Distributed Meetings: A Meeting Capture and Broadcasting System

Ross Cutler, Yong Rui, Anoop Gupta, JJ Cadiz
Ivan Tashev, Li-wei He, Alex Colburn, Zhengyou Zhang, Zicheng Liu, Steve Silverberg
Microsoft Research
One Microsoft Way, Redmond, WA, USA
{rcutler, yongrui, anoop, jjcadiz, ivantash, lhe, alexco, zhang, zliu, v-ssilve}@microsoft.com

ABSTRACT
The common meeting is an integral part of everyday life for most workgroups. However, due to travel, time, or other constraints, people are often not able to attend all the meetings they need to. Teleconferencing and recording of meetings can address this problem. In this paper we describe a system that provides these features, as well as a user study evaluation of the system. The system uses a variety of capture devices (a novel 360º camera, a whiteboard camera, an overview camera, and a microphone array) to provide a rich experience for people who want to participate in a meeting from a distance. The system is also combined with speaker clustering, spatial indexing, and time compression to provide a rich experience for people who miss a meeting and want to watch it afterward.

General Terms
Algorithms, Measurement, Performance, Design, Experimentation, Human Factors.

Keywords
360 degree video, microphone array, meeting capture, meeting indexing, teleconferencing

1. INTRODUCTION
Meetings are an important part of everyday life for many workgroups. Often, due to scheduling conflicts or travel constraints, people cannot attend all of their scheduled meetings. In addition, people are often only peripherally interested in a meeting such that they want to know what happened during it without actually attending; being able to browse and skim these types of meetings could be quite valuable.

This paper describes a system called Distributed Meetings (DM) that enables high quality broadcasting and recording of meetings, as well as rich browsing of archived meetings. DM has a modular design and can use combinations of a variety of input devices (360º camera, overview camera, whiteboard capture camera, and microphone array) to capture meetings. For live meetings, the system broadcasts the multimedia meeting streams to remote participants, who use the public telephone system for low-latency duplex voice communication. The meetings can also be recorded to disk and viewed on-demand. Post-processing of recorded meetings provides on-demand viewers with indexes of the whiteboard content (e.g., jump to when this was written) and speakers (e.g., only show me the parts when this person speaks). On-demand viewers can also use time compression to remove pauses in the meeting and speed up playback without changing the audio pitch of the speakers.

While the DM system is designed to support remote viewing of meetings as they occur and viewing of meetings after they have finished, most of the recent work on the DM system focuses on the latter scenario. Thus, this paper focuses primarily on recording meetings and providing rich functionality for people who watch these recordings after the fact.

The rest of the paper is organized as follows: Section 2 describes a typical scenario for how we envision the DM system being used. Section 3 gives a brief overview of related work in terms of existing systems, capturing devices and associated software. Section 4 presents in detail the hardware equipment and software modules used in the system. System performance is detailed in Section 5. The system was tested by 10 groups who had their meetings recorded using the system; this test and its results are described in Section 6. Conclusions and future work are given in Section 7.

2. SCENARIO
This section describes a scenario of how we envision people utilizing the DM system to record, broadcast, and remotely participate in meetings.

Fred needs to schedule a meeting for this week to discuss the status of a current project. He checks everyone's calendars and tries to find an open time, but there is no common free time during which everyone can meet. However, he finds an hour when only one person, Barney, cannot make it. He decides to schedule the meeting during that time, and he lets Barney know that he will be able to watch it afterward.

Fred sends out the meeting request using Microsoft Outlook. The meeting request includes the DM-enabled meeting room as a scheduled resource. When Fred shows up for the meeting, he walks over to the DM kiosk (Figure 2) and touches the "record a meeting" button on the screen. Because Fred's meeting request included the meeting room, the kiosk automatically fills in the meeting description and participants.

Betty is working in an office on the other side of the corporate campus and receives an Outlook reminder about the meeting. She needs to attend the meeting, but does not want to commute to and from the meeting. So she clicks a link in the notification to view the broadcast from the meeting, and calls in to the meeting room to establish an audio link.

