Viewing Meetings Captured by an Omni-Directional Camera

Yong Rui, Anoop Gupta and JJ Cadiz

September 21, 2000

Technical Report
MSR-TR-2000-97

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

Viewing Meetings Captured by an Omni-Directional Camera

Yong Rui, Anoop Gupta and JJ Cadiz

Collaboration and Multimedia Systems Group, Microsoft Research

One Microsoft Way

Redmond, WA 98052-6399

{yongrui, anoop, jjcadiz}@microsoft.com

ABSTRACT

One vision of future technology is the ability to easily and inexpensively capture any group meeting that occurs, store it, and make it available for people to view anytime and anywhere on the network. One barrier to achieving this vision has been the design of low-cost camera systems that can capture all important aspects of the meeting without needing a human camera operator. A promising solution that has emerged recently is omni-directional cameras that can capture a 360-degree video of the entire meeting.

The panoramic capability provided by these cameras raises both new opportunities and new issues for the interfaces provided for post-meeting viewers – for example, do we show all meeting participants all the time or do we just show the person who is speaking, how much control do we provide to the end-user in selecting the view, and will providing this control distract them from their task. These are not just user-interface issues, they also raise tradeoffs for the client-server systems used to deliver such content. They impact how much data needs to be stored on the disk, what computation can be done on the server vs. the client, and how much bandwidth is needed. We report on a prototype system built using an omni-directional camera and results from user studies of interface preferences expressed by viewers.

Keywords

Omni-directional camera systems, On-demand meeting watching

1 INTRODUCTION

In corporate and university environments today, capture of audio-video of presentations for subsequent online viewing has become commonplace. As examples, He *et. al.* [8] report on widespread use within Microsoft, and numerous universities are making their courses available online [21, 24]. The online recordings provide benefits of anytime anywhere viewing and potential for time-saving as only relevant portions of the presentation may be watched [8,9].

While online presentations are becoming commonplace, the same is not true for audio-video recording of group meetings in the workplace. The reason in large part is that the cost-benefit economics surrounding meetings is different. The benefits, of course, are many – we can go back and review meetings for critical decisions and memory jogging (e.g., as used at Xerox [14]), we can catch-up with happenings if we had to miss a meeting due to travel or other reasons, and may be most importantly, we can save time by skipping meetings of limited relevance to us and just browsing them online later [8,9,17,22]. On the cost side, first there is the overhead of planning that we intend to have the meeting recorded. Second, there is a significant cost in recruiting a camera operator to come and video tape the meeting and then post it online. On the social side, the presence of a camera operator in the group meetings can also perturb the group dynamics.

While these costs are substantial today, emerging technological advances will significantly lower the costs, making benefits exceed costs. We believe that in the future recording a meeting will become almost as simple as turning ON the light switch in the meeting room, and the recurring cost will be negligible (few dollars for the disk storage for a one-hour meeting). In this paper, we explore design, user-interface, and related system-tradeoff issues for one such prototype system, targeting small group meetings.

We report on an autonomous meeting capture system built using emerging omni-directional cameras. We use the latest generation of these cameras that can capture the 360-degree view at high-resolution of 1300x1030 pixels and 11 frames per second. We have built the image processing software that can locate where the people are in this 360-degree field, and can extract and frame a person in a rectangular video window appropriately.

This omni-directional camera system allows us to easily explore various user interface choices. For example, it provides all meeting participants' videos simultaneously which only multiple conventional cameras can provide. It also provides a panoramic overview of the entire meeting site almost effortlessly (top portion of Figure 5). Our primary focus in this paper is to study new opportunities and new issues raised by the new system for capturing small group meetings. Specifically, we study the interfaces used to present the captured meeting video to online users, and associated systems tradeoffs, including:

- View of meeting participants: Do users prefer to view all participants all the time or do they prefer to view just a single "active" person (say the person speaking) appropriately framed.
- Amount of user involvement: Do users like to control whom they want to see during the meeting or would they rather let the computer choose the camera shots?
- Rules for camera control: If users prefer that the computer control the camera, what are some desirable

rules the computer should follow to choose camera shots?

- Providing meeting context: Can a 360-degree view of a meeting be used to provide users with added context about a meeting?

The rest of the paper is organized as follows. First, we will discuss related work in Section 2. In Section 3, we present the detailed design of our omni-directional camera system, including both hardware setup and software development, for capturing small group meetings. In Section 4, we describe five carefully chosen interfaces to study the questions we raised earlier. In Sections 5 and 6, we present the user study methods and results for the five interfaces. We discuss important findings of the user study in Section 7. In Section 8, we give concluding remarks and future research directions.

