Requirements and Recommendations for an Enhanced Meeting Viewing Experience

Sasa Junuzovic
Computer Science Department
University of North Carolina at Chapel Hill
Chapel Hill, NC
sasa@cs.unc.edu

Rajesh Hegde, Zhengyou Zhang, Phil A. Chou,
Zicheng Liu and Cha Zhang
Microsoft Research
Redmond, WA
{rajeshh,zhang,pachou,zliu,chazhang}@microsoft.com

ABSTRACT
We have found that viewing recorded meetings using traditional meeting viewers whose interfaces consist of an automatic speaker and a fixed context view does not provide sufficient information and control to the users. In particular, a survey of users who watch meeting recordings on a regular basis revealed that it is also useful to provide (1) speaker-related information, including who the speaker is talking to, looking at, and being interrupted by, and (2) more control of the interface, including changing the relative sizes of the speaker and context views and navigating within the context view. We present a 3D interface prototype designed specifically to meet these requirements when viewing recorded meetings. We describe in detail the results of a user study comparing the effectiveness of the new and traditional style interfaces with respect to these requirements. Based on this study, we present a set of guidelines for future interfaces.

Categories and Subject Descriptors
H.4.3 [Communication Applications]: Computer conferencing, teleconferencing, and videoconferencing. H.5.1 [Multimedia Information Systems]: Video; Audio input/output; H.5.2 [User Interfaces]: Interaction styles; User-centered design.

General Terms
Design, Human Factors, Experimentation.

Keywords
Remote Meeting Viewer, Requirements, Recommendations.

1. INTRODUCTION
Meetings are an integral part of workplace dynamics. However, due to travel, time, or other constraints, attending a meeting in person may not be practical for some invitees. As a result, a number of commercial systems, such as Microsoft's LiveMeeting and Cisco's WebEx, and research systems [1][2][3] have been developed for remote meeting attendance or offline viewing of recorded meetings. In either case, one of the main goals of these systems is capturing the relevant aspects of the meeting, without which people may not actually view the meeting. Capturing the relevant aspects is more important for offline meeting viewing, which is the case we focus on, than for remotely attending a meeting. The reason is that during an ongoing meeting, remote attendees can interrupt the conversation to ask for clarifications, which is not possible in the case of a person watching a recorded meeting.

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.

The aspects of a meeting that are important are meeting-dependent. In general, meetings can be roughly classified into two types. In one type of meeting, there are a large number of attendees, but only a few of them are active. An example of such a meeting is a lecture in which there is one lecturer and a large audience. In the other type of meeting, there are a small number of attendees, but the majority of them are active. Examples of such meetings are brainstorming sessions, team weekly status meetings, and new hire discussions. We focus on viewing recordings of the second, more interactive type of meeting.

A key aspect of interactive meetings is the current speaker, who is, by definition, changing frequently. Thus, traditional meeting viewing interfaces for such meetings have an automatic speaker view, which always shows the current speaker. Previous work [3] has also identified that the speaker view should be coupled with a fixed context view, which shows an overview of all of the attendees. The overview may show a small thumbnail-size video of each attendee or a panoramic view of the meeting room that captures all of the attendees.

Our research indicates that for viewing recorded meetings, the combination of an automatic speaker and fixed context views does not provide sufficient information. For example, we found that users desire speaker-related information such as who the speaker is talking to, looking at, and being interrupted by. We also found that they desire control of the context views. For instance, a user may want to focus on a non-speaking attendee in the context view. In this paper, we present a 3D interface prototype designed specifically to meet these requirements. A twenty participant user study revealed that the viewing experience is better with the 3D prototype than with the traditional interface.

The rest of this paper is organized as follows. We first motivate the need for each additional kind of information and control in the interface. Following this, we describe the 3D interface prototype. We then present in detail the results of a study comparing the 3D interface to a typical existing system. From the results, we extract several new guidelines for future meeting viewing systems. Finally, we end with related work, brief conclusions, and directions for future work.

2. SYSTEM REQUIREMENTS
The requirements of a meeting viewing system are a function of the type of meeting. As mentioned above, we focus on the type of meeting in which there is a small number of attendees (fewer than

ten) most of whom equally contribute to the discussion. Previous studies have identified the need for an automatic speaker and context views when viewing such meetings. To illustrate a typical traditional interface, consider a meeting in which five attendees, Alice, Bob, Charlie, Dave, and Eve, are seated around a table as shown in Figure 1 (top). Each attendee's face is captured from the front by a different camera. We will use this example as a running example in the rest of the paper. (The meetings used in our study also had five attendees in the same seating arrangement, and a separate camera captured each attendee from the front.) In this example, a typical traditional interface shows a large view of the current speaker above thumbnail size videos of all of the attendees as shown in Figure 1 (bottom). Figure 1 (bottom-left) and Figure 1 (bottom-right) show the traditional interface when Alice and Eve are the speakers, respectively. A panoramic video [3] of the entire room can be shown instead of the thumbnails if an omni-directional camera is available to capture the meeting.

Based on previous videoconferencing research and our experience with viewing recorded meetings, we have identified two sets of issues with the traditional interfaces for viewing recorded meetings. One set of issues are the difficulties in interpreting speaker-oriented information, such as who the speaker is speaking to, looking at, or being interrupted by. While some or all of this information can be implicit in the dialogue, there are times when it is not. For instance, suppose that during the meeting, Bob asks "*What are the fourth quarter profits like?*" which sparks a discussion on the company performance in general. Eventually, to answer Bob's question, Alice says "*Returning to your question ...*" and looks at Bob. At that point, everyone present in the meeting room knows that Alice wants to discuss fourth-quarter profits. However, an observer, such as a remote viewer, without the knowledge of who Alice is looking at may get confused. Moreover, as a part of her answer to Bob's question, Alice says "*They are better than we expected*" and briefly looks at Charlie, who is the accountant, for confirmation. Charlie agrees by nodding without interrupting. At the same time, Eve agrees by saying "*Much much better,*" which causes Alice to glance in Eve's direction. Everyone in the meeting knows that Charlie and Eve agree with Alice's claim. Someone who is later viewing a recording of this meeting, on the other hand, may neither realize that Charlie agrees with Alice as Charlie does not say anything

