. In *ACM CHI*, pages 170–177, 1995.

[DB92] P. Dourish and S. Bly. Portholes: supporting awareness in a distributed work group. In *ACM CHI*, pages 541–547, 1992.

[DeV03] R.F. DeVellis. *Scale development: theory and applications*. Sage Publications, Thousand Oaks, 2003.

[DGZ92] S. M. Drucker, T. A. Galyean, and D. Zeltzer. Cinema: a system for procedural camera movements. In *SI3D '92: Proceedings of the 1992 symposium on Interactive 3D graphics*, pages 67–70, New York, NY, USA, 1992.

[DJMW98] O. Daly-Jones, A. Monk, and L. Watts. Some advantages of video conferencing over high-quality audio conferencing: fluency and awareness of attentional focus. *Int. J. Hum.-Comput. Stud.*, 49(1):21–58, 1998.

[Dru94] S. M. Drucker. *Intelligent camera control for graphical environments*. PhD thesis, Cambridge, MA, USA, 1994.

[DS00] R. Donald and T. Spann. *Fundamentals of TV Production*. Blackwell Publication, Ames, IA, 2000.

[DV02] D. A. De Vaus. *Survey in Social Research*. Routledge, 2002.

[DZ95] S. M. Drucker and D. Zeltzer. CamDroid: A system for implementing intelligent camera control. In *Symposium on Interactive 3D Graphics*, pages 139–144, 1995.

[Fin97] K. E. Finn. *Video-Mediated Communication*. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, USA, 1997.

[FK00] J. Foote and D. Kimber. Flycam: practical panoramic video. In *ACM Multimedia*, pages 487–488, 2000.

[FKC90] R. S. Fish, R. E. Kraut, and B. L. Chalfonte. The videowindow system in informal communication. In *ACM CSCW*, pages 1–11, 1990.

[FKRR92] R. S. Fish, R. E. Kraut, R. W. Root, and R. E. Rice. Evaluating video as a technology for informal communication. In *ACM CHI*, pages 37–48, 1992.

[FKS00] S. R. Fussell, R. E. Kraut, and J. Siegel. Coordination of communication: effects of shared visual context on collaborative work. In *ACM CSCW*, pages 21–30, 2000.

[FSK03] S. R. Fussell, L. D. Setlock, and R. E. Kraut. Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks. In *ACM CHI*, pages 513–520, 2003.

[FSP03] S. R. Fussell, L. D. Setlock, and E. M. Parker. Where do helpers look?: gaze targets during collaborative physical tasks. In *ACM CHI*, pages 768–769, 2003.

[Fur00] G. Furnas. Future design mindful of the moras. *Human-Computer Interaction*, 15:205–261, 2000.

[G.99] James R. G. Challenges for spoken dialogue systems. In *In Proceedings of 1999 IEEE ASRU Workshop*, 1999.

[Gav92] W. W. Gaver. The affordances of media spaces for collaboration. In *ACM CSCW*, pages 17–24, 1992.

[Ger06] D. Gergle. *The Value of Shared Visual Information for Task-Oriented Collaboration*. PhD thesis, Carnegie Mellon University, 2006.

[GHW02] M. L. Gleicher, R. M. Heck, and M. N Wallick. A framework for virtual videography. In *SMARTGRAPH '02: Proceedings of the 2nd international symposium on Smart graphics*, pages 9–16, New York, NY, USA, 2002.

[GKE90] J. Galegher, R. E. Kraut, and C. Egido, editors. *Intellectual teamwork: social and technological foundations of cooperative work*. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, USA, 1990.

[GKF04] D. Gergle, R. E. Kraut, and S. R. Fussell. Action as language in a shared visual space. In *ACM CSCW*, pages 487–496, 2004.

[GM03] D. M. Grayson and A. F. Monk. Are you looking at me? eye contact and desktop video conferencing. *ACM Trans. Comput.-Hum. Interact.*, 10(3):221–243, 2003.

[GMM+92] W. W. Gaver, T. Moran, A. MacLean, L. Lövstrand, P. Dourish, K. Carter, and W. A. S. Buxton. Realizing a video environment: Europarc's rave system. In *ACM CHI*, pages 27–35, 1992.

[GMR95] H. Gajewska, M. Manasse, and D. Redell. Argohalls: adding support for group awareness to the argo telecollaboration system. In *ACM UIST*, pages 157–158, 1995.

[God60] J. Godard. À bout de souffle, Les Productions Georges de Beauregard and Socit Nouvelle de Cinématographie, 1960.

[GRK07] D. Gergle, C. P. Rose, and R. E. Kraut. Modeling the impact of shared visual information on collaborative reference. In *ACM CHI*, pages 1543–1552, 2007.

[GSHL93] W. W. Gaver, A. Sellen, C. Heath, and P. Luff. One is not enough: multiple views in a media space. In *ACM CHI*, pages 335–341, 1993.

[GSO95] W. W. Gaver, G. Smets, and K. Overbeeke. A virtual window on media space. In *ACM CHI*, pages 257–264, 1995.

[GTHT+06] N. Gupta, G. Tur, D. Hakkani-Tur, S. Bangalore, G. Riccardi, and M. Gilbert. The at and t spoken language understanding system. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(1):213–222, Jan. 2006.

[GYW00] R. Gross, J. Yang, and A. Waibel. Face recognition in a meeting room. *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 294–299, 2000.

[HB02] A. J. Howell and H. Buxton. Visually mediated interaction using learnt gestures and camera control. In *GW '01: Revised Papers from the International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction*, pages 272–284. Springer-Verlag, 2002.

[HFS96] L. W. He, Cohen M. F., and D. H. Salesin. The virtual cinematographer: a paradigm for automatic real-time camera control and directing. In *ACM SIGGRAPH*, pages 217–224, 1996.