2 RELATED WORK

As we will elaborate, the majority of commercial and research systems in this area have focused on remote teleconferencing environments, where the meeting is "live" and majority of participants remote. This makes the user-interface requirements somewhat different than our focus here, where we are targeting a person watching a locally held meeting at a later time.

2.1 Commercial Teleconferencing Systems

Today a large variety of teleconferencing systems are available commercially from PictureTel, PolyCom, CUSeeMe, Sony, V-Tel, and so on. Given similarity of many of these products we focus on PictureTel's systems.

PictureTel's products come in both a personal system version (e.g., PictureTel 550) and a group system version (e.g., PictureTel 900) [18]. For PictureTel's personal system, Microsoft NetMeeting is used as the interface. NetMeeting provides a picture-in-picture view (large video of the remote person, and small of the local person) that is nice for the "live" conference.

For their group system, PictureTel provides a controllable pan-tilt-zoom camera and a microphone array, which is either built into the camera or placed elsewhere in the meeting room. Because human voice reaches different microphones at slightly different time, the microphone array can determine the position of the sound source. PictureTel's group systems use TV-based interfaces. PictureTel has developed an "Enhanced Continuous Presence" technique, which allows remote users to display multiple meeting sites on the screen at the same time [19]. Remote users can choose six different layouts to best meet the needs of their meetings. The six layouts are: full screen, 2-way (side-by-side), 2-way (above/below), 4-way quad, 1+5 (1 large window and 5 smaller windows), and 9-way. Out of the six layouts, the 2-way, 4-way, 9-way are similar to our "all-up" interface and the 1+5 interface is similar to our "user-controlled + overview" interface that we describe later. For both PictureTel 550 and 900, they also support meeting recording and on-demand viewing, where they use the same interface as the "live" situation.

2.2 Research Systems

Buxton, Sellen, and Sheasby present an excellent overview of interface issues for multiparty videoconferences in the book *Video Mediated Communication* [3]. The book chapter brings together systems and research effort presented in earlier conferences, including Hydra, LiveWire, Portholes and BradyBunch [3,20]. They explore interfaces from the perspective of establishing eye contact, awareness of others and who is attending to whom, parallel conversations and ability to hold side conversations, perception of group as a whole, and ability to deal with shared documents and artifacts.

Many of the interfaces we examine in this paper are common with their work, though there are differences in detail either because of hardware configuration (e.g., omni-directional camera) or choice of parameters – e.g., in our voice-activated video window we have an explicit rule that does not allow the camera to switch too often, something that bothered their subjects [3,20]. Also, since our focus is on on-demand review of captured meetings rather than remote participation in live meetings, the issues faced by users are quite different. For example, while gaze awareness and ability to have side conversations are very important for a "live" meeting [3], clearly these are not an issue for on-demand viewing.

Vic-Vat is a network-based tele-conferencing system developed at UC Berkeley [13]. Its interface displays multiple participants as thumbnail videos. If remote users are interested in any of the thumbnail videos, they can click and open a bigger video window. In a later version of Vic-Vat [25], the interface would cycle through different images based either on a timer or on which participant was talking. Again, their research focus is on "live" meeting.

Finally, there is a large literature in video mediated communication for live meetings [2,6,20,23]. However, given its loose relation to this work, we do not elaborate on it here.

2.3 Omni-Directional Cameras

Recent technology advances in omni-directional vision sensors have inspired many researchers to rethink the way images are captured and analyzed [5]. The applications of omni-directional camera span the spectrum of 3D-reconstruction, visualization, surveillance and navigation in the computer vision research community [5,11,15,26]. Omni-directional cameras have also found their way to the consumer market. BeHere Corporation [1] provides 360° Internet video technology in entertainment, news and sports webcasts. With its interface, remote users can control personalized 360° camera angles independent of other viewers to gain a "be here" experience. The system is however designed with a low-resolution camera system (~

¼ of system used here) and the interface is not targeted for meetings.

Omni-directional camera systems most related to our work are described in [10,16,22]. However, in these systems, the omni-directional camera technology is used to determine where participants are located, and then a conventional camera is used to zoom in on the participant. That is, their systems use the omni-directional camera for monitoring but separate conventional cameras for capturing while ours monitors meeting participants and captures the meeting at the same time. Their research also does not explore the user interface options evaluated here or report user study results.

To summarize, substantial research has been done on real-time teleconferencing systems, including system architecture, interface design and capturing devices. We complement this previous research by examining user interface issues and systems implications, focusing on the perspective of people watching pre-recorded small group meetings and exploiting emerging 360-degree omni-directional cameras.

3 MEETING CAPTURE ENVIRONMENT

Meeting capture environment and user interfaces go hand in hand. In fact, user interface functionalities are fundamentally limited by the underlying meeting capturing system. On the other hand, the requirements from user interface functionalities will impact the design of a meeting capture system.