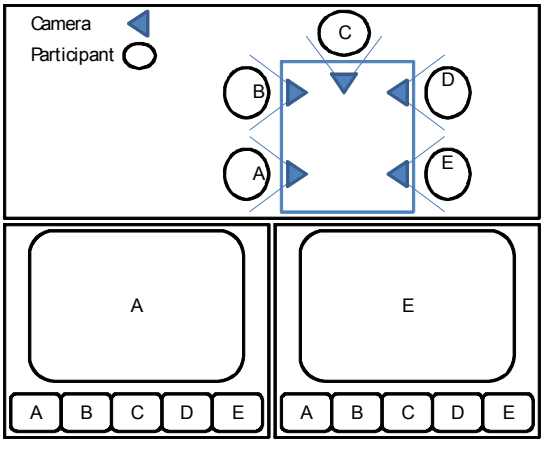

Figure 1. (top) A birds-eye view of a meeting with five attendees, Alice, Bob, Charlie, Dave, and Eve, who are seated around a table; (bottom) the traditional interface showing when (left) Alice and (right) Eve is the speaker.

nor figure out that it was Eve who confirmed what Alice said. The reason is that in the traditional interface, it is difficult to interpret the direction in which the speaker, or any other attendee, is looking. For instance, in our running example, when Alice is looking at Eve, she is looking straight-ahead (Figure 1 (top)), which to the users appears as if she is looking at them. Moreover, the users cannot easily tell that Alice first looks at Charlie and then at Eve, regardless of whether the interface displays thumbnail-size videos of the attendees or a panoramic video of the meeting room.

Another set of issues with the traditional interfaces is that the context views do not allow the user to control any aspect of the overview. The lack of control, combined with the fact that these views are small, makes it difficult to focus on a non-speaking attendee even though there are instances when the speaker is not the focus of attention. In our running example, while Alice is answering Bob's question, it is useful to be able to focus on Bob to see his reaction – perhaps he does not agree with what Alice is saying or is confused by it. Moreover, suppose that Alice's answer triggers a side conversation between Charlie and Dave. In this case, it is useful to be able to focus on the part of the meeting room in which Charlie and Dave are sitting in order to see both of them at the same time.

Based on these issues, we abstract out two high-level requirements for meeting viewing interfaces, which are not fully addressed by current meeting viewing systems:

> *Requirement 1:* the interface should correctly convey to the user who an attendee is looking at.
>
> *Requirement 2:* the context view should allow the user to focus on any attendee or any part of the meeting room.

In the next section, we describe a meeting viewing system we developed specifically to meet these requirements.

3. SYSTEM DESCRIPTION

In general, a meeting viewing system consists of three parts: the recording equipment; a system for processing the recordings; and the user-interface for viewing the processed recordings.

3.1 Recording Equipment

The recording equipment has a large impact on the rest of the system. For example, if a single, regular camera captures the meeting, the video presented in the interface is simply the one captured by the camera. On the other hand, if a high-resolution, wide-angle camera is used, then the video can be processed to extract separate video streams of all of the attendees, which can be presented as thumbnails in the interface. In fact, most of the systems require substantial infrastructure or hardware setup, such as an omni-directional camera [3], specially positioned IP camera and microphones, and carefully designed rooms and dedicated high-speed connections among remote sites as in HP's Halo and Cisco's TelePresence systems. There are two problems with such systems – they are expensive and difficult to set up.

We assume that such heavy infrastructure for recording meetings is not available. Instead, we assume only that there is a camera in front of each attendee in the meeting room. This assumption is not unrealistic. For example, meeting attendees often bring with them laptops to, for example, take notes, and most modern laptops are equipped with a built-in camera and microphone. Alternatively,

(cheap) IP cameras can also be mounted on the table such that a camera faces each chair.

While the recording equipment impacts the processing and interface components, these two components have a more symbiotic relationship between them – new interface requirements drive the design of the processing component, and new processing components enable previously unrealizable interfaces. We next describe an interface that meets the requirements derived above and the processing component required to realize the interface.

3.2 Interface Prototypes

As mentioned above, a meeting viewing interface should allow for an easy and correct interpretation of who an attendee is looking at. To meet this requirement, we created an interface in which the attendees' videos are positioned in a manner that replicates the spatial relationships (scaled to fit the interface) between the attendees in the meeting. Such an interface for our running example when Charlie is the current speaker is shown in Figure 2 (top). As Figure 2 (top) shows, the video of the speaker is large and the videos of all of the other attendees are small. Unlike in the traditional interface, in which the videos of the non-speaking attendees are shown side-by-side, in the new interface, the positions and angles of these videos reflect the actual seating arrangement of the attendees. For example, because Alice sits to the right of Bob, her video is positioned to the right of his. Similarly, because Eve sits to the left of Dave, her video is positioned to the left of his. In addition, because Bob and Dave face each other, their videos also face each other. Similarly, Alice's and Eve's videos face each other. Moreover, the angles and positions are set with respect to the current speaker, whose video is always shown from the front, as if the user viewing the meeting is sitting across the table from the speaker. For instance, when Charlie is the speaker, the user sees the meeting from the place on the table directly opposite of Charlie. Therefore, Charlie's video is positioned so that it appears to be farther away than Bob's and Eve's videos, which are in turn, positioned so that they appear to be farther away than Alice's and Eve's videos. In order to prevent one video from obscuring another video, we use alpha blending techniques and show the video in front as semi-transparent. We refer to this interface as the ***3D interface*** as it uses a 3D analogy to preserve the spatial relationships among the attendees, and we refer to the traditional interface as the ***2D interface*** because it displays the videos in non-3D environment.

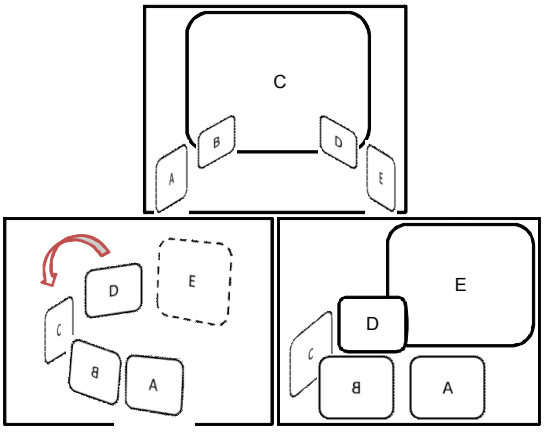

Figure 2. The 3D interface (top) when Charlie is the speaker; (bottom-left) during the transition to Eve the speaker; and (bottom-right) when Eve is the speaker.