Wilma and Dino receive the notification and come to the meeting. On the way to the meeting, Dino realizes that Pebbles might be

able to help address a few of the tough issues the team is trying to solve. Pebbles agrees to attend. As she walks in the room, she swipes her employee cardkey on a reader next to the kiosk; the system adds her as a participant to the meeting.

During the meeting, Betty is able to see a panoramic image of the meeting, a higher resolution image of the current speaker, an overview of the room from a camera in one of the top corners, and an image of the whiteboard. Betty asks about the status of the project implementation. Wilma draws a few implementation diagrams on the whiteboard, which gets erased several times during the discussion of various components. Toward the end of meeting, Fred writes several action items on the whiteboard to summarize the meeting. At the end of the meeting, Fred presses the "stop recording" link on the kiosk. The Meeting Archive Server processes the recording and sends email to all of the meeting attendees with a URL that points to the archived meeting.

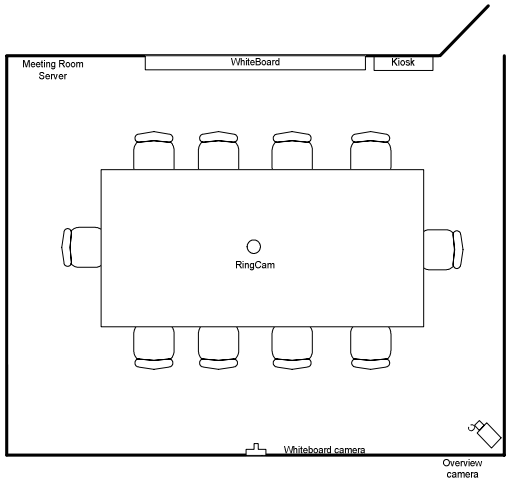

Figure 1: DM room diagram, which contains a RingCam, whiteboard and overview camera, meeting room server and kiosk.

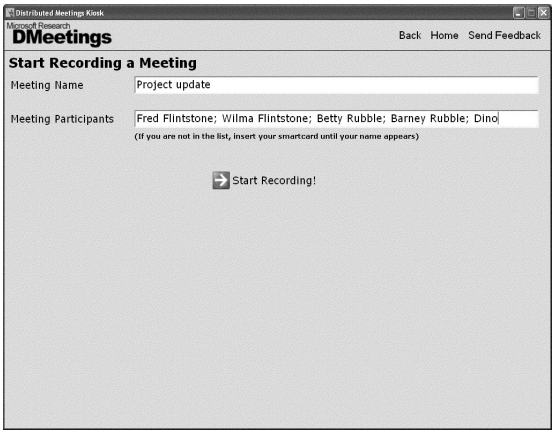

Figure 2: The DM kiosk is used to control the DM system in the meeting room.

Later that day, Barney gets back to his office and sees the e-mail about the recorded meeting. He clicks the link in the mail to start the archived meeting viewing client (Figure 3). While watching the meeting, he uses time compression to view the meeting faster. Barney also uses the whiteboard chapter frames to jump directly to the discussion on the implementation, and then clicks individual strokes on the whiteboard to listen to the detailed conversation on each specific point. He has yet to attend a meeting where Dino says anything intelligible, so in the speaker timeline, he unchecks Dino so that the client skips all the times he talks. Fred often makes good points but then talks about random things afterward. When Fred does this, Barney uses the timeline to see where Fred stops talking and skips to that point. With these features, Barney is able to view the meeting in much less time than would have been required to attend the meeting in person.

Figure 3: Distributed Meetings archived meeting client: Panorama window (bottom), speaker window (upper left), whiteboard (upper right), timeline (bottom).

3. RELATED WORK

While the focus of this paper is on recording meetings and watching them afterward, a considerable overlap exists between this domain and the domain of live teleconferencing. For example, both require audio-visual capturing equipment, and both can use sound source localization (SSL) to track the person who is speaking. Today, a variety of live teleconferencing systems are available commercially from PolyCom (including PictureTel) [13][14], Tandberg [23], and Sony, among others. Given the similarity of these products, we primarily focus on PolyCom/PictureTel's systems. We review related work in capturing devices, the associated software, and existing meeting recording systems.