3.1 Hardware

System designers typically have three choices when designing a meeting capture environment: using a static camera, using a camera that moves based on information from a microphone array, or using multiple cameras. Unfortunately, a single static camera rarely can cover enough area, a camera that moves based on a microphone array can often be slow and/or distracting, and multiple cameras are difficult to calibrate and set up.

These issues can be overcome using an omni-directional camera. In contrast to previous omni-directional camera systems [10,16,22], our environment uses a high resolution (1300x1030 = 1.3 Mega pixels) camera to both track meeting participants and capture video at 11 frames per second. This single camera has the resolution of 10+ normal video conferencing cameras – each 320x240 CIF video is only 76,800 pixels. Of course some of the resolution is being wasted capturing non-interesting portions of the scene. The system is shown in Figure 1. An example image captured by the system is shown in Figure 2.

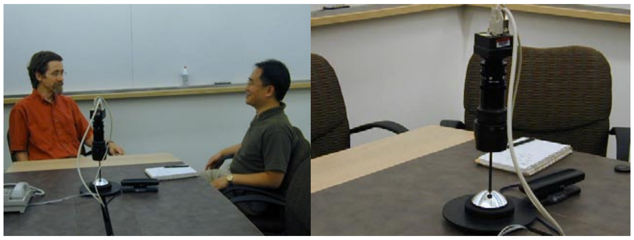
(a) (b)

Figure 1. The omni-directional camera meeting capture environment. (a) People seated around the meeting table. (b) Close-up of the parabolic mirror and camera.

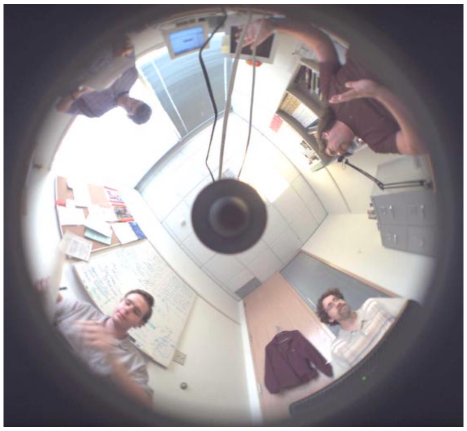
Figure 2. An example image captured by the omni-directional camera (shrunk to fit the page). While the captured image is warped, since the geometry of the mirror can be calibrated, it can be un-warped by the computer vision techniques.

3.2 Software

To create a completely autonomous meeting capture system, three companion software modules were developed: the omni-image rectifying software, the person-tracking software, and the virtual video director software.

3.2.1 Omni-Image Rectifying Software

As shown in Figure 2, the raw image captured by the omni-directional camera is warped. Fortunately, because the geometry of the parabolic mirror can be computed by using computer vision calibration techniques [11], the rectifying software can de-warp the image to normal images. Example de-warped images are shown in Figure 4. It is also almost effortlessly to construct a 360-degree overview image of the entire meeting site from the omni-image (top portion of Figure 5).

3.2.2 Person-Tracking Software

The person-tracking software decides how many people are in a meeting and tracks them. Dozens of person-tracking algorithms exist in the computer vision research community, each of which is designed for a different application. Some are designed for very accurate pixel-resolution tracking, but require initial identification of objects [12]. Others do not require initialization but are only good for frontal faces [4]. In our system, we cannot assume that initialization is possible or that faces are always frontal. Thus, we used motion-detection and skin-color tracking algorithms. Because people rarely sit still, motion can be used to detect the regions of a video that contain a person. A statistical skin-color face tracker [22] can be used to locate a person's face in the region so that the video frame can be properly centered. This person-tracker does not require initialization, works in cluttered background, and runs in real time.

3.2.3 Virtual Video Director Software

The virtual video director software decides on the best camera shot to display to the user. Note, because our omni-directional camera covers an area that multiple normal cameras can cover, we use "*camera shot*" to refer to a portion of the omni-image, e.g., the peoples' images that the person-tracking module has extracted, as shown in Figure 4.

There are many strategies the director can take. The simplest one is to cycle through all the participants, showing each person for a fixed amount of time. A more natural strategy is to show the person who is talking, as implemented in LiveWire [3,20] and later versions of Vic-Vat [25]. However, sometimes users want to look at other participants' reaction instead of the person talking, especially when one person has been talking for too long.