[HL91] C. Heath and P. Luff. Disembodied conduct: communication through video in a multi-media office environment. In *ACM CHI*, pages 99–103, 1991.

[HL92] C. Heath and P. Luff. Media spaces and communicative asymmetries: Preliminary observations of video mediated interaction. *Human Computer Interaction*, 7:315–346, 1992.

[HLS97] C. Heath, P. Luff, and A. Sellen. Reconfiguring media space: Supporting collaborative work. *Video-Mediated Communication*, pages 323–347, 1997.

[HS08] Human-Synergistics. The subarctic survival simulation, http://www.human-synergistics.com.au/content/products/simulations/survival.asp, 2008.

[IOM95] T. Inoue, K. Okada, and Y. Matsushita. Learning from tv programs: application of tv presentation to a videoconferencing system. In *ACM UIST*, pages 147–154, 1995.

[IOM96] T. Inoue, K. Okada, and Y. Matsushita. Evaluation of a videoconferencing system based on tv programs. In *IEEE Proceedings of the 19th International Convention of Electrical and Electronics Engineers in Israel*, pages 436–439, 1996.

[IOM97] T. Inoue, K. Okada, and Y. Matsushita. Integration of face-to-face and video-mediated meetings: Hermes. In *ACM SIGGROUP*, pages 405–414, 1997.

[IPKK06] J. Ilonen, P. Paalanen, J.-K. Kamarainen, and H. Kalviainen. Gaussian mixture pdf in one-class classification: computing and utilizing confidence value. In *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, pages 577–580. IEEE Computer Society, 2006.

[ISOM04] A. Inoue, H. Shigeno, K. Okada, and M. Matsushita. Introducing grammar of the film language into automatic shooting for face-to-face meetings. In *2004 International Symposium on Applications and the Internet, 2004*, pages 277–280, 2004.

[IT93] E. A. Isaacs and J. C. Tang. What video can and can't do for collaboration: a case study. In *ACM Multimedia*, pages 199–206, 1993.

[JMFV05] T. Jenkin, J. McGeachie, D. Fono, and R. Vertegaal. eyeView: focus+context views for large group video conferences. In *ACM CHI*, pages 1497–1500, 2005.

[Jon71] P. Jones. *The Technique of The Television Cameraman*. Communication Arts Books, Focal Press Limited, New York, 1971.

[Jou02] N. P. Jouppi. First steps towards mutually-immersive mobile telepresence. In *ACM CSCW*, pages 354–363, 2002.

[KFS03] R. E. Kraut, S. R. Fussell, and J Siegel. Visual information as a conversational resource in collaborative physical tasks. *Human Computer Interaction*, 18(1):13–49, 2003.

[KGF02] R. E. Kraut, D. Gergle, and S. R. Fussell. The use of visual information in shared visual spaces: informing the development of virtual co-presence. In *ACM CSCW*, pages 31–40, 2002.

[KKT94] H. Kuzuoka, T. Kosuge, and M. Tanaka. Gesturecam: a video communication system for sympathetic remote collaboration. In *ACM CSCW*, pages 35–43, 1994.

[KKY+04] H. Kuzuoka, J. Kosaka, K. Yamazaki, Y. Suga, A. Yamazaki, P. Luff, and C. Heath. Mediating dual ecologies. In *ACM CSCW*, pages 477–486, 2004.

[KMS96] R. E. Kraut, M. D. Miller, and J. Siegel. Collaboration in performance of physical tasks: effects on outcomes and communication. In *ACM CSCW*, pages 57–66, 1996.

[KOF06] A. D. I. Kramer, L. M. Oh, and S. R. Fussell. Using linguistic features to measure presence in computer-mediated communication. In *ACM CHI*, pages 913–916, 2006.

[KOY+00] H. Kuzuoka, S. Oyama, K. Yamazaki, K. Suzuki, and M. Mitsuishi. Gestureman: a mobile robot that embodies a remote instructor's actions. In *ACM CSCW*, pages 155–162, 2000.

[KSK+04] T. Kurata, N. Sakata, M. Kourogi, H. Kuzuoka, and M. Billinghurst. Remote collaboration using a shoulder-worn active camera/laser. *ISWC*, 00:62–69, 2004.

[Kun90] J. Kuney. *Take One: Television Directors on Directing*. Praeger Publishers, New York, 1990.

[Kuz92] H. Kuzuoka. Spatial workspace collaboration: a sharedview video support system for remote collaboration capability. In *ACM CHI*, pages 533–540, 1992.

[KZMG85] S. Kiesler, D. Zubrow, A. M. Moses, and V. Geller. Affect in computer-mediated communication: An experiment in synchronous terminal-to-terminal discussion. *Human-Computer Interaction*, 1:77–104, 1985.

[LKF+02] Q. Liu, D. Kimber, J. Foote, L. Wilcox, and J Boreczky. Flyspec: a multi-user video camera system with hybrid human and automatic control. In *ACM Multimedia*, pages 484–492, 2002.

[LLK+03] C. Liao, Q. Liu, D. Kimber, P. Chiu, J. Foote, and L. Wilcox. Shared interactive video for teleconferencing. In *ACM Multimedia*, pages 546–554, 2003.

[Lot07] D. Lottridge. Hedonic affective response as a measure of human performance. Technical report, University of Toronto, 2007.

[LRG+00] M. Lombard, R. D. Reich, M. E. Grabe, C. C. Bracken, and T. B. Ditton. Presence and television. *Human Communication Research*, 26(1):75–98, 2000.

[LRGC01] Q. Liu, Y. Rui, A. Gupta, and J. J. Cadiz. Automating camera management for lecture room environments. In *ACM CHI*, pages 442–449, 2001.