Compared to the 2D interface, the 3D interface makes it easier to understand who the speaker is looking at. For example, during the meeting, when Charlie looks at Alice or Eve, his eyes or head turn slightly to the right or left, respectively. When he looks at Bob or Dave, his eyes and head turn significantly right or left, respectively. Since in the interface, the user is looking directly at Charlie when Charlie is the speaker, the user can tell who Charlie is looking at by following his eyes.

When the speaker changes, a large video of the new speaker is shown from the front. The videos of all of the other attendees are small. Moreover, the positions of the videos are rearranged to display the correct spatial information of the attendees with respect to the new speaker. The question is how to perform this rearrangement. Suppose that the current speaker changes from Charlie to Eve. In the 2D interface, the large video simply cuts from Charlie to Eve – the interface snapshots just before and after Eve becomes the speaker are shown in Figure 1 (bottom). The only aspect that changes in the interface is the large video at the center. The 3D interface could also "cut" to Eve by instantaneously switching the sizes and positions of all of the videos. However, the user may get confused when many parts of the interface change suddenly. Hence, the 3D interface instead gradually rotates around the center of the table as shown in Figure 2 (bottom). During the rotation, the large video of Charlie fades out and the large video of Eve fades in.

Another issue that occurs when, in our running example, Bob (or anyone other than Charlie) is the speaker is how to display the video of the attendees who are on the opposite side of the table from the speaker. Recall that in our example, all of the cameras are on the table. Hence, no camera is capturing the attendees from behind. Nevertheless, we still show the video of these attendees captured by cameras on the table. In this solution, however, when these attendees look left, such as when Alice looks at Dave, it appears to the user as if they are looking to the right. To correct the issue, we horizontally flip the videos of these attendees. For example, when Eve is the speaker, videos of Alice and Bob are horizontally flipped in the interface as shown in Figure 2 (bottom-right). As our study results will show, the users had no difficulties interpreting such positions correctly.

So far, we have addressed only one of the requirements, namely, the requirement for correctly displaying the direction in which an attendee is looking. The second requirement is to provide the users of the system with the ability to focus on any attendee or any part of the meeting room. One way to do this is to allow the user to click on any video and then show a larger version of that video in the interface. Since each interface already shows the current speaker in a large video, showing an additional large video may be confusing. A less confusing solution is to separate the auto speaker view from context view. The 3D interface in which the speaker and context views are separated is shown in Figure 3. As Figure 3 shows, the speaker view is displaying Alice, the current speaker, while the context view shows the seating arrangement of the attendees through the locations of their videos. The context view allows a user to focus on a particular attendee by clicking on that attendee's video. Figure 3 shows the case when the user focuses on Charlie. The environment rotates around the center of the table until Charlie's video is positioned at the center.

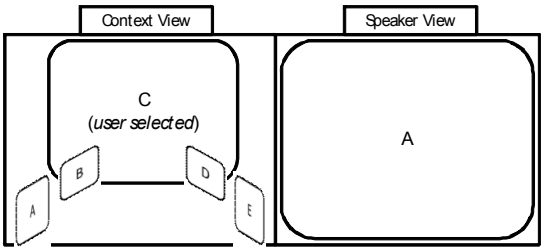

Figure 3. The 3D interface with speaker and context views shown separately.

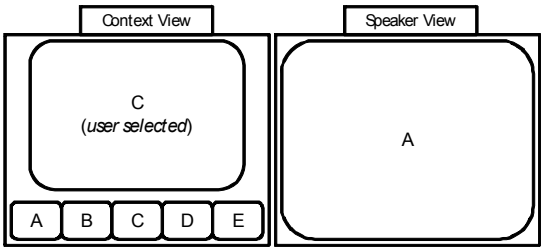

Figure 4. The traditional 2D interface with speaker and context views shown separately.

Moreover, users can zoom into, pan, rotate, and tilt the environment shown in the context view. These navigation controls are similar to the popular map applications such as Virtual Earth and Google Maps. As with the 3D models in those systems, the context view in our 3D interface supports 1) panning by dragging any part of the context view, 2) tilting or rotating by holding down the CTRL key and dragging, and 3) zooming by holding down the SHIFT key and dragging. Users only need to use the navigation controls to get a better view of the meeting or to focus on a particular attendee in the context view. For example, recall that in our running example, Alice's comments on the profits triggered a side conversation between Charlie and Dave. Since Alice continued to talk, the auto speaker view continues to show Alice. To see the side conversation between Charlie and Dave, the user can either click on Charlie's video to see his expression or zoom into the environment and adjust the tilt and pan to see Charlie's and Dave's videos clearly in the context view. By providing the ability to focus on any attendee or any part of the meeting room, the 3D interface satisfies the second requirement.

Hence, the 3D interface has two modes. In one mode, the speaker and context views are combined into a single view as shown in Figure 2. We call this mode the ***Automatic*** mode since the user cannot interact with the combined view. In the second mode, the speaker and context views are shown separately as illustrated in Figure 3. We call this mode the ***User-controlled*** mode because the user can control the context view. We explore which mode is more useful through a user study which we describe shortly.

One issue with separating the speaker and context views is that the users may have different preferences on the relative sizes of these two views. We addressed this by allowing the users to adjust the relative sizes of the two views.

3.3 Audio-Video Processing

Regardless of whether the speaker and context views are shown together or separately, the 3D interface needs the positions of the videos that preserve the spatial relationship among the attendees. In our case, the interface needs to know where the attendees are sitting. Since we assume that a camera located on the table captures each person in the meeting from the front, we can get their approximate locations from the camera locations. The camera locations can be automatically extracted by processing the captured media. There are two ways to obtain the camera locations. One approach is to use video information. If there are overlaps between cameras, the so-called structure from motion technique in computer vision can be used [4][5]. The second approach is to use audio information. For example, if cameras and microphones are close together, as is the case when using laptops to record the meeting, microphone, and hence, camera locations can be determined based on relative audio energy decay [6]. Since camera localization is beyond the scope of this paper, we manually configure the locations of the videos in the interface.

4. USER STUDY

So far, we have motivated new requirements for offline meeting viewing systems and have presented a new 3D interface that is designed specifically to better meet these requirements. It is also important to (a) verify that the users also find the requirements we set forth important and (b) evaluate how well the traditional 2D and our new 3D interfaces satisfy these requirements. In this section, we present the results of a study of these issues.