3.1 Capturing Devices

Capturing devices tend to focus on four major sources of data that are valuable for videoconferencing and meeting viewing: video data, audio data, whiteboard marks, and documents or presentations shown on a computer monitor. Given that software solution exists to share documents, we focus on the first three in this section.

3.1.1 Video Capturing Devices

Three different methods exist to capture video data: pan/tilt/zoom (PTZ) cameras [13], mirror-based omni-directional cameras [19] , and camera arrays [6]. While PTZ cameras are currently the most popular choice, they have two major limitations. First, they can only capture a limited field of view. If they zoom in too fast, the context of the meeting room is lost; if they zoom out too much,

people's expressions become invisible. Second, because the controlling motor takes time to move the camera, the camera's response to the meeting (e.g., switching between speakers) is slow. In fact, the PTZ cameras cannot move too much, otherwise people watching the meeting can be quite distracted.

Given these drawbacks and recent technological advances in mirror/prism-based omni-directional vision sensors, researchers have started to rethink the way video is captured and analyzed [5]. For example, 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 mirror-based omni-directional system was also used in our previous system for meeting capturing and viewing [19]. While this approach overcomes the two difficulties faced by the PTZ cameras, these type of devices tend to be too expensive to build given today's technology and market demand. For example, our previous system cost approximately $7,000. Although the cost may have dropped to $3000 today, it remains a very expensive capturing device. In addition, the mirror omnicams suffer from low resolution (even with 1MP sensors) and defocusing problems, which result in inferior video quality.

Multiple inexpensive cameras can be assembled to form an omni-directional camera array. For example, in [6] four NTSC cameras are used to construct a panoramic view of the meeting room. Two important features distinguish this design from the design (RingCam) described in this paper. First, NTSC cameras provide a relatively low quality video signal. In addition, the four cameras require four video capture boards to digitize the signal before it can be analyzed, transmitted or recorded. In contrast, we use five IEEE 1394 cameras that provide superior video quality and only require a single 1394 card. Second, the RingCam integrates a microphone array, used for sound source localization and beamforming.

3.1.2 Audio Capturing Devices

Capturing high-quality audio in a meeting room is challenging. The audio capturing system needs to remove a variety of noises, remove reverberation, and adjust the gain for different levels of input signal. In general, there are three approaches to address these requirements. The simplest approach is to use close-up microphones, but it is cumbersome. Placing a microphone on the meeting room table to prevent multiple acoustic paths is currently the most common approach, e.g., PolyCom's VoiceStation series and Digital Microphone Pad [14]. These systems use several (usually three) hypercardioid microphones to provide omni-directional characteristics. The third approach is provided in PictureTel's desktop teleconferencing system iPower 600 [13]. A unidirectional microphone is mounted on top of a PTZ camera, which points at the speaker. The camera/microphone group is controlled by a computer that uses a separate group of microphones to do sound source localization. This approach, however, requires two separate sets of microphones.

For the DM system, instead of using several directional microphones with complex construction to provide 360^{o} acoustic capture, a microphone array with omni-directional microphones is used. This solution allows the system to capture the audio signal from around the meeting room, use sound source localization to find the direction of the speaker, and beam-form to enhance the sound quality. This solution is a seamless integration of the last two solutions with low-cost hardware.

3.1.3 Whiteboard Capturing Device

Many technologies have been created to capture the whiteboard content automatically. One category of whiteboard capture devices captures images of the whiteboard directly. One of the earliest of whiteboard capture technologies, the whiteboard copier from Brother [2], is a special whiteboard with a built-in copier. With a click of a button, the whiteboard content is scanned and printed. Video cameras are also used, e.g., ZombieBoard system at Xerox PARC [20] and the Hawkeye system from SmartTech [21].

A second category of whiteboard capture devices track the location of the pen at high frequency and infer the content of the whiteboard from the history of the pen coordinates. Mimio [11] is one of the best systems in this category. Since the history of the pen coordinates is captured, the content on the whiteboard at any given moment can be reconstructed later. The user of whiteboard recording can play back the whiteboard like a movie. The content is captured in vector form so it can be transmitted and archived with low bandwidth and storage requirement.