Based on discussions with four professional video producers from the corporate video studios, we decided to incorporate the following two rules into our director:

1. When a new person starts talking, switch the camera to the new person, unless the camera has been on the previous person for less than 4 seconds.
2. If the camera has been on the same person for a long duration (e.g., more than 30 seconds), then switch to one of the other people (randomly chosen) for a short duration (e.g., 5 seconds), and switch back to the talking person, if he/she is still talking.

Inspired by the virtual cinematographer work by He *et. al.* [7], the underlying logic for the virtual director is based on probabilistic finite state machines. These provide a flexible control framework. The parameters to the rules above are easily changeable, plus many of the parameters are sampled from distributions, so that the director does not seem mechanical to the human viewers.

3.3 Determining Who Is Talking

From the previous discussion, it's clear that knowing who is talking is important. Several approaches exist to address this problem. If each person is associated with one microphone, the problem can easily be resolved by examining the signal strength from each microphone. In the more difficult case where several people are in one room and each person is not associated with a single microphone, microphone arrays can detect who is talking using sound source localization algorithms, as used in PictureTel systems [18]. Limited by resources, we decided to manually annotate who is talking in this study. We consider the quality of human annotation to be the upper bound of the automatic speaker-detection algorithms.

4 INTERFACES EVALUATED

The focus of this research is to examine interfaces for viewing meetings captured by our omni-directional camera system, and to understand users' preferences and system implications. Figure 3 shows a high-level view of the client-server organization of such a meeting viewing system – on the left is the meeting capture camera system, in the middle is the video server where the captured meeting is stored, and on the right is the client system for on-demand viewing.

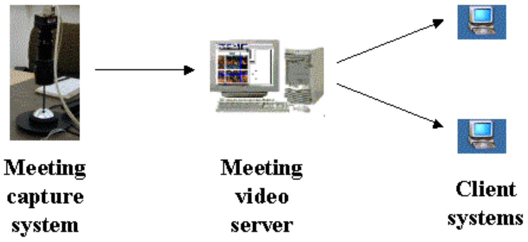

Figure 3. System block diagram

For our user studies, we have carefully chosen and implemented the following five interfaces to understand people's preference on seeing all the meeting participants all the time vs. seeing the "active" participant only; on controlling the camera themselves vs. letting the computer take control; and on the usefulness of the overview window provided by the 360-degree panoramic view:

- **All-up**: All members in the meeting are displayed side-by-side, each at the resolution of 280x210 pixels as shown in Figure 4. This is a common interface used in many current video conferencing systems [3].

  If there are N people in the meeting, this interface requires that all N video streams (one corresponding to each person) be stored on the video server, and all N be delivered to the client. In our specific case, assuming 4 people and each stream requiring 256Kbps bandwidth, it requires 1 Mbps of storage (~500Mbytes/hour) on the server and 1 Mbps of bandwidth to the client. While this should be easy to support on corporate intranets, it would be difficult to get to home even over DSL lines.

- **User-controlled + overview**: This is the interface

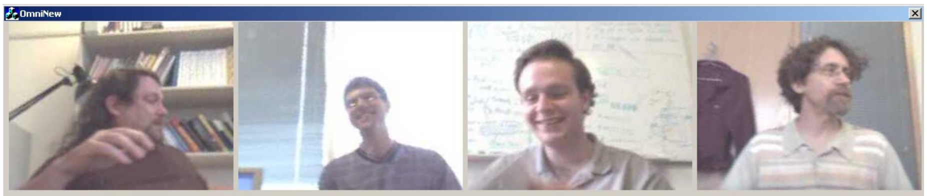

Figure 4. The all-up interface. Each video window has 280x210 pixels.

shown in Figure 5. There is a main video window (280x210 pixels) showing the person selected by a user, and an overview window whose total pixel area (648x90 pixels) is the same as that of the main video window. Note that the overview window is a full 360-degree panorama; so spatial relationships/interactions between people can be seen.

Users can click the five buttons at the bottom of the window to control the camera. The interface shows the "speaker" icon above the person who is talking. Clicking the rightmost button gave control of the camera to the virtual video director. It is worth pointing out that even though we name this interface a user-controlled interface, it actually *combines* both a user-controlled and a computer-controlled interface.

Given that the user can control whom he/she sees, this interface requires that the video server store all N video streams (one corresponding to each person) as in the all-up interface, plus the overview stream separately. From the bandwidth perspective, the bandwidth used is only 2x of what is needed by one person's video, thus 512 Kbps using the parameters mentioned earlier.

- **User-controlled**: This interface is exactly the same as the user-controlled + overview interface, without the overview window. The storage requirements on the server are the same as all-up interface, but the bandwidth to the client is 1/Nth that needed by the all-up interface, i.e., only 256Kbps using our parameters.
- **Computer-controlled + Overview**: This interface is exactly the same as the user-controlled + overview interface, except that the user cannot press the buttons to change the camera shot -- the video in the main window is controlled by our virtual camera director based on the rules described in the previous section.