For a fair comparison between the 2D interface and the 3D interface in which the speaker and context views are shown separately, we created automatic and user-controlled modes of the 2D interface. The automatic mode is simply the traditional 2D interface. The user-controlled mode, on the other hand, shows the speaker and context views separately as shown in Figure 4. The user can click on the thumbnail video of any attendee in the context view to display the attendee's video in the main video rectangle in the context view. In Figure 4, the speaker view is showing Alice while the user has focused on Charlie in the context view.

4.1 User Study Description

To carry out the study, we recruited twenty people from our organization who remotely attend or view recorded meetings on a regular basis. The group was gender balanced and consisted of people from different job functions and experience.

In a controlled lab study that compares the different interfaces, ideally the study participants should view semantically consistent but syntactically different meetings using each interface in order to obtain valid comparisons. Moreover, these meetings should be realistic and highly interactive. To ensure that the consistency requirement is met, we recorded two such meetings taking place in the same room with the same attendees. In both meetings, the attendees had to select the top three alternatives out of five. In one meeting, they chose three out of five cars to recommend to others, while in the second meeting, they chose three out of five graduate students to hire as interns. These meetings are similar to actual meetings in which the attendees have to prioritize items. Each meeting had five people seated around a table as shown in Figure 1. To make the discussion as interactive as possible, the attendees were provided with initial data. In the car-selection meeting, they were given information such as mileage, horse power, and the amenities of each car. In the intern-selection meeting, they were given information such as education, publications, and work experience of each student. In our experience, prioritization-type meetings often involve personal preferences which can lead to conflict. To mimic this aspect of an actual meeting in our recorded sessions, we purposely led some attendees to introduce conflicting

opinions. For instance, in the car selection meeting, two of the people in the meeting were told that they were getting kick-backs from one of the car companies and should therefore push for that company's car to be selected even though it was inferior compared to the other cars. Similarly, in the intern-selection meeting, two attendees were told that they should prefer an intern from the schools they attended for college. None of the meetings were scripted and they were free to conduct the meetings the way they wanted. The meetings lasted about six minutes.

As the user-controlled modes of the interfaces have a certain degree of customization, we recorded a third meeting, which the study participants watched to figure out their interface preferences. As this recording was only for interface setup, it was not as important for the meeting to be realistic. However, it was still important that the meeting is interactive. Hence, we recorded five people (seated as shown in Figure 1) as they discussed their favorite movies for two minutes. Most of them had many favorite movies, so the discussion was very interactive.

In all three of the meetings we recorded, the attendees were the same five people from our organization. As each study participant watched all three meetings, we shuffled the attendees' seating arrangement prevent any learning factors from carrying over from one viewing to another. Regardless of the meeting, however, the five people in the meeting were seated around the table in the same manner as shown in Figure 1 (top). Moreover, we placed a tabletop IP camera in front of each intern.

As described earlier, our goal for the study was to evaluate (a) the relevance of the system requirements and (b) how well the 2D and 3D interfaces satisfy these requirements. In the remainder of this section, we describe the five steps each participant performed during a session: 1) complete an initial questionnaire; 2) participate in a training session; 3) setup the interface preferences and complete a questionnaire regarding the preferences; 4) watch the car and intern selection meetings, one with the 2D and the other with a 3D interface, and complete questionnaires about the viewing experience; and 5) complete a closing questionnaire and go through a short debrief interview. All questionnaire questions were answered using a 5-point Likert scale where 1 = "strongly disagree" and 5 = "strongly agree". We interleaved positive and negative questions in the questionnaires so the participants did not follow a specific answer pattern.

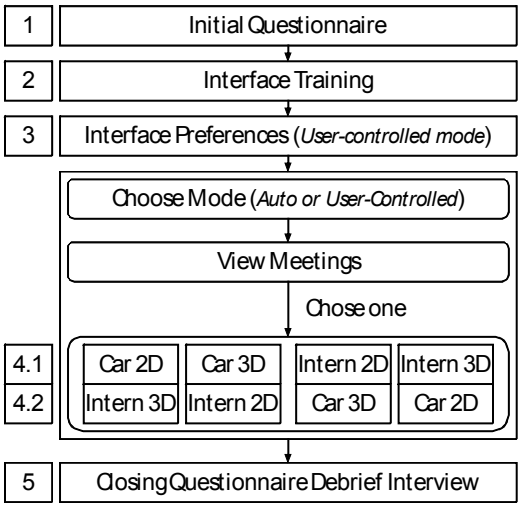

Figure 5. User study session procedure. Arrows leading from one step to the next are taken randomly.

We next describe the study procedure steps in more detail. To counter any ordering effects, the steps were not identical for all participants. The set of possible procedures is displayed as a flow-chart shown in Figure 5.

1. Initial Questionnaire: At the beginning of each session, the study participants completed the questionnaire shown in Table 1, which evaluated if they felt the system requirements we derived earlier are important. The participants did not view a meeting prior to completing the questionnaire. Their responses were based entirely on their prior experience with traditional systems.

2. Training: Following the completion of the questionnaire, each participant took part in a short (ten minute) training session. During training, the study administrator demonstrated all of the features of the 2D and 3D interfaces.

3. Interface Customization: After getting accustomed to the interfaces, the participants customized the interfaces. For instance, for both 2D and 3D interfaces in the user-controlled mode, the participants selected their preferred relative sizes of the speaker and context views. To do this, they adjusted the sizes of the views using a slider bar control. There were seven different discrete settings on the slider bar. The left-most setting (value=1) showed context view in full with no speaker view, while the right-most setting (value=7) showed speaker view in full with no context view. The middle setting (value=4) showed the two views in equal size. Other intermediate values of the slider showed the two views

Figure 6. 3D interface in the user-controlled mode configured to show a larger context view.

Figure 7. Traditional interface (2D) in the user-controlled mode configured to show a larger context view.

in the corresponding intermediate sizes. In addition, in the 3D case, they also selected the zoom level. The settings chosen by each participant were used as the initial settings when watching the other two meetings.

Figure 6 and Figure 7 show the actual 3D and 2D interfaces, respectively, in the user-controlled mode. In each interface, the speaker and context views are shown on the right and left, respectively. The tool bar at the top contains the slider bar that controls the relative sizes of the speaker and context views. Moreover, in the 3D case, the toolbar also contains a check box for showing the room boundaries as a wire-frame mesh and a check box for showing the virtual camera positions on the table. At the completion of the meeting, the participants indicated their preferences in a questionnaire shown in Table 2, which included questions regarding the importance of the speaker and context views and their sizes.