[Lum02] S. Lumet. 12 angry men, Orion-Nova Productions, 2002.

[LZHC07] Z. Liu, Z. Zhang, L. He, and P. Chou. Energy-based sound source localization and gain normalization for ad hoc microphone arrays. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume vol.2, pages II–761–II–764, 2007.

[Man88] M. Mantei. Capturing the capture concepts: a case study in the design of computer-supported meeting environments. In *ACM CSCW*, pages 257–270, 1988.

[McG90] J. E. McGrath. Time matters in groups. *Intellectual teamwork: social and technological foundations of cooperative work*, pages 23–61, 1990.

[MM94] J. C. McCarthy and A. F. Monk. Measuring the quality of computer-mediated communication. *Behaviour and Information Technology*, 13(5):311–319, 1994.

[MR02] E. Machnicki and L. A. Rowe. Virtual director: Automating a webcast. In *Proceedings of the SPIE Multimedia Computing and Networking 2002, Vol. 4673*, 2002.

[MRS+91] M. Mantei, Baecker R.M., A.J. Sellen, W.A.S. Buxton, and T. Milligan. Experiences in the use of media space. In *ACM CHI*, pages 203–208, 1991.

[MS99] S. Mukhopadhyay and B. Smith. Passive capture and structuring of lectures. In *ACM Multimedia*, pages 477–487, 1999.

[MW98] A. F. Monk and L. A. Watts. Peripheral participants in mediated communication. In *ACM CHI*, pages 285–286, 1998.

[NC05] D. Nguyen and J. Canny. Multiview: spatially faithful group video conferencing. In *ACM CHI*, pages 799–808, 2005.

[NKW+95] B. A. Nardi, A. Kuchinsky, S. Whittaker, R. Leichner, and H. Schwarz. Video-as-data: technical and social aspects of a collaborative multimedia application. *Comput. Supported Coop. Work*, 4(1):73–100, 1995.

[NS03] K. Nickel and R. Stiefelhagen. Pointing gesture recognition based on 3-d tracking of face, hands and head orientation. In *International Conference on Multimodal Interfaces*, pages 140–1460. ACM Press, 2003.

[Nun78] J.C. Nunally. *Psychometric theory*. McGraw-Hill, New York, 1978.

[OMIM94] K. Okada, F. Maeda, Y. Ichikawaa, and Y. Matsushita. Multiparty videoconferencing at virtual social distance: Majic design. In *ACM CSCW*, pages 385–393, 1994.

[OO01] G. M. Olson and J. S. Olson. Distance matters. *Human-Computer Interaction*, 15:139–179, 2001. TY - JOUR ID: 58 M1 - Journal.

[OOF+05] J. Ou, L. M. Oh, S. R. Fussell, T. Blum, and J. Yang. Analyzing and predicting focus of attention in remote collaborative tasks. In *ICMI '05: Proceedings of the 7th international conference on Multimodal interfaces*, pages 116–123. ACM Press, 2005.

[OOYF05] J. Ou, L. M. Oh, J. Yang, and S. R. Fussell. Effects of task properties, partner actions, and message content on eye gaze patterns in a collaborative task. In *ACM CHI*, pages 231–240, 2005.

[PC98] E. Paulos and J. Canny. Prop: personal roving presence. In *ACM CHI*, pages 296–303, 1998.

[PE97] S. E. Poltrock and G. Engelbeck. Requirements for a virtual collocation environment. In *ACM SIGGROUP*, pages 61–70, 1997.

[PG05] S. E. Poltrock and J. Grudin. Videoconferencing: Recent experiments and reassessment. *hicss*, 4:104a, 2005.

[Pol08] Polycom. "polycom, http://www.polycom.com/", 2008.

[RBB06] A. Ranjan, J.P. Birnholtz, and Ravin Balakrishnan. An exploratory analysis of partner action and camera control in a video-mediated collaborative task. In *ACM CSCW*, pages 403–412, 2006.

[RBB07] A. Ranjan, J. P. Birnholtz, and R. Balakrishnan. Dynamic shared visual spaces: Experimenting with automatic camera control in a remote repair task. In *ACM CHI*, pages 1177–1186, 2007.

[RBB08] A. Ranjan, J. P. Birnholtz, and R. Balakrishnan. Improving meeting capture by applying television production principles with audio and motion detection. In *ACM CHI*, pages 227–236, 2008.

[RGC01] Y. Rui, A. Gupta, and J. J. Cadiz. Viewing meeting captured by an omni-directional camera. In *ACM CHI*, pages 450–457, 2001.

[RGG03] Y. Rui, A. Gupta, and J. Grudin. Videography for telepresentations. In *ACM CHI*, pages 457–464, 2003.

[RHGL01] Y. Rui, L. He, A. Gupta, and Q. Liu. Building an intelligent camera management system. In *ACM Multimedia*, pages 2–11, 2001.

[Ros99] B. Rose. *Directing for Television: Conversations with American TV Directors*. Scarecrow Press, 1999.

[Rub02] A. M. Rubin. The uses-and-gratifications perspective of media effects. *Media Effects: Advances in theory and persuasion*, pages 525–548, 2002.

[SB98] R. S. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 1998.

[Sch00] D. W. Schloerb. A quantitative measure of telepresence. *Presence: Teleoperators and Virtual Environments*, 4(1):64–80, 2000.

[Sel92] A. J. Sellen. Speech patterns in video-mediated conversations. In *ACM CHI*, pages 49–59, 1992.

[SG00] J. Sherrah and S. Gong. Vigour: A system for tracking and recognition of multiple people and their activities. *ICPR*, 01:1179, 2000.