  Because the user has no control over the camera, only the view selected by virtual director needs to be stored on the server. Thus the storage needed on server is only 2x of that needed by single stream (1x for main video, and 1x for overview), and the bandwidth needed is only 2x of single stream. The fact that storage and bandwidth requirements are independent of the number of people in the meeting makes this interface more scalable than the previous ones.
- **Computer-controlled**: This interface is exactly the same as the computer-controlled + overview interface, without the overview window. Thus the user sees the video selected by our virtual director. For this interface, both the storage requirements and bandwidth requirements are only 1x of that required by single video stream – roughly translating to 125Mbytes/hour for storage (less than $1) and 256Kbps of bandwidth using our parameters.

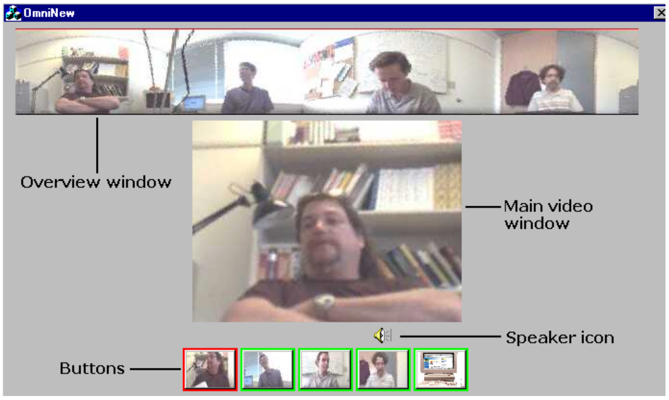

Figure 5: The user-controlled + overview interface. The window at the top is a 360-degree panoramic overview of the meeting room. The five buttons at the bottom are static images. Pressing these buttons changes the direction of the virtual camera. Pressing the fifth button gives control of the camera to the computer. The speaker icon above the buttons indicates who is currently talking.

Among the five interfaces, some show full-resolution video of all participants while others have only one main video window. Some have overview windows while others do not. Some are user-controlled while others are computer-controlled. These five interfaces are carefully chosen so as to allow us to effectively study the questions raised at the beginning of the paper.

5 STUDY METHODS

Following is the method we used to test the meeting interfaces.

5.1 Scenario

Subjects were told they had been out of town on business when four of their group members interviewed two candidates. Their task was to watch a 20-minute meeting held by the four group members the day before and decide which candidate to hire. Subjects were asked to take notes so that they could justify their hiring decision to upper management at the end of the study.

5.2 Study Procedure

Before watching the 20-minute meeting, subjects watched a five-minute training video captured by the same camera system in which each of the five interfaces was explained. Subjects then watched the meeting using the five interfaces. Each interface was used for four minutes of the 20-minute meeting. The order in which the interfaces were used was randomized to counterbalance ordering effects. After subjects watched the 20-minute video, they completed a survey.

5.3 Pilot Study

The whole study consists of a pilot study and a main study, separated by one week. 12 people participated in the pilot study and 13 people participated in the main study.

After reviewing data from the pilot study, we decided to make a few refinements to the interfaces. First, the pilot study subjects told us that the overview window was too small to be useful. We therefore increased it from 324x64 pixels in the pilot study to 648x90 pixels in the main study.

Second, subjects said that in the computer-controlled interfaces, the virtual director did not switch to the current speaker fast enough. Thus, the system was improved so that the virtual video director would switch to the speaker about 0.3 seconds quicker than the pilot study.

6 USER STUDY RESULTS

Unless otherwise noted, all of the following results are from the main study only.

6.1 Want to See All Participants or Not?

The all-up, computer-controlled + overview and user-controlled + overview interfaces show all the meeting participants all the time, though at different image resolutions. On the other hand, the computer-controlled and user-controlled interfaces only show a single meeting participant, selected either by the video director or by the user. Interface preference was measured using both rankings and ratings and summarized in Table 1. It is interesting that both results suggest a general trend that the interfaces showing all the meeting participants were favored over the interfaces showing only a single participant. This seems to indicate that remote users prefer to have a global context of the meeting, which agrees with the findings in the "live" teleconferencing systems [3,20]. It is worth pointing out that the high-preference is at the cost of more server storage, more network bandwidth and more screen real estate.

6.2 User-Control vs. Computer-Control?

For the user-controlled interfaces, all button clicks were logged. Figure 6 shows a histogram of subjects grouped by number of button presses. Two groups seem to emerge from this figure: those who like to control the camera, and those who don't. The top 5 subjects in terms of button presses account for 76% of all button presses, while the rest of the subjects only account for 24% of the button presses.