4. Interface Evaluation: Once the participants indicated the preferences for the two interfaces, they watched the car and intern selection meetings. For each participant, we randomly decided if they watched the meetings using the user-controlled or automatic interface mode. Moreover, we randomly decided if the car or the intern selection meeting was shown first. Finally, we randomly decided if the participant used the 2D interface for the first and the 3D interface for the second meeting or vice versa. The set of possible combinations is illustrated in Figure 5. By doing random selections at each step, we counter-balanced all ordering effects for each participant. After watching each meeting, the participants completed the questionnaire shown in Table 3. The questions evaluated the interfaces with respect to the requirements and were exactly the same for both the automatic and user-controlled modes, with an additional question in the user-controlled case to evaluate the navigation controls.

5. Closing Questionnaire and Debrief Interview: At the end of each session, the participants completed a final questionnaire, shown in Table 5, which contained questions regarding the viewing experience with the 3D interface. Finally, each participant took part in a debrief interview.

4.2 Validation of System Requirements

As mentioned above, we derived new speaker-oriented and control-oriented requirements for meeting viewing systems. The former included who the speaker is looking at, talking to, and being interrupted by, while the latter included the ability to focus on any person in the meeting or any part of the meeting space at any time. To evaluate if these requirements are actually relevant to them, the participants completed the questionnaire shown in Table 1 as the first step of the study.

For all of the questionnaires, we performed a single tail t-test on each question. We assumed that the population mean as the center of our ranking scale at $\boldsymbol{\mu = 3.0}$. We use the null hypothesis that our observed mean stays close to the population mean and an alternate hypothesis that the observed mean $\mathbf{X > \mu}$ and $\mathbf{X < \mu}$ for positively and negatively phrased questions, respectively. Using the t-test, we compute the probability $\boldsymbol{p}$ that our observed mean is away from the population mean by chance. Traditionally, if $\mathbf{p < 0.05}$, then the difference between the means is considered significant and not by chance. Additionally, for the single tailed test, a positive (negative) t value indicates the observed mean **X** is significantly higher (lower) than the population mean $\boldsymbol{\mu}$.

As Table 1 shows, the study participants on average agreed with the following statements: "1) It is useful to know who the speaker is" (Mean, Median = 4.65, 5); "4) It is useful to know who is interrupting the speaker" (4.2, 4); "5) It is useful to know participants' seating arrangement in the room" (3.45, 4); "7) It is useful to focus on a non-speaking participant sometimes" (3.75, 4); and "8) It is useful to have control over what/who you can focus on (for viewing)" (4.3, 4). Moreover, the study participants

Table 1. Initial questionnaire results (Sample Size = 20).

#	Survey Question	Mean	Med	STD	t(19)	p
1	It is useful to know who the speaker is.	4.65	5	0.93	7.91	≈ 0
2	It is **NOT** useful to know who the speaker is talking to.	2.00	2	0.79	-5.63	≈ 0
3	It is **NOT** useful to know who the speaker is looking at.	2.25	2	0.91	-3.68	0.0008
4	It is useful to know who is interrupting the speaker.	4.20	4	0.77	6.99	≈ 0
5	It is useful to know participants' seating arrangement in the room.	3.45	4	1.10	1.83	0.0414
6	It is **NOT** useful to see all the participants.	2.30	2	1.03	-3.04	0.0034
7	It is useful to focus on a non-speaking participant sometimes.	3.75	4	0.64	5.25	≈ 0
8	It is useful to have control over what/who you can focus on (for viewing).	4.30	4	0.73	7.93	≈ 0

Table 2. Interface preferences questionnaire results (MD=Median, SD=Standard Deviation). Ranking Scale: 1=Strongly Disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly Agree, p column cell shown in bold if NOT found significant at 0.05.

#	Survey Question	UI	Mean	MD	STD	t	p
1	I liked being able to resize the views.	2D	4.3	4	0.733	7.93	≈ 0
		3D	3.85	4	1.226	3.10	0.002941
2	My favorite slider value was:	2D	3.35	4	1.089	-2.67	0.007596
		3D	3.65	4	1.226	-1.28	**0.108512**
3	The automatic speaker view is important.	2D	4.45	5	0.759	8.54	≈ 0
		3D	4.5	5	0.761	8.82	≈ 0
4	The context view is **NOT** important.	2D	1.95	2	0.605	-7.76	≈ 0
		3D	2	2	0.858	-5.21	0.000025

disagreed with the following *negatively phrased statements*: "2) It is **not** useful to know who the speaker is talking to" (2, 2); "3) It is **not** useful to know who the speaker is looking at" (2.25, 2); and "6) It is **not** useful to see all the participants" (2.3, 2). Based on the single tail t-test analysis, the average response for each question is significantly different from the null hypothesis, leading us to reject the null hypothesis for all questions. Thus, the responses indicate that the two new sets of requirements are indeed important.

4.3 Interface Customization Results

As the interfaces we developed had many customization options, our next goal was to find out 1) the relative importance of context view and the automatic speaker view and 2) the participants' preferences regarding the relative sizes of these views. Therefore, each participant watched the two-minute movie discussion, during which the study administrator recorded the participant's preferences. Following this, each participant filled out the questionnaire shown in Table 2. As with the questions in the previous questionnaire, we checked whether or not the responses were significantly different from the null hypothesis using t-tests. As Table 2 shows, the participants "1) liked being able to resize the views" in both 2D (4.3, 4) and 3D (3.85, 4) interfaces. There was a decrease in the mean from 2D to 3D, but Mann-Whitney test *(z = 0.89, p=0.5)* suggests this difference is not significant. Some participants preferred a larger speaker view and a smaller context view because they felt that the speaker is the most important person in the meeting, while others preferred a larger context and a smaller speaker view because the videos of all of the attendees are larger. Running t-tests on the responses suggests that the results are statistically significant.

As question 2, "My favorite slider value was:" in Table 2 shows, the mean (2D=3.35, 3D=3.65) fell slightly below the middle setting of 4 which shows the same size speaker and context. This makes sense because at slider value of 3, the video sizes of all the attendees are large enough to show enough detail. Moreover, slider values below 3 cause the speaker view to appear quite small. However, t-tests reveal that only the 2D case responses are significantly different from the null hypothesis (*Mean=3.35, t(19)=-2.67, p=0.0076*). This may be because a zoom control was provided in the 3D view, which the participants used to adjust the thumbnail video size rather than adjusting the view size itself.