[SGHB00] J. Sherrah, S. Gong, A. J. Howell, and H. Buxton. Interpretation of group behavior in visually mediated interaction. *icpr*, 01:1266, 2000.

[Sim96] H. Simon. *The Sciences of the Artificial*. MIT Press, Cambridge, MA, 1996.

[SZ02] R. Stiefelhagen and J. Zhu. Head orientation and gaze direction in meetings. In *CHI '02: CHI '02 extended abstracts on Human factors in computing systems*, pages 858–859, 2002.

[TBN00] B. Tomlinson, B. Blumberg, and D. Nain. Expressive autonomous cinematography for interactive virtual environments. In *AGENTS '00: Proceedings of the fourth international conference on Autonomous agents*, pages 317–324, New York, NY, USA, 2000.

[TM90] J. C. Tang and S. L. Minneman. Videodraw: a video interface for collaborative drawing. In *ACM CHI*, pages 313–320, 1990.

[TOM03] Y. Takemae, K. Otsuka, and N. Mukawa. Video cut editing rule based on participants' gaze in multiparty conversation. In *ACM Multimedia*, pages 303–306, 2003.

[TOM04] Y. Takemae, K. Otsuka, and N. Mukawa. Impact of video editing based on participants' gaze in multiparty conversation. In *ACM CHI*, pages 1333–1336, 2004.

[TOY05] Y. Takemae, K. Otsuka, and J. Yamato. Automatic video editing system using stereo-based head tracking for multiparty conversation. In *ACM CHI*, pages 1817–1820, 2005.

[Ver99] Verizon. Meetings in america: a study of trends, costs, and attitudes toward business travel and teleconferencing, and their impact on productivity, http://emeetings.verizonbusiness.com/meetingsinamerica/uswhitepaper.php, 1999.

[Ver02] G. Verbinski. The ring, Dreamworks SKG Productions, MacDonald/Parkes Productions, and Benderspink, 2002.

[Ver03] Verizon. Meetings in america v: Meeting of the minds (mci). http://emeetings.verizonbusiness.com/meetingsinamerica/pdf/mia5.pdf, 2003.

[Vic08] Vicon. http://www.vicon.com/, 2008.

[VJ04] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, 2004.

[VOOF99] E. S. Veinott, J. Olson, G. M. Olson, and X. Fu. Video helps remote work: speakers who need to negotiate common ground benefit from seeing each other. In *ACM CHI*, pages 302–309, 1999.

[VWSC03] R. Vertegaal, I. Weevers, C. Sohn, and C. Cheung. Gaze-2: conveying eye contact in group video conferencing using eye-controlled camera direction. In *ACM CHI*, pages 521–528, 2003.

[Wei99] K. Weick. Organizing for high reliability: Processes of collective mindfulness. *Research in organizational behavior*, 21:81–123, 1999.

[Whi03] S. Whittaker. Theories and methods in mediated communication. *The Handbook of Discourse Processes*, 2003.

[WHK92] Y. Wakita, S. Hirai, and T. Kino. Automatic camera-work control for intelligent monitoring of telerobotic tasks. In *Proceedings of the 1992 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1130–1135, 1992.

[WO97] S. Whittaker and B. O'Conaill. The role of vision in face-to-face and mediated communication. *Video-Mediated Communication*, pages 23–49, 1997.

[WS98] B. G. Witmer and M. J. Singer. Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments*, 7(3):225–240, 1998.

[YCNB96] K. Yamaashi, J. R. Cooperstock, T. Narine, and W. A. S. Buxton. Beating the limitations of camera-monitor mediated telepresence with extra eyes. In *ACM CHI*, pages 50–57, 1996.

[YKA02] M. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Transactions on Pattern Aanalysis and Machine Intelligence*, 24(2):34–58, 2002.

[Zet05] H. Zettl. *Television Production Handbook*. Wadsworth Publishing, Belmont, CA, 2005.

Appendix A

Consent forms

This appendix includes the consent form used in the studies described in the dissertation.

CONSENT FORM

I agree to participate in a study that is comparing the effectiveness of various camera control techniques in video conferencing systems. I understand that my participation is entirely voluntary.

The following points have been explained to me:

1. The purpose of this research is to compare the effectiveness of different camera control techniques in the completion of certain types of tasks. I understand I will be asked questions about my previous computer experience, whether or not I have any prior relationship with the other subject I am paired with in this experiment, and about my satisfaction with different systems used. The primary benefits I may expect from the study are: (a) an appreciation of research on user interfaces, (b) an opportunity to contribute to scientific research.

2. **I will receive $10 for my participation in this research.**

3. The procedure will be as follows: During a single session lasting approximately 1 hour, I will perform various experimental tasks using a video conferencing system.

4. The researchers do not foresee any risks to me for participating in this study, nor do they expect that I will experience any discomfort or stress.

5. I understand that I may withdraw from the study at any time.

6. I understand that I will receive a copy of this consent form.

7. All of the data collected will remain strictly confidential. Only people associated with the study will see my responses. My responses will not be associated with my name; instead, my name will be converted to a code number when the researchers store the data.

8. The experimenter will answer any other questions about the research either now or during the course of the experiment. If I have any other questions or concerns, I can address them to the research director**,** Prof. Ravin Balakrishnan of the Department of Computer Science. He can be contacted by phone: 416-978-5359 or email: ravin@cs.toronto.edu. Directions to his office can be found on his website: www.dgp.toronto.edu/~ravin

9. Upon completion of my participation, I will receive an explanation about the rationale and predictions underlying this experiment.

_____	_____	_____
Participant's Printed Name	Participant's Signature	Date

_____	_____
Experimenter Name	Participant Number

Figure A.1: Consent form.

Appendix B

Questionnaires

This chapter of the appendix includes all the questionnaires used in the studies described in the dissertation.