The notion that people can be divided into two groups is further supported by comments made in the post-study survey. One subject who controlled the camera a lot wrote,

Questions		Mean	Median	Std Dev
Rank order of the interface. (1 = like the most, 5 = like the least)	All-up	2.54	2.00	1.56
	Comp. controlled	3.85	4.00	1.14
	User controlled	3.54	4.00	1.33
	Comp. Contr. + overview	2.62	2.00	1.50
	User controlled + overview	2.46	2.00	1.13
Ratings: I liked this interface. (1 = strongly disagree, 7 = strongly agree)	All-up	5.08	5.00	1.93
	Comp. controlled	4.23	4.00	1.96
	User Controlled	4.54	5.00	1.85
	Comp. controlled + overview	5.38	6.00	1.89
	User controlled + overview	4.54	5.00	1.76

Table 1: Results from participants' rankings and ratings of the five interfaces.

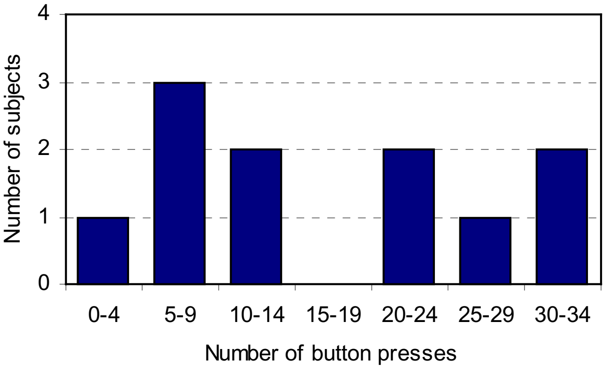

Figure 6. Histogram for button presses

> *The computer control although probably giving a better perspective, doesn't allow the user to feel in control.*

On the other hand, one subject who didn't control the camera much wrote,

> *I like having the computer control the main image so that I didn't have to think about who was talking and which image to click on – I could spend my time listening and watching instead without the distraction of controlling the image.*

The "two-group" hypothesis and finding is important in that when we design user interfaces for on-demand meetings, we need to take both groups into account.

6.3 Does the Virtual Camera Director Do a Good Job?

Because a large percentage of people like to have the computer control the camera, it is important to design and implement a good virtual video director. We made several improvements over LiveWire's design [3,20]. For example, we encoded two rules into the virtual director's knowledge and we explicitly made sure that the minimum shot length should be greater than four seconds (Section 3.2.3).

Feedback was quite positive (Main study in Table 2). In fact, in the user-controlled interfaces, seven out of thirteen subjects chose to use the computer-controlled mode for more than 30% of their viewing time.

Clearly, the success of virtual video director depends heavily on the accuracy of the "speaker detection" technique. From the pilot study to the main study, based on feedback, we have made speaker detection more prompt (about 0.3 second faster). This seemingly minor

(7 = strongly agree, 1 = strongly disagree)	Study	Mean	Median	Std Dev
The computer did a good job of controlling the camera.	Pilot (n = 12)	3.92	3.5	1.83
	Main (n = 13)	5.46	6.0	1.27

Table 2: Difference in perception of quality of camera control in the pilot study and the main study. In the main study, the speaker detection data were improved so that lag time to focus on the currently speaker decreased by about 0.3 seconds.

(7 = strongly agree, 1 = strongly disagree)	Mean	Median	Std Dev
When using interfaces with the overview window, I thought the overview window was helpful	5.69	6.0	1.70

Table 3. Study results on the usefulness of the overview window.

modification created a substantial change in attitude toward the virtual director's control of the camera, as shown in Table 2. A Mann-Whitney U test found that the feeling that the computer did a good job of controlling the camera increased significantly from the pilot study to the main study (z = -2.18, p = .035). These data indicate that people are *quite* sensitive to rather small delays in virtual director's switching camera to the currently speaking person.

6.4 Is the Overview Window Useful?

A unique feature of the omni-directional camera is its ease of constructing the overview video. In this section, we will study its usefulness from different perspectives.

6.4.1 Interfaces With and Without the Overview Window

By comparing the computer-controlled interface with the computer-controlled + overview interface, and the user-controlled interface with the user-controlled + overview interface, we can see the impact that the overview window makes. Overall rankings and ratings of the interfaces were provided earlier in Table 1.

Using the Wilcoxon Signed Ranks test, we found that rankings of both the user-controlled and computer-controlled interfaces were significantly higher with the overview than without (user-controlled: z = 2.03, p = .042; computer-controlled: z = 2.23, p =.026).