But the pen tracking devices have several inherent disadvantages: 1) People have to use special dry-ink pen adapters, which make them much thicker, and press the pens harder; 2) If the system is not on or the user writes without using the special pens, the content cannot be recovered by the device; 3) Many people often use their fingers to correct small mistakes on the whiteboard in stead of the special eraser. This common behavior causes extra strokes to appear on the captured content; 4) Imprecision of pen tracking sometimes causes misregistration of adjacent pen strokes.

An image-based whiteboard capture device does not have these problems. In addition, an image-based system captures the context of the whiteboard (e.g., who is writing, and pointing gestures). Our system uses a high-resolution digital camera to capture a whiteboard image sequence. By using intelligent algorithm to analyze the image sequence, time stamps of the strokes and key frames can be automatically computed.

3.2 Speaker Detection Techniques

Knowing who is talking and where that person is located are important for both live teleconferencing and meeting recording scenarios. For example, if a PTZ camera is used, the system can direct the camera to focus on the correct person. If an omni-directional camera is used, the system can cut directly to that person. All commercial VTC systems we are aware of use only audio-based SSL to locate the speaker. While this approach works most of the time, it has two limitations. First, its spatial resolution is not high enough. Second, it may lose track and point to the wrong direction due to room noise, reverberation, or multiple people talking at the same time. The DM system uses both audio-based SSL and vision-based person tracking to detect speakers, which results in higher accuracy.

3.3 Meeting Recording Systems

There has been recent interest in automatic meeting recording systems, e.g., from FX PAL [4], Georgia Tech [16], and PolyCom's StreamStation [15]. The former two mainly focus on recording slides, notes and annotations. In addition, they study more on the UI side of the system instead of the technology, e.g., how

to identify who is talking when in the meeting. Two features distinguish our system from theirs. First, we not only record notes and drawings on the whiteboard, we also capture rich 360º video and audio. Second, we focus on the technology, in addition to the UI, that can enable the rich meeting indexing, e.g., robust person tracking and SSL.

StreamStation [15] is a simple extension of PolyCom's live teleconferencing system in the recording domain. Little has been done to construct the rich indexes as we will present in this paper. There also exist web-based conferencing systems such as WebEx [24], though their meeting playback experience is extremely limited.

4. SYSTEM OVERVIEW

An overview of the DM system is shown in Figure 4. We describe the hardware and software components below.

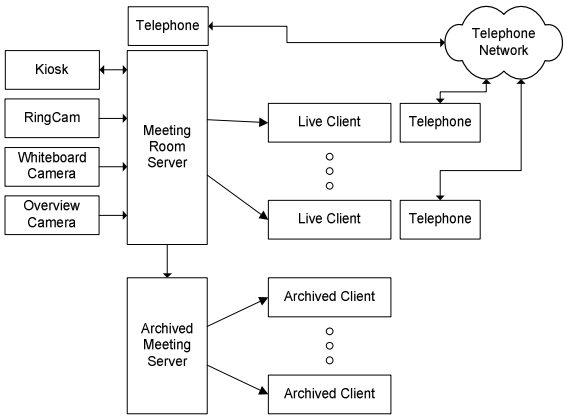

Figure 4: The DM architecture. Meetings are captured and broadcasted by the meeting room server, and stored for offline access by the archived meeting server.

4.1 Hardware Overview

4.1.1 RingCam

The RingCam (see Figure 5) is an inexpensive 360º camera with integrated microphone array. A 360º camera placed in the center of a meeting table generally provides a better viewpoint of the meeting participants than a camera placed in the corner or side of the room. By capturing a high resolution panoramic image, any of the meeting participants can be viewed simultaneously, which is a distinct advantage over traditional PTZ cameras. Because there are no moving parts, the RingCam is also less distracting than a PTZ camera.

Before developing the RingCam, we used a single sensor omnicam with a hyperbolic mirror [19] to first determine the utility of such cameras for meeting capture. As discussed in Section 3, these types of cameras currently suffer from insufficient resolution and high costs. The RingCam, on the other hand, outputs a 3000x480 panoramic image, which provides sufficient resolution for small to medium size meeting rooms (e.g., a 10x5' table).

The RingCam consists of five inexpensive ($60) 1394 board cameras arranged in a pentagonal pattern to cover a 360º horizontal and 60º vertical field of view. The individual images are corrected for radial distortion and stitched together on the PC using an image remapping table. The video is transmitted to the meeting server via a 1394 bus.

At the base of the RingCam is an 8-element microphone array used for beamforming and sound source localization. The microphone array has an integrated preamplifier and uses an external 1394 A/D converter (Motu828) to transmit 8 audio channels at 16-bit 44.1KHz to the meeting room server via a 1394 bus.

The RingCam was primarily designed for capturing meetings. Some of the design goals of RingCam include the following:

(1) The camera head is sufficiently high from the table so that a near frontal viewpoint of the participants can be imaged, but low enough not to be obtrusive to the meeting participants.

(2) The microphone array is as close to the table as possible so that sound reflections from the table do not complicate audio processing.

(3) The camera head and microphone array are rigidly connected to allow for fixed relative geometric calibration.

(4) The rod connecting the camera head and microphone array is thin enough to be acoustically invisible to the frequencies of human speech (200-4000Hz).

(5) The camera and microphone array has a privacy mode, which is enabled by turning the cap of camera. In this mode, the cameras are optically occluded and the microphone preamp is powered off. Both a red light on top of the camera and the meeting room kiosk indicates whether the camera is in privacy mode.

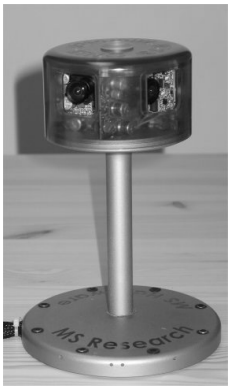

Figure 5: RingCam: an inexpensive omnidirectional camera and microphone array designed for capturing meetings.

4.1.2 Whiteboard Camera

The DM system uses a digital still camera to capture whiteboards. The whiteboard camera is a consumer-level 4MP digital still camera (Canon G2), which takes images of the whiteboard about once every five seconds. The images are transferred to the meeting room server via a USB bus as a compressed MJPEG image.

4.1.3 Overview camera

The overview camera is used to provide a view of the entire meeting room. It is currently used by meeting viewers, but in the future it could also be used to automatically detect events such as a person entering or exiting the room, or a person pointing to the whiteboard. The overview camera used is 640x480 15FPS, with a 90º HFOV lens. The camera is connected to the meeting room server via a 1394 bus.

4.1.4 Meeting Room Server

A dual CPU 2.2GHz Pentium 4 workstation is used for the meeting room server. It uses a 15" touchscreen monitor, and a keycard reader for user authentication. It interfaces with the live clients with a 100Mbps Ethernet network.

4.1.5 Archived Meeting Server

A dual CPU 2.2GHz Pentium 4 workstation is used for the archive meeting server. It interfaces with the archived clients with a 100Mbps Ethernet network, and contains a RAID to store the archived meetings.

4.2 Software Overview

4.2.1 Meeting Room Server

The meeting room server performs all processing required to broadcast and record meetings. A dataflow of the meeting room server is shown in Figure 6. The input devices are the RingCam, overview camera, whiteboard camera, and microphone array (Motu828). The server runs Windows XP and is implemented using C++ and DirectShow. Each node in the dataflow is a DirectShow filter and is described below.

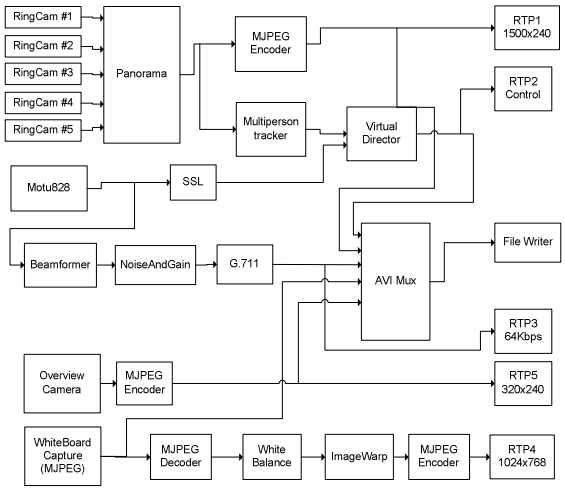

Figure 6: Meeting Room Server dataflow

4.2.1.1 Panorama Stitcher

The Panorama filter takes five video stream inputs (each 320x240 15FPS) from the RingCam and outputs a single panorama image of size 1500x240 (3000x480 is possible in full resolution mode, but requires additional computation). Since each camera uses a wide field of view lens, the images have significant radial distortion. The radial distortion model used is [10]:

$$x_u = x_d + x_d \sum_{i=1}^{\infty} \kappa_i R_d^i ; y_u = y_d + y_d \sum_{i=1}^{\infty} \kappa_i R_d^i$$

where the κ's are the radial distortion parameters, (x_u, y_u) is the theoretical undistorted image point, (x_d, y_d) is the measured distorted image point, and $R_d = x_d^2 + y_d^2$. We use a calibration pattern to determine the first 5 radial distortion parameters, and correct for the radial distortion.

The images are then transformed into cylindrical coordinates, and the translation and scaling between each pair of adjacent cameras is determined. The cylindrical mappings are then combined to form a panoramic image, cross-fading the overlapping regions to improve the panoramic image quality [22]. The images are corrected for vignetting and color calibrated to further enhance the panoramic image quality (see [12] for details). All of these operations (radial distortion correction, cylindrical mapping, panoramic construction, cross-fading, devignetting) are combined into a single image remapping function for computational efficiency.

4.2.1.2 Sound Source Localization

In the DM context, the goal for sound source localization is to detect which meeting participant is talking. The most widely used approach to SSL is the generalized cross-correlation (GCC). Let $s(n)$ be the source signal, and $x_1(n)$ and $x_2(n)$ be the signals received by the two microphones:

$$\begin{aligned} x_1(n) &= as(n-D) + h_1(n) * s(n) + n_1(n) \\ x_2(n) &= bs(n) + h_2(n) * s(n) + n_2(n) \end{aligned} \quad (1)$$

where D is the time delay of the signal arriving at the two microphones, a and b are signal attenuations, $n_1(n)$ and $n_2(n)$ are the additive noise, and $h_1(n)$ and $h_2(n)$ represent the reverberations. Assuming the signal and noise are uncorrelated, D can be estimated by finding the maximum GCC between $x_1(n)$ and $x_2(n)$:

$$D = \arg\max_{\tau} \hat{R}_{x_1 x_2}(\tau)$$

$$\hat{R}_{x_1 x_2}(\tau) = \frac{1}{2\pi} \int_{-\pi}^{\pi} W(\omega) G_{x_1 x_2}(\omega) e^{j\omega\tau} d\omega$$

where $\hat{R}_{x_1 x_2}(\tau)$ is the cross-correlation of $x_1(n)$ and $x_2(n)$, $G_{x_1 x_2}(\omega)$ is the Fourier transform of $\hat{R}_{x_1 x_2}(\tau)$, i.e., the cross power spectrum, and $W(\omega)$ is the weighting function.

In practice, choosing the right weighting function is of great significance for achieving accurate and robust time delay estimation. As can be seen from equation (1), there are two types of noise in the system, i.e., the background noise $n_1(n)$ and $n_2(n)$ and reverberations $h_1(n)$ and $h_2(n)$. Previous research suggests that the maximum likelihood (ML) weighting function is robust to background noise and phase transformation (PHAT) weighting function is better at dealing with reverberations:

$$W_{ML}(\omega) = \frac{1}{|N(\omega)|^2}, \qquad W_{PHAT}(\omega) = \frac{1}{|G_{x_1 x_2}(\omega)|}$$

where $|N(\omega)|^2$ is the noise power spectrum.

It is easy to see that the above two weighting functions are at two extremes. That is, $W_{ML}(\omega)$ puts too much emphasis on "noiseless" frequencies, while $W_{PHAT}(\omega)$ completely treats all the frequencies equally. To simultaneously deal with background noise and reverberations, we have developed a technique that integrates the advantages of both methods:

$$W(\omega) = \frac{1}{\gamma |G_{x_1 x_2}(\omega)| + (1-\gamma) |N(\omega)|^2}$$

where $\gamma \in [0,1]$ is the proportion factor. Once the time delay D is estimated from the above procedure, the sound source direction α can be estimated given the microphone array's geometry:

$$\alpha = \arcsin \frac{D \times v}{|AB|}$$

where D is the time delay, |AB| is the distance between the two microphones, and v = 342 m/s is the speed of sound traveling in air.

4.2.1.3 Person Detection and Tracking

Although audio-based SSL can detect who is talking, its spatial resolution is not high enough to finely steer a virtual camera view. In addition, occasionally it can lose track due to room noise, reverberation, or multiple people speaking at the same time. Vision-based person tracking is a natural complement to SSL. Though it does not know who is talking, it has higher spatial resolution and tracks multiple people at the same time.

However, robust vision-based multi-person tracking is a challenging task, even after years of research in the computer vision community. The difficulties come from the requirement of being fully automatic and being robust to many potential uncertainties. After careful evaluation of existing techniques, we implemented a fully automatic tracking system by integrating three important modules: auto-initialization, multi-cue tracking and hierarchical verification [18].

1. **Auto-Initialization:** We use three different ways to achieve auto-initialization: when there is motion in the video, we use frame difference to decide if there are regions in the frame that resemble head-and-shoulder profiles; when there is audio, we use SSL to initialize the tracker; when is neither motion nor audio, we use a state-of-the-art fast multi-view face detector [18] to initialize the tracker.

2. **Hierarchical Verification:** No vision-based tracker can reliably track objects all the time. Each tracked object therefore needs to be verified to see if the tracker has lost track. To achieve real-time performance, we have developed a hierarchical verification module. At the lower level it uses the object's internal color property (e.g., color histogram in HSV color space) to conduct a faster but less accurate verification. If a tracked object does not pass the low-level verification, it will go through a slower but more accurate high-level verification. If it fails again, the tracking system will discard this object.

3. **Multi-Cue Tracking:** Because of background clutter, single visual tracking cues are not robust enough individually. To overcome this difficulty, we have developed an effective multi-cue tracker based on hidden Markov models (HMM) [18]. By expanding the HMM's observation vector, we can probabilistically incorporate multiple tracking cues (e.g., contour edge likelihood, foreground/background color) and spatial constraints (e.g., object shape and contour smoothness constraints) into the tracking system.

Working together, these three modules achieve good tracking performance in real-world environment. A tracking example is shown in Figure 3 with white boxes around the person's face.

4.2.1.4 Beamforming

High quality audio is a critical component for remote participants. To improve the audio quality, beamforming and noise removal are used. Microphone array beamforming is a technique used to "aim" the microphone array in an arbitrary direction to enhance the S/N in that direction. For computational efficiency and low latency (compared to adaptive filters), we use delay and sum beamforming [3]. The beamformer also helps dereverberate the audio, which significantly improves the audio quality.

4.2.1.5 Noise Reduction and AGC

The audio signal is band filtered to [200,4000] Hz to remove non-speech frequencies, and a noise reduction filter removes stationary background noise (e.g., noise from projector fan and air conditioners). The gain is automatically adjusted so that speakers sitting close to the RingCam have similar amplitudes to those sitting further away. Details are provided in [9].

4.2.1.6 Virtual Director

The responsibility of the virtual director (VD) module is to gather and analyze reports from the SSL and multi-person tracker and make intelligent decisions on what the speaker window (the top left window in Figure 3) should show. Just like video directors in real life, a good VD module observes the rules of the cinematography and video editing in order to make the recording more informative and entertaining [19]. For example, when a person is talking, the VD should promptly show that person. If two people are talking back and forth, instead of switching between these two speakers, the VD may decide to show them together side by side (note that our system captures the entire 360º view). Another important rule to follow is that the camera should not switch too often; otherwise it may distract viewers.

4.2.1.7 RTP

All multimedia streams are transmitted (multicast) to live remote clients via the Real-Time-Protocol [17].

4.2.1.8 Whiteboard Processing

For live broadcasting, the images are white-balanced, cropped and a bilinear warp is used correct for a non-frontal camera viewpoint. The images are then recompressed and broadcasted to the remote participants.

For archived meetings, offline image analysis is performed to detect the creation time for each pen strokes. Further analysis is performed to detect key frames, which are defined as the whiteboard image just before a major erasure happens. See [8] for more details about whiteboard processing done in DM.

4.2.1.9 Speaker Segmentation and Clustering

For archived meetings, an important value-added feature is speaker clustering. If a timeline can be generated showing when each person talked during the meeting, it can allow users to jump between interesting points, listen to a particular participant, and better understand the dynamics of the meeting. The input to this preprocessing module is the output from the SSL, and the output from this module is the timeline clusters. There are two components in this system: pre-filtering and clustering. During pre-filtering, the noisy SSL output will be filtered and outliers thrown away. During clustering, K-mean's clustering is used during the first a few iterations to bootstrap, and a mixture of Gaussians clustering is then used to refine the result. An example timeline cluster is shown in the lower portion of Figure 3.

4.2.1.10 Meeting Room Kiosk

The meeting room kiosk is used to setup, start, and stop the DM system. The setup screen is shown in Figure 2. The meeting description and participants are automatically initialized using information gleaned from the Microsoft Exchange server and any schedule information that known for that meeting room at that

time. All entries can be modified and new users can be quickly added using the keycard reader attached to the system.

4.2.2 Remote Client

The DM Remote Client supports both live and asynchronous viewing of meetings. The user interface for the archived client is shown in Figure 3. The live client is similar, but does not include the timeline or whiteboard key frame table of contents.

A low resolution version of the RingCam panorama image is shown in the lower part of the client. A high resolution image of the speaker is shown in the upper left, which can either be automatically selected by the virtual director or manually selected by the user (by clicking within the panoramic image).

The whiteboard image is shown in the upper right window. Each pen stroke is timestamped, and clicking on any stroke in the whiteboard synchronizes the meeting to the time when that stroke was created. Pen strokes that will be made in the future are displayed in light gray, while pen strokes in the past are shown in their full color. Key frames for the whiteboard are displayed to the right of the full whiteboard image and provide another index into the meeting. The transparency of the current key frame and the current image can be adjusted so that remote viewers can even view pen strokes occluded by a person.

The timeline is shown in the bottom of the window, which shows the results of speaker segmentation. The speakers are automatically segmented and assigned a unique color. The person IDs have been manually assigned, though this process could be automated by voice identification. The remote viewer can select which person to view by clicking on that person's color. The speakers can also be filtered, so that playback will skip past all speakers not selected.

The playback control section to the left of the panorama allows the remote view to seek to the next or previous speaker during playback. In addition, time compression [7] can be used to remove pauses to and increase the playback speed without changing the speaker's voice pitch.

Just above the playback control is the tab control, which allows the user to display meeting information (time, location, duration, title, participants), meeting statistics (who led the meeting, number of active participants), the overview window, and whiteboard statistics.

5. SYSTEM PERFORMANCE

For a system to be of practical use, it is important to benchmark and analyze the system performance. The bandwidth per stream is summarized in Table 1 for meeting broadcasting. Each meeting takes about 2GB/hour ($4/hour) to store, most of it for the RingCam video stream. By recompressing the streams using the Windows Media 8 CODEC for asynchronous access, the storage requirements are reduced to about 540MB/hour (1.2Mbps) for similar quality as the MJPEG CODEC.

The live meeting bandwidth can be significantly reduced by not broadcasting the full resolution panoramic image, but rather a lower resolution panoramic image and a client specific speaker window. When this is combined with the Windows Media 8 CODEC, bandwidth is reduced from 4.68Mbps to under 1Mbps. The primary reasons this was not initially done are (1) Windows Media 8 requires significantly more computational processing than MJPEG; (2) Windows Media 8 has significantly more latency than MJPEG; (3) a client specific speaker window would require additional computational processing for each new additional client connected to the meeting room server. We are investigating solutions for each of these problems.

Stream	Width	Height	FPS	Compression	Mbits/s
RingCam	1500	240	15	30	4.12
Overview	320	240	7.5	30	0.44
Whiteboard	1024	768	0.1	30	0.06
Audio					0.06
				Total	**4.68**
1 Hour storage (GB)	2.06				

Table 1