Pre-Experiment Questionnaire

Please provide your name and contact information so that we may contact you in the event that you are eligible for one of the gift certificate prizes:

Name: _____

Email: _____

Phone: _____

Please check the appropriate spaces or fill in the requested information:

1. Your age: _____
2. Dominant Hand: Right___ Left___
3. Year in school/university: 1st___ 2nd___ 3rd___ 4th___ Other:____
4. Your native/primary language: _____
5. Your second language (if applicable): _____
6. Over the past 12 months, about how frequently have you used videoconferencing to communicate with others (e.g., via a home webcam, meeting room conferencing, etc.)?

   Never | Once or Twice | A few times | Once a month | Once a week or more

7. Over the past 12 months, about how frequently have you built objects or played with Lego plastic bricks?

   Never | Once or Twice | A few times | Once a month | Once a week or more

8. As a child, about how frequently did you build objects or play with Lego plastic bricks?

   Never | Once or Twice | A few times | Once a month | Once a week or more

Figure B.1: Pre-Questionnaire (Chapter 4).

Subject ID #_____

Post-Experiment Questionnaire

Please indicate the extent of your agreement with the following statements by circling one of the options:

1. My partner and I completed the Lego construction tasks effectively.

   Strongly Disagree Disagree Neutral Agree Strongly Agree

2. It was easy for me to do my part of the tasks.

   Strongly Disagree Disagree Neutral Agree Strongly Agree

3. I had trouble understanding what my partner wanted.

   Strongly Disagree Disagree Neutral Agree Strongly Agree

4. It was easy for me to do my part of the task.

   Strongly Disagree Disagree Neutral Agree Strongly Agree

5. It was easy to see what my partner was doing.

   Strongly Disagree Disagree Neutral Agree Strongly Agree

6. It was difficult to hear my partner.

   Strongly Disagree Disagree Neutral Agree Strongly Agree

7. I had too much to do in completing these tasks.

   Strongly Disagree Disagree Neutral Agree Strongly Agree

We are interested in your experience with the videoconferencing system being used here. Are there any specific improvements you would make that you think would make the task easier?

Figure B.2: Post-Questionnaire (Chapter 4).

Helper Questionnaire – Condition 1

Subject ID: _____

	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
My partner and I completed this task effectively	1	2	3	4	5	6	7
My partner and I completed this task faster than most people could.	1	2	3	4	5	6	7
My partner and I communicated well in completing this task.	1	2	3	4	5	6	7
This task would have gone more smoothly if my partner and I were in the same place.	1	2	3	4	5	6	7
I was effective in doing what I needed to do for my partner and I to complete this task.	1	2	3	4	5	6	7
I performed my role in this task better than most people could.	1	2	3	4	5	6	7
I was happy with my performance in this task.	1	2	3	4	5	6	7
It was easy to do my part of this task.	1	2	3	4	5	6	7
My partner was effective in doing what he/she needed to do for us to complete this task..	1	2	3	4	5	6	7
I was able to examine objects in great detail.	1	2	3	4	5	6	7
I was able to tell all of the Lego pieces apart by looking at the video screen.	1	2	3	4	5	6	7
I relied primarily on the video view, and not conversation with my partner, to tell pieces apart.	1	2	3	4	5	6	7
Most of the time, I saw exactly what I wanted on the video screen.	1	2	3	4	5	6	7
I could usually tell what my partner was doing.	1	2	3	4	5	6	7

Figure B.3: Page 1 of Helper questionnaire (Static camera condition in Chapter 5).

	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
I had no trouble seeing what I needed to see in completing this task.	1	2	3	4	5	6	7
I usually knew where in the workspace my partner was working	1	2	3	4	5	6	7
It was hard to tell where in the workspace my partner was working	1	2	3	4	5	6	7
There were times when I had no idea what my partner was doing.	1	2	3	4	5	6	7
I adjusted quickly to the process of working with this system.	1	2	3	4	5	6	7
I felt much more comfortable with this system at the end than at the start	1	2	3	4	5	6	7
I had to work hard to learn how to work with this system	1	2	3	4	5	6	7
It was frustrating to learn how to use this system	1	2	3	4	5	6	7
This system made it hard for me to do this task.	1	2	3	4	5	6	7
	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree

Figure B.4: Page 2 of Helper questionnaire (Static camera condition in Chapter 5).

Helper Questionnaire – Condition 2

Subject ID: _____

	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
My partner and I completed this task effectively	1	2	3	4	5	6	7
My partner and I completed this task faster than most people could.	1	2	3	4	5	6	7
My partner and I communicated well in completing this task.	1	2	3	4	5	6	7
This task would have gone more smoothly if my partner and I were in the same place.	1	2	3	4	5	6	7
I was effective in doing what I needed to do for my partner and I to complete this task.	1	2	3	4	5	6	7
I performed my role in this task better than most people could.	1	2	3	4	5	6	7
I was happy with my performance in this task.	1	2	3	4	5	6	7
It was easy to do my part of this task.	1	2	3	4	5	6	7
My partner was effective in doing what he/she needed to do for us to complete this task..	1	2	3	4	5	6	7
I was able to examine objects in great detail.	1	2	3	4	5	6	7
I was able to tell all of the Lego pieces apart by looking at the video screen.	1	2	3	4	5	6	7
I relied primarily on the video view, and not conversation with my partner, to tell pieces apart.	1	2	3	4	5	6	7
Most of the time, I saw exactly what I wanted on the video screen.	1	2	3	4	5	6	7
I could usually tell what my partner was doing.	1	2	3	4	5	6	7

Figure B.5: Page 1 of Helper questionnaire (Dynamic camera condition in Chapter 5).