It is quite interesting that for the rating question, adding the overview window changed subjects' ratings for the computer-controlled interface (z = 1.85, p = .064) but provided almost no change for the user-controlled interfaces.

6.4.2 Survey Questions about the Overview Window

We also measured the usefulness of the overview window by asking a direct question if the overview window was helpful. A mean of 5.69 and median of 6.0 in a scale of 1.0 to 7.0 (Table 3) indicate most subjects thought the overview was indeed helpful.

6.4.3 Effect on Button Presses

In addition to examining survey data with regard to the overview window, we also explored whether the addition of the overview window changed the number of button presses that subjects made to change their camera shots for the user-controlled interfaces. The total number of button presses dropped from 103 in the user-controlled interface to 61 in the user-controlled + overview interface, although this difference was not significant (t(16.5)=1.20, p = 0.249).

7 DISCUSSION

Having examined all five interfaces on several dimensions, we now return to the questions we raised at the beginning of the paper.

7.1 Want to See All Participants or Not?

Most subjects preferred to see all the meeting participants. The all-up, computer-controlled + overview and user-controlled + overview interfaces were favored over the computer-controlled and user-controlled interfaces. It has been observed that in "live" meetings, remote audiences want to see all meeting participants to establish a global context [3]. Our finding indicates that for on-demand meetings this preference is still true.

7.2 User-Control vs. Computer-Control?

The data indicate a fairly clear split among subjects who like to control the camera and subjects who prefer to let the computer do the work. Based on this finding, it is important for us to take into account the different needs from these two groups of people when we design on-demand meeting user interfaces.

7.3 Does the Virtual Camera Director Do a Good Job?

The data show that the virtual video director did an excellent job of controlling the camera. Ratings of the computer's camera control were high (median of 6 on the 7-point scale) and when using the user-controlled interfaces, and seven out of thirteen subjects chose to use the computer-controlled mode for more than 30% of the time. Note, however, that these high ratings only appeared in the main study after we made the speaker-detection 0.3 second more prompt. This highlights the importance of spending the resources necessary to get speaker detection as fast as possible.

7.4 Is the Overview Window Useful?

In human vision system, people use their peripheral vision to monitor the global environment and use their fovea vision to concentrate on the object of interest. In our study, we hypothesized that the overview window would be helpful to provide contextual information to the users about what's happening in the entire meeting. The data from our study indicate that the overview window is worth the added bandwidth and screen real estate. The benefit of the overview window was also apparent in subjects' surveys. They wrote:

- *I liked having the overview so that I could see everybody's reactions (body language and facial expressions) to what was being said.*
- *I felt that the computer controlled with overview gave a good overall feel of the meeting. I could see who was talking and also quickly see the other's reactions to the speaker.*

It is quite interesting to see that the impact of the overview window is much bigger on the computer-controlled interfaces than that on the user-controlled interfaces (see

Table 1). This could be due to the fact that in the user-controlled interfaces, subjects' attention was distracted by clicking the control buttons.

7.5 Discussion Summary

To summarize, "user-controlled + overview" seems to be the winning interface for our targeted small group meetings. Its overview window provides a global meeting context for the user. Its design supports the "two-group" hypothesis: users can control the camera either by themselves or let the virtual video director take control. In addition, even though this interface uses the same storage as the all-up interface, its bandwidth is significantly lower. As the cost for storage is becoming negligible, network bandwidth is the main factor in system design tradeoffs.

The findings on the omni-directional camera system itself are also quite interesting and exciting: both of its unique features are proven to be important. First, its easy construction of the overview window provides the users with great added value. Second, a good virtual video director needs to switch cameras instantaneously as discovered in Section 6.3. While this is quite difficult to achieve for a single moving camera (e.g., slow) or for multi-camera systems (e.g., needs calibration), it is almost effortless for our omni-directional camera system.

8 CONCLUDING REMARKS AND FUTURE WORK

In this paper, we reported the design of an omni-directional camera system for capturing small group meetings. We studied various on-demand meeting viewing user interfaces that the system supports. Specifically, we focused on the issues of "viewing all meeting participants or not", "amount of user involvement", "how to design a good virtual video director" and "the usefulness of the overview window". Study results reveal that the subjects liked our omni-directional camera system and the features it offers in the interfaces (e.g., overview window). One subject wrote "Cool concept, I really liked the ability to view a meeting I could not attend so as to get a broader view of the topic."

There are still many interesting topics that remain to be explored. For example, we want to make the virtual video director more intelligent. Detecting the head orientation of the speaker will be valuable. When the speaker has been talking for too long, instead of switching to a random person, the video director can switch to the person the speaker is talking to. Second, because of the importance of fast speaker detection, we are developing microphone array techniques to achieve this goal. Finally, we want to integrate various meeting browsing techniques (e.g., time-compression [17], summarization [8,9], and indexing [22]) into our system to make on-demand group meetings more valuable and enjoyable to watch.