As for the other two questions, the participants agreed that both the automatic speaker and context views were important for both the interfaces, which is in-line with the findings presented in [1][3]. T-tests revealed that the results are statistically significant.

4.4 Interface Evaluation Results

So far, we have 1) validated the importance of the requirements we derived above and 2) gathered the preferences regarding the interface settings. Finally, we can evaluate how well the 2D and 3D interfaces satisfy the requirements. To find out if one interface satisfies them better than the other, each participant viewed the intern-selection and car-ranking meeting recordings. One recording was viewed using the 2D interface and the other using the 3D interface. Of the twenty participants, ten viewed the meetings in the automatic mode and ten viewed the meetings in the user-controlled mode. At the end of each meeting, they filled a questionnaire regarding how well the interface satisfied the requirements, results of which are shown in Table 3.

As before, we use the t-test to determine if the observed mean is significantly greater or less than the hypothesized population mean and the Mann-Whitney test to compare if the means between two treatments are significantly different. Since we randomized the order of showing the two interfaces and also the two modes, the Mann-Whitney test applies. We use the typical **$p < 0.05$** for testing significance of the differences. In the rest of this section, we discuss some of the more interesting 1) responses, which are shown in Table 3, and 2) their Mann-Whitney interface comparison results, shown in Table 4. Both sets of results are analyzed on a per mode basis.

Q1: Easy to tell who the speaker was? The participants found it easy to tell who the speaker was in all cases. This is expected since the automatic speaker views in both modes always showed the current speaker.

Q2: Easy to tell who the speaker was looking at? We expected the 3D interface to do better here since it had incorporated the relative seating arrangement information of the attendees. From the t-tests, it was clear that figuring out who the speaker was looking at using a 2D interface was difficult. The Mann-Whitney tests indicated 3D interface did well for the user-controlled mode. ($z= -1.74$, $p = 0.041$).

Q6: Easy to figure out the seating arrangement? The participants strongly agreed that the 3D interfaces successfully conveyed the seating arrangement in both the modes. T-tests suggest a strong significance as well. The Mann-Whitney tests indicates that the 3D interface did much better in both the modes ($z=2.83$, $p=0.002$ in automatic mode and $z=3.33$, $p\approx0$ in user-controlled mode). Clearly, the 3D interface better conveys the seating arrangement than its 2D counterpart.

Q7: Easy to see all of the participants' faces: In the automatic mode, 2D actually did better (Mann-Whitney $z= -1.78$, $p=0.038$) whereas in the other mode, both 2D and 3D did equally well. We attribute this to the fact that the speakers in our recorded meetings switched very often due to the highly interactive nature of the meetings causing the 3D interface to rotate often in order to bring the speaker to the center. This confused some study participants, and hence they were unable to pay attention to the rest of the attendees in the meeting. We believe that better camera switching rules [7] will reduce the confusion.

Q8: Easy to follow speaker transitions? The results suggested that both interfaces did equally well in conveying what happened during speaker transitions.

Q9: Easy to focus on a non-speaking participant? In the automatic mode, the t-test results suggested no significant difference, while in the user-controlled mode, the participants found it easy to focus on non-speaking attendees. However, the Mann-Whitney test suggested both 2D and 3D did equally well at $p=0.05$ confidence level. But, at $p=0.10$ level, 2D did slightly better. This is probably because when the participants tilted or rotated the context view in the 3D interface, some of the videos overlapped causing a blurred effect. A better algorithm for arranging the videos when the participants navigate the interface should help improve the 3D perception. One solution is to adjust the spatial positions of each video to utilize the available space better so that for any tilt and zoom level, the videos are arranged in a way that minimizes the overlap among them. On several occasions, this would amount to altering the exact spatial relationships. But we believe as long as we maintain the proper ordering and orientation, this should not present any problems.

Table 3. Interface Evaluation Questionnaire Results (SD=Standard Deviation). Ranking Scale: 1=strongly Disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=strongly Agree, p column cell value shown in bold if NOT found significant at 0.05.

#	Question	UI	Automatic Mode				User Controlled Mode			
			Mean	STD	t(9)	p	Mean	STD	t(9)	p
1	It was easy to tell who the speaker was.	2D	4.3	0.95	4.33	0.0009	4.3	0.95	4.33	0.0009
		3D	4.3	0.48	8.51	≈ 0	4.7	0.48	11.13	≈ 0
2	It was easy to tell who the speaker was looking at.	2D	1.7	0.48	-8.51	≈ 0	1.9	0.88	-3.97	0.0016
		3D	2.3	1.16	-1.91	0.0443	2.7	0.95	-1.00	**0.1717**
3	It was difficult to tell who the speaker was talking to.	2D	3.5	1.18	1.34	**0.1063**	4	0.82	3.87	0.0019
		3D	3.8	1.14	2.23	0.0264	3.2	1.03	0.61	**0.2777**
4	It was easy to tell who was interrupting the speaker.	2D	3.1	0.88	0.36	**0.3632**	3.3	1.16	0.82	**0.2172**
		3D	2.9	1.29	-0.24	**0.4057**	3.4	1.07	1.18	**0.1347**
5	It was difficult to tell who was going to be the next speaker.	2D	3.5	0.85	1.86	0.0479	4	0.82	3.87	0.0019
		3D	3.7	0.95	2.33	0.0223	3.4	1.17	1.08	**0.1546**
6	It was difficult to tell the participants' seating arrangement.	2D	4	1.25	2.54	0.0160	4.2	0.92	4.13	0.0013
		3D	2	0.94	-3.35	0.0042	1.7	0.95	-4.33	0.0009
7	It was difficult to see all the participants' faces.	2D	1.4	0.52	-9.80	≈ 0	2.1	0.74	-3.86	0.0019
		3D	2.1	0.88	-3.25	0.0050	2.2	1.03	-2.45	0.0184
8	It was easy to follow speaker transitions.	2D	4	0.47	6.71	≈ 0	3.6	0.97	1.96	0.0406
		3D	3.3	1.42	0.67	**0.2602**	3.7	0.82	2.69	0.0124
9	It was hard to focus on a non-speaking participant.	2D	2.5	1.58	-1	**0.1717**	2.3	1.25	-1.77	0.0554
		3D	3.3	1.34	0.71	**0.2481**	1.9	0.88	-3.97	0.0016
10	It was useful to navigate around the meeting room.	2D	-	-	-	-	3.9	0.74	3.86	0.0019
		3D	-	-	-	-	4.4	0.97	4.58	0.0007

Table 4. Mann-Whitney Test results. '-' indicates the results were not significant.

#	Question	Automatic Mode			User Controlled Mode		
		z	p	3D Better?	z	p	3D Better?
1	It was easy to tell who the speaker was.	0.45	0.326	-	-0.8	0.203	-
2	It was easy to tell who the speaker was looking at.	-1	0.154	-	-1.7	0.041	Yes
3	It was difficult to tell who the speaker was talking to.	-0.6	0.284	-	1.55	0.061	Yes α=0.10
4	It was easy to tell who was interrupting the speaker.	0.38	0.352	-	-0.1	0.456	-
5	It was difficult to tell who was going to be the next speaker.	-0.5	0.326	-	1.1	0.136	-
6	It was difficult to tell the participants' seating arrangement.	2.83	0.002	Yes	3.33	0.000	Yes
7	It was difficult to see all of the participants' faces.	-1.8	0.038	No	0	0.500	-
8	It was easy to follow speaker transitions.	0.76	0.224	-	-0.1	0.468	-
9	It was hard to focus on a non-speaking participant.	-1.4	0.087	No α=0.10	0.6	0.274	-
10	It was useful to navigate around the meeting room.				-1.5	0.066	Yes α=0.10

Table 5. Closing questionnaire results.

#	Question	Mean	MD	STD	t(19)	p
1	I liked the spatial arrangement of participants in the viewer.	3.7	4	1.03	3.04	0.0034
2	I think spatial arrangement is a better experience than current systems (2D).	4.2	4	0.89	6.00	0

Q10: Useful to navigate around the meeting room? Since the automatic mode did not have any navigation capabilities, this question was not asked for that mode. For the user-controlled mode, the Mann-Whitney test showed that 3D did better (*z=-1.51, p=0.066*).

4.5 Closing Questionnaire and Debrief

After the second task in our interface evaluation, the participants filled out a closing survey containing questions about the spatial interface and discussed their general impression about both the interfaces with the study administrator. We report the results from this survey in Table 5. As Table 5 shows, the participants overwhelmingly agreed they liked the 3D interface, and that the interface was a step in the right direction for improving the traditional 2D interface. Some specific comments were

> *"Very contextual, can relate to what's being attempted in this project. I deal with a lot of distributed teams and anything that works and scales well is appreciated."*

> *"Gives a feeling of being in the room."*

However, there were clearly still some issues with the 3D interface. One of the key observations we made during the study and the debrief discussion was the fact that most of the participants did not like the frequent automatic rotation to the current speaker in the 3D interface. Some of the participants' comments were:

"My head started spinning."

"I was very distracted by the rotations. Please don't change the positions too often."

"Hard to focus on non-speaking participant because of distractions (many videos, rotations), too much attention is given in the interface to the current speaker, even in the 2D interface."

"In 3D view whatever orientation I tried, one of the faces would be so inclined that I couldn't see it properly."

The participants also offered a number of encouraging comments with regard to the 3D interface and suggestions to improve it.

"Like to use mouse wheel to change the relative sizes of the views."

"I would like to be able to control what/how I see but only until I know I have a good view, once set, I don't like anything to change at all."

"Interface needs to scale well with number of people. I often go to meetings where there is room full of people, some even standing."

Finally, the participants suggested that multiple simultaneous speakers case can be handled by using picture-in-picture techniques, showing multiple videos in the speaker view, adding a uniform background to all the videos [3], and showing a perspective cone from the thumbnail video to the large video to indicate the position at the table from which the video came from.

5. GUIDELINES FOR FUTURE SYSTEMS

Based on the study participants' comments and the study results, we extract several guidelines for future meeting viewing systems. First, these systems need to satisfy the two high-level requirements we derived: 1) the interface should correctly convey to a user who a person in the meeting is looking at, and 2) the context view should allow the user to focus on any person in the meeting or any part of the meeting room.

In order to satisfy the first requirement, the spatial relationship among the attendees must be captured. Such a capture should adapt to the infrastructure available. For example, when an omni-directional camera is available, the capture consists of a 360 degree view of the room which can then be presented easily on a curve-like surface to convey the spatial relationship as opposed to showing a flat panoramic view. When only laptop mounted cameras are available, other methods must be used to capture the spatial relationships [4][6]. From our discussion with the study participants, creating a smooth background for the meeting attendees helps convey the fact that they are all in the same room. Our 3D interface does not have such a background, which made it slightly unnatural. With state-of-the-art computer graphics and vision algorithms, creating a smooth background is possible. In fact, it is even possible to entirely replace the background of the meeting room with a virtual one [8].

The second requirement can be met by providing various navigation controls to the users so they can choose to focus on any part of the meeting space. Such controls must be kept optional because some users wish to watch recordings without much manual navigation, while others prefer to follow the meetings more closely by carefully watching different parts of the meeting. Some study participants wanted additional controls, such as the ability to drag and arrange the videos at different positions in the interface similar to the way files on a computer desktop can be arranged. While this is a good idea, care must be taken to ensure that the spatial relationship among the videos is maintained.

Meeting viewing systems must represent the speaker oriented information prominently. For example, most interfaces contain a speaker view that always shows the current speaker. However, there are some meetings where more than one person may speak at the same time. The viewing systems should be able to accommodate this. One way of handling this situation is to show two videos in the speaker view which would also avoid the need to constantly switch the speakers as in [9]. In fact, one of the key learning from our study was that frequent movement in the interface is highly distracting and discourages the use of the interface.

Moreover, speaker-oriented information consists not only of the speaker, but also of the person the speaker is looking at, the person interrupting the speaker, and any side conversations occurring while the speaker is talking. Viewer interface must be able to represent these aspects as well. For example, videos of non-speaking attendees can be highlighted when they interrupt the speaker or "talk" bubbles can be shown above their videos using a comic-book metaphor.

Finally, the interface should be able to easily add additional modalities like whiteboard captures, transcripts, and gestures.

6. RELATED WORK

There have been many meeting viewing systems developed in the past. Some focus on live video conferencing scenarios while others, like us, focus on browsing recorded meetings. Because of lack of space, we cannot discuss all of them. Instead, we discuss only those most relevant to our work.

The most relevant system to ours is the system described in [3], which uses an omni-directional camera to capture the meeting and present a panoramic view of the room in the context view. As mentioned above, one issue with this system is that it is difficult to tell who an attendee is looking at in the panoramic view.

A number of previous studies have considered the speaker-oriented information, such as who a speaker is looking at, for live video conferencing systems. Examples include a study on the importance and effectiveness of gaze [10], systems such as GAZE-2 [11], Hydra [12], and MultiView [13]. All of them consider a conferencing scenario in which the attendees are physically distributed across different locations, which is different from our scenario in which the attendees are collocated and the users view the meeting recordings offline. Nevertheless, they have all identified the importance of gaze awareness, which is the knowledge of the direction in which an attendee is looking. Accurate gaze is very important for live video conferencing. We show a different result. In particular, we show that when viewing a recording of a meeting in which the attendees are collocated, it is also important to know who an attendee is looking at. However, unlike live conferencing systems that require substantial infrastructure (projectors, multiple cameras per participant, half-silvered mirrors, etc.) to accurately convey gaze, we show that in the case of recorded meetings, it is possible to do so with little or no infrastructure.

The user study most relevant to ours was done by Rui et al [3], who as mentioned above, developed a system for viewing meetings recorded using omni-directional camera. The context view in their system displayed a panoramic 360 degree view of the meeting room, and an automatic speaker view. They found

that the users liked the overview window and our results agree with theirs. However, they did not study the importance of speaker-oriented information nor did they explore richer navigation controls that we found users desired to have, such as panning, tilting, zooming, and rotating. Although some spatial information can be added to 2D interface by carefully stitching the panoramic videos in a way that the attendees' videos are always arranged on the correct side of the enhanced speaker video, it is still hard to figure out who exactly the speaker is looking at. This is because a 2D interface can only give the direction in which the speaker is looking (e.g. left, right, straight) and not the actual person. To convey the direction correctly, the videos need to be angled, which then leads to a 3D style interface.

A large body of work has focused on meeting browsing. These systems provide either meeting summaries [14][15] or indices into the meeting, often based on the meeting transcripts [16], whiteboard capture [17], and various other modalities. This work is orthogonal to our work. In particular, when the user decides to view a part of a meeting using, for example, a transcript-based index, the meeting system must then replay the videos of the attendees for that part of the meeting. We focus on the presentation of the videos, which is necessary, both when viewing an entire meeting or just browsing the meeting.

Another issue identified by previous research [18][19] is privacy. There are a number of ways to address privacy concerns, such as informing the meeting participants of the fact that the meeting will be recorded, the quality of the recorded data, the locations of recording devices, and the potential distribution and viewership of the recorded content. The mitigation of privacy concerns is orthogonal to our work.

7. CONCLUSIONS AND FUTURE WORK

In today's busy work environments, many people cannot attend all of the meetings to which they are invited because of travel and scheduling issues. When travel is the issue, people have turned to remotely attending meetings. When scheduling is the issue, the only option available is to view a recording of the missed meeting. However, viewing recorded meetings is not popular today. One of the main reasons is a poor viewing experience. This paper focuses on improving the experience, and thus, increasing the popularity of viewing remote meetings.

Our contributions can be described at a number of levels. Our highest-level contribution is identifying and verifying through a user study that (1) speaker-oriented information, such as who the speaker is looking at, talking to, and being interrupted by and (2) navigation-oriented requirements, such as the ability to focus on any person in the recording, are important for the person viewing the meeting. The next lower level contribution are the recommendations for future meeting viewing systems we have identified based on these requirements and the user study results. The lowest level contribution of our work is the 3D interface we have described that captures the spatial arrangement of the people in the meeting. As the results of the user study show, such an interface better meets the speaker oriented and navigation oriented requirements.

In our future work, we would like to apply what we have learned to develop a system that captures the meetings using an ad-hoc array of cameras and microphones available in the room and automatically provides calibration information to the viewer. We also like to adapt the system to live conferencing and support multiple simultaneous remote attendees.

8. ACKNOWLEDGEMENTS

This research was funded in part by Microsoft and NSF grants ANI 0229998, EIA 03-03590, IIS 0312328, and IIS 0712794.

9. REFERENCES

[1] Cutler, R., et al. "Distributed meetings: a meeting capture and broadcasting system," 2002. MM.

[2] Foote, J., et al. "Immersive Conferencing Directions at FX Palo Alto Laboratory," 2004. MLMI.

[3] Rui, Y., Gupta, A. and Cadiz, J. "Viewing meeting captured by an omni-directional camera," 2001. CHI.

[4] Hartley, R. and Zisserman, A. *Multiple View Geometry in Computer Vision.* s.l. : Cambridge University Press, 2004.

[5] "Photosynth," [Online] Microsoft Live Labs, 2008. http://labs.live.com/photosynth/.

[6] Liu, Z., et al. "Energy-based sound source localization and gain normalization for ad hoc microphone arrays," 2007. ICASSP.

[7] Strubbe, H and Lee, M S. "UI for a videoconference camera," Seattle, Washington : ACM, 2001.

[8] Ahmed, R., Karmakar, G.C. and Dooley, L.S. "Automatic video background replacement using shape-based probabilistic spatio-temporal object segmentation," 2007. ICICS.

[9] Bianchi, M. "Automatic video production of lectures using an intelligent and aware environment," 2004. MUM '04.

[10] Garau, M, et al. "The impact of eye gaze on communication using humanoid avatars," 2001. CHI.

[11] Vertegaal, R., Weevers, I. and Sohn, C. "GAZE-2: an attentive video conferencing system," 2002. CHI.

[12] Sellen, A. and Buxton, B. "Using Spatial Cues to Improve Video Conferencing," s.l. : CHI, 1992.

[13] Nguyen, D. and Canny, J. "MultiView: Improving Trust in Group Video Conferencing Through Spatial Faithfulness," 2007. CHI.

[14] Chiu, P., et al. "LiteMinutes: an Internet-based system for multimedia meeting minutes," 2001. WWW.

[15] Waibel, A, et al. "Meeting Browser: Tracking and Summarizing Meetings," 1998. Broadcast News Transcription and Understanding Workshop.

[16] Wellner, P., Flynn, M. and Guillemot, M. *Browsing Recorded Meetings With Ferret.* IDIAP Research Institute. 2004.

[17] He, L. and Zhang, Z. "Real-time whiteboard capture and processing using a video camera for teleconferencing," 2005. ICASSP.

[18] Adams, Anne. "Multimedia information changes the whole privacy ballgame," 2000. Computers, Freedom and Privacy: Challenging the Assumptions.

[19] Olson, J, Grudin, J and Horvitz, E. "A study of preferences for sharing and privacy," 2005. CHI.