	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
I had no trouble seeing what I needed to see in completing this task.	1	2	3	4	5	6	7
I usually knew where in the workspace my partner was working	1	2	3	4	5	6	7
I usually knew when the camera was going to change shots	1	2	3	4	5	6	7
It was frustrating sometimes when the camera changed shots	1	2	3	4	5	6	7
When the camera changed shots, it usually changed to something I wanted to see	1	2	3	4	5	6	7
I had no idea when the camera was going to change shots	1	2	3	4	5	6	7
It was hard to tell where in the workspace my partner was working	1	2	3	4	5	6	7
There were times when I had no idea what my partner was doing.	1	2	3	4	5	6	7
I adjusted quickly to the process of working with this system.	1	2	3	4	5	6	7
I felt much more comfortable with this system at the end than at the start	1	2	3	4	5	6	7
I had to work hard to learn how to work with this system	1	2	3	4	5	6	7
It was frustrating to learn how to use this system	1	2	3	4	5	6	7
This system made it hard for me to do this task.	1	2	3	4	5	6	7
	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree

Figure B.6: Page 2 of Helper questionnaire (Dynamic camera condition in Chapter 5).

Worker Questionnaire – Condition 1

Subject ID: _____

	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
My partner and I completed this task effectively.	1	2	3	4	5	6	7
My partner and I completed this task faster than most people could.	1	2	3	4	5	6	7
My partner and I communicated well in completing this task..	1	2	3	4	5	6	7
This task would have gone more smoothly if my partner and I were in the same place.	1	2	3	4	5	6	7
I was effective in doing what I needed to do for my partner and I to complete this task..	1	2	3	4	5	6	7
I performed my role in this task better than most people could.	1	2	3	4	5	6	7
I was happy with my performance in this task.	1	2	3	4	5	6	7
My partner was able to examine objects in great detail.	1	2	3	4	5	6	7
My partner was effective in doing what he/she needed to do for us to complete this task..	1	2	3	4	5	6	7
It was easy to do my part of this task.	1	2	3	4	5	6	7
My partner was able to tell all of the pieces apart by looking at the video screen.	1	2	3	4	5	6	7
My partner relied primarily on the video view, and not on our conversation, to tell pieces apart.	1	2	3	4	5	6	7
My partner saw exactly what he/she wanted on the video screen most of the time.	1	2	3	4	5	6	7

Figure B.7: Page 1 of Worker questionnaire (Static camera condition in Chapter 5).

	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
My partner usually knew exactly what I was doing.	1	2	3	4	5	6	7
My partner had no trouble seeing what he/she needed to see in completing this task..	1	2	3	4	5	6	7
My partner usually knew where in the workspace I was working	1	2	3	4	5	6	7
It was hard for my partner to tell where in the workspace I was working	1	2	3	4	5	6	7
There were times when my partner had no idea what I was doing.	1	2	3	4	5	6	7
I adjusted quickly to the process of working with this system.	1	2	3	4	5	6	7
I felt much more comfortable with this system at the end than at the start	1	2	3	4	5	6	7
I had to work hard to learn how to work with this system	1	2	3	4	5	6	7
It was frustrating to learn how to use this system	1	2	3	4	5	6	7
This system made it hard for me to do this task.	1	2	3	4	5	6	7

Figure B.8: Page 2 of Worker questionnaire (Static camera condition in Chapter 5).

Worker Questionnaire – Condition 2

Subject ID: _____

	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
My partner and I completed this task effectively.	1	2	3	4	5	6	7
My partner and I completed this task faster than most people could.	1	2	3	4	5	6	7
My partner and I communicated well in completing this task..	1	2	3	4	5	6	7
This task would have gone more smoothly if my partner and I were in the same place.	1	2	3	4	5	6	7
I was effective in doing what I needed to do for my partner and I to complete this task..	1	2	3	4	5	6	7
I performed my role in this task better than most people could.	1	2	3	4	5	6	7
I was happy with my performance in this task.	1	2	3	4	5	6	7
My partner was able to examine objects in great detail.	1	2	3	4	5	6	7
My partner was effective in doing what he/she needed to do for us to complete this task..	1	2	3	4	5	6	7
It was easy to do my part of this task.	1	2	3	4	5	6	7
My partner was able to tell all of the pieces apart by looking at the video screen.	1	2	3	4	5	6	7
My partner relied primarily on the video view, and not on our conversation, to tell pieces apart.	1	2	3	4	5	6	7
My partner saw exactly what he/she wanted on the video screen most of the time.	1	2	3	4	5	6	7

Figure B.9: Page 1 of Worker questionnaire (Dynamic camera condition in Chapter 5).

	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
My partner usually knew exactly what I was doing.	1	2	3	4	5	6	7
My partner had no trouble seeing what he/she needed to see in completing this task..	1	2	3	4	5	6	7
My partner usually knew where in the workspace I was working	1	2	3	4	5	6	7
I usually knew when the camera was going to change shots	1	2	3	4	5	6	7
It was frustrating sometimes when the camera changed shots	1	2	3	4	5	6	7
When the camera changed shots, it usually changed to something my partner wanted to see.	1	2	3	4	5	6	7
I had no idea when the camera was going to change shots.	1	2	3	4	5	6	7
It was hard for my partner to tell where in the workspace I was working	1	2	3	4	5	6	7
There were times when my partner had no idea what I was doing.	1	2	3	4	5	6	7
I adjusted quickly to the process of working with this system.	1	2	3	4	5	6	7
I felt much more comfortable with this system at the end than at the start	1	2	3	4	5	6	7
I had to work hard to learn how to work with this system	1	2	3	4	5	6	7
It was frustrating to learn how to use this system	1	2	3	4	5	6	7
This system made it hard for me to do this task.	1	2	3	4	5	6	7

Figure B.10: Page 2 of Worker questionnaire (Dynamic camera condition in Chapter 5).

Video Recording Feedback (Administered half-way through the video)

1. Do you think any of the participants spent more or less time talking than the others? If so, which one(s)?

2. Who do you think is the most influential participant so far? Why?

3. How often did you wish you could have seen something, but could not. (please mark an 'X' beside your selection)

a) Never
b) Hardly ever
c) Sometimes
d) Often
e) Always

4. How often did you see something, but wondered why you were seeing it or wished you could see something else? (please mark an 'X' beside your selection)

a) Never
b) Hardly ever
c) Sometimes
d) Often
e) Always

5. Indicate the extent to which it was easy for you to tell who was talking to whom. (please mark an 'X' beside your selection)

a) Very easy
b) Easy
c) Neither easy, nor difficult
d) Difficult
e) Very difficult

Figure B.11: Video evaluation questionnaire (Chapter 6).

Post-Questionnaire

1. If you had to find three main differences in the quality of video presentation between the first video and the second video, what would those be? Please consider the way the video was presented, and ignore the discussion content difference and video/audio artifacts.

   i. First difference:

   ii. Second difference:

   iii. Third difference:

2. Based on the quality of video presentation, how do you think the first video was shot?

   i. A professional TV production crew
   ii. A novice crew
   iii. Automatically using computers
   iv. Other

3. How do you think the second video was shot?

   i. A professional TV production crew
   ii. A novice crew
   iii. Automatically using computers
   iv. Other

Figure B.12: Post-questionnaire (Chapter 6).

Appendix C

Arctic survival task scenario

This chapter of the appendix includes the Arctic survival task scenario used in the study described in Chapter 6.

Instructions

1. The Situation

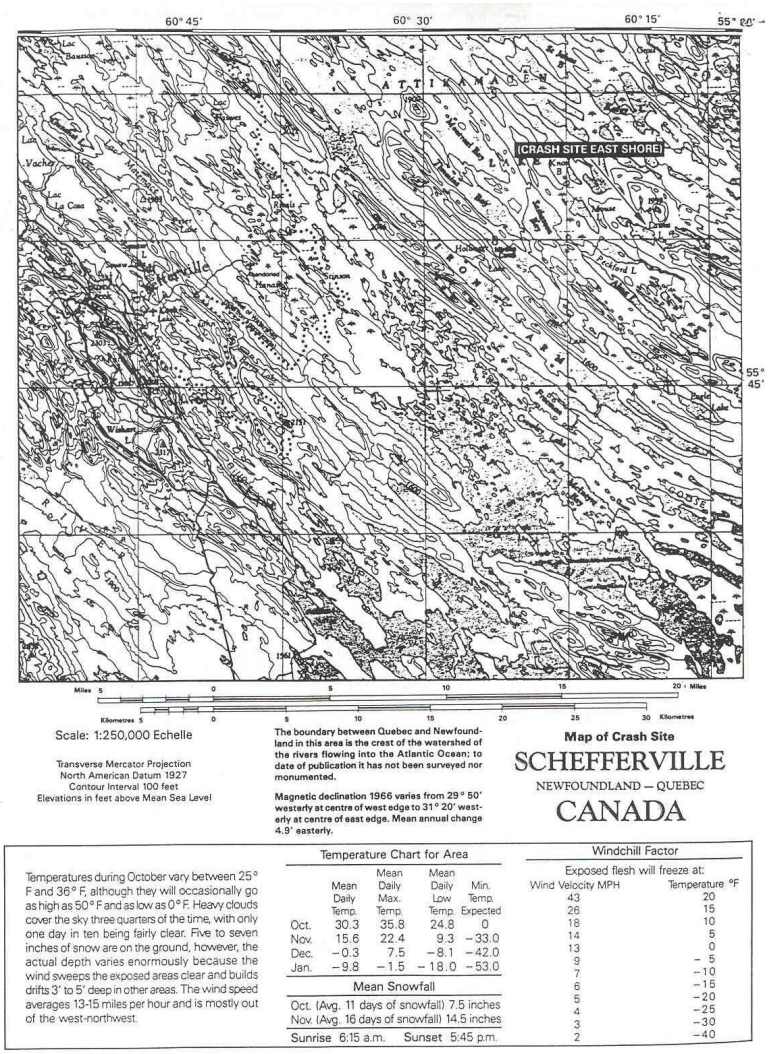
It is approximately 2:30pm, October 5 and you have just crash-landed in a float plane on the east shore of Laura Lake in the sub arctic region of the northern Quebec-Newfoundland border. The pilot was killed in the crash, but the rest of you are uninjured. Each of you is wet up to the waist and perspiring heavily. Shortly after the crash, the plane drifted into deep water and sank with the pilot's body pinned inside.

The pilot was unable to contact anyone before the crash. However, ground sightings indicated that you are 30 miles south of your intended course and approximately 22 air miles east of Schefferville, your original destination and the nearest known habitation. (The mining camp on Hollinger Lake was abandoned years ago when a fire destroyed the buildings.) Schefferville (pop. 5,000) is an iron ore mining town approximately 200 air miles north of the St. Lawrence, 450 miles east of the James Bay/Hudson Bay area, 800 miles south of the Arctic Circle, and 300 miles west of the Atlantic coast. It is reachable only by air or rail, all roads ending a few miles from town. Your party was expected to return from northwestern Labrador to Schefferville no later than October 19 and filled a Flight Notification Form to that effect with the Department of Transportation via Schefferville radio.

The immediate area is covered small evergreen trees(1 ½ to 4 inches in diameter). Scattered in the area are a number of hills with rocky and barren tops. Tundra (arctic swamps) make up the valleys between the hills and consist only of small scrubs. Approximately 25 percent of the region is covered by long, narrow lakes which run northwest to southeast. Innumerable streams and rivers flow into and connect the lakes.

You are all dressed in insulated underwear, socks, heavy wool shirts, pants, knit gloves, sheepskin jackets, knitted wool caps, and heavy leather hunting boots. Collectively, your group possessions include: $152 in bills and 2 half dollars, 4 quarters, 2 dimes, 1 nickel and 3 new pennies; 1 pocket knife (2 blades and an awl which resembles an ice pick): one stub lead pencil; and an air map (shown on opposite page).

Figure C.1: Arctic survival task scenario (Page 1).



Temperatures during October vary between 25° F and 36° F, although they will occasionally go as high as 50° F and as low as 0° F. Heavy clouds cover the sky three quarters of the time, with only one day in ten being fairly clear. Five to seven inches of snow are on the ground, however, the actual depth varies enormously because the wind sweeps the exposed areas clear and builds drifts 3' to 5' deep in other areas. The wind speed averages 13-15 miles per hour and is mostly out of the west-northwest.

Temperature Chart for Area

	Mean Daily Temp.	Mean Daily Max. Temp.	Mean Daily Low Temp.	Min. Temp. Expected
Oct.	30.3	35.8	24.8	0
Nov.	15.6	22.4	9.3	−33.0
Dec.	−0.3	7.5	−8.1	−42.0
Jan.	−9.8	−1.5	−18.0	−53.0

Mean Snowfall

Oct. (Avg. 11 days of snowfall) 7.5 inches
Nov. (Avg. 16 days of snowfall) 14.5 inches

Sunrise 6:15 a.m. Sunset 5:45 p.m.

Windchill Factor

Exposed flesh will freeze at:	
Wind Velocity MPH	Temperature °F
43	20
26	15
18	10
14	5
13	0
9	− 5
7	−10
6	−15
5	−20
4	−25
3	−30
2	−40

Figure C.2: Arctic survival task scenario (Page 2).

2. Your Task

Before the plane drifted away and sank, you and your partners were able to salvage the 15 items listed on the next page, Your task is to rank these items according to their importance to your survival, starting with “1” as the most important, to “15” as the least important.

You may assume ---

1. the number of survivors is just 3: you and your partners;
2. you are the actual people in the situation;
3. you and your partners have agreed to stick together;
4. all items are dry and in good condition.

Step 1: You are to individually rank each item.

Figure C.3: Arctic survival task scenario (Page 3).

Step 1. Individual ranking.

The following table provides 15 items. You are required to ***individually*** rank all the items according to their importance to survival in the very area.

Items	**Individual ranking**
A magnetic compass	
A gallon can of maple syrup	
A sleeping bag per person (arctic type down filled with liner)	
A bottle of water purification tablets	
A 20' x 20' piece of heavy duty canvas	
13 wood matches in a metal screw top waterproof container	
250 ft. of ¼ inch braided nylon rope, 50 lb test	
An operating 4 battery flashlight	
3 pairs of snowshoes	
A fifth Bacardi rum (151 proof)	
Safety razor shaving kit with mirror	
A wind-up alarm clock	
A hand axe	
One aircraft inner tube for a 14 inch wheel (punctured)	
A book entitled, Northern Star Navigation	

Figure C.4: Arctic survival task scenario (Page 4).

Step 2. Team ranking.

First, please copy your individual ranking in *Step1* to the table below. Once discussion begins, do not change your individual ranking.

You are now to discuss this with your partner, who happens to be far from you, but you can reach him by video. You will have up to 40 minutes to negotiate a best solution.

Items	**Your Individual Ranking**	**Team Ranking**
A magnetic compass		
A gallon can of maple syrup		
A sleeping bag per person (arctic type down filled with liner)		
A bottle of water purification tablets		
A 20' x 20' piece of heavy duty canvas		
13 wood matches in a metal screw top waterproof container		
250 ft. of ¼ inch braided nylon rope, 50 lb test		
An operating 4 battery flashlight		
3 pairs of snowshoes		
A fifth Bacardi rum (151 proof)		
Safety razor shaving kit with mirror		
A wind-up alarm clock		
A hand axe		
One aircraft inner tube for a 14 inch wheel (punctured)		
A book entitled, Northern Star Navigation		

Figure C.5: Arctic survival task scenario (Page 5).

Step 3. Select a Team Leader

If one of you should be a leader, select a leader _____(name one of you only).

Figure C.6: Arctic survival task scenario (Page 6).

Appendix D

Brief biographies of collaborating professional directors

D.1 Jeremy Birnholtz

Jeremy has an undergraduate degree in Radio/TV/Film production from Northwestern University. While at Northwestern, he took a variety of television production courses, including one on studio directing. He also spent four years, two of them as a director, working in the journalism's school's newscast studio, in which students produced professional-caliber news programs in a 3-camera studio. Prior to all of this, Jeremy was actively involved with his high school's cable TV channel for 2.5 years, serving as a crew member and/or director for a range of studio programs, sporting events and live coverage of school board meetings.

D.2 Dana Lee

Dana Lee has 30 years experience in television production, technical operations and university teaching, and has been involved in research, design and development, both technical and aesthetic. During his 17 year tenure with Chum Television as operations supervisor of MuchMusic, he helped research and design the original Much facility within the ChumCity Building on Queen Street West in Toronto, optimizing its layout for maximum flexibility and functionality within an eclectic shooting environment. Dana M. Lee is now an assistant professor, teaching technical theory and practicum in the Radio and Television Arts program at Ryerson University. He has also written a comprehensive technical training textbook for the RTA program (which is published on the Web) and is used by operations departments in television facilities and teaching facilities worldwide.