9 ACKNOWLEDGEMENT

The authors would like to thank Sing-Bing Kang, Gavin Jacke, Michael Boyle, and Dan Bersak for helping implement the system, and John McGrath, John Conrad, Travis Petershagen, and Greg Small for sharing their filming and directing rules with us.

10 REFERENCES

1. BeHere, http://www.behere.com/
2. Brave, S., Ishii, H. and Dahley, A., Tangible interface for remote collaboration and communication, *Proc. of CSCW'98*, 169-178.
3. Buxton, W., Sellen, A., & Sheasby, M., Interfaces for multiparty videoconferences, *Video-mediated communication* (edited by Finn, K., Sellen, A., & Wilbur, S.), Lawrence Erlbaum Associates, Publishers, 1997
4. Colmenarez, A. and Huang, T., Face detection with information-based maximum discrimination, *Proc. of IEEE CVPR*, June 17, 1997, 782-789
5. Daniilidis, K., Preface, *Proc. of IEEE Workshop on Omnidirectional Vision*, June, 12, 2000.
6. Elrod, S., Bruce, R., Gold, R. Goldberg, D. Halasz, F., Janssen, W., Lee, D., McCall, K. Pederson, E., Pier, K., Tang, J., & Welch, B., Liveboard: a large interactive display supporting group meetings, presentations and remote collaboration, *Proc. of CHI'92*, 599-607.
7. He, L., Cohen, M., & Salesin, D., The virtual cinematographer: a paradigm for automatic real-time camera control and directing, *Proc. of ACM SIGGRAPH'96*, New Orleans.
8. He, L., Grudin, J., & Gupta, A., Designing presentations for on-demand viewing, *Proc. of CSCW'00*, Dec. 2000
9. He, L. Sanocki, E., Gupta, A, Grudin, J., Comparing presentation summaries: slides vs. reading vs. listening, *Proc. of CHI'00,* 177-184.
10. Huang, Q., Cui, Y., and Samarasekera S., Content based active video data acquisition via automated cameramen, *Proc. IEEE ICIP'98, 808-811.*
11. Kang, S.-B., Catadioptric self-calibration, *Proc. of IEEE CVPR*, June 12, 2000, I:201-208.
12. Liu, Z., Zhang, Z., Jacobs, C., and Cohen, M., Rapid Modeling of Animated Faces From Video. *Technical Report, Microsoft Research 99-21*, April 1999.
13. McCanne, S. and Jacobson, V., Vic: a flexible framework for packet video, *Proc. ACM multimedia'95*, 511-522.
14. Moran, T. et al., I'll get that off the audio: a case study of salvaging multimedia meeting records. *Proc. of CHI'97*, 202-209.
15. Nicolescu, M., Medioni, G., and Lee, M., Segmentation, tracking and interpretation using panoramic video, *Proc. of IEEE Workshop on Omnidirectional Vision*, June, 12, 2000, 169-174.
16. Nishimura, T., Yabe, H., and Oka, R., Indexing of human motion at meeting room by analyzing time-varying images of omni-directional camera, *Proc. IEEE ACCV'00*, 1-4.
17. Omuigui, N., He, L., Gupta, A., Grudin, J., & Sanock, E., Time-compression: system concerns, usage, and benefits, *Proc. CHI'99*, 136-143
18. PictureTel, http://www.picturetel.com/
19. PictureTel, Enhanced continuous presence, http://www.picturete.com
20. Sellen, A., Remote conversations: the effects of mediating talk with technology, *Human-Computer Interaction*, 10(4), 401-444.
21. Stanford Online, http://stanford-onlines.stanford.edu/
22. Stiefelhagen, R., Yang, J., and Waibel, A., Modeling focus of attention for meeting indexing, *Proc. ACM Multimedia'99*, 1-10.
23. Tang, J. C., & Rua, M. Montage: providing teleproximity of rdistibuted groups. *Proc. of CHI'94*, 37-43.
24. USC Integrated Media Systems Center, http://imsc.usc.edu/education/education.htm
25. Wong, T., Hand and ear: enhancing awareness of others in MASH videoconferencing tools, Project report, University of California,

Berkeley, http://www.cs.berkeley.edu/~twong/classes/cscw/report.html

26. Zhu, Z., Rajasekar, K., Riseman, E., and Hanson, A., Panoramic virtual stereo vision of cooperative mobile robots for localizing 3D moving objects, *Proc. of IEEE Workshop on Omnidirectional Vision*, June, 12, 2000, 29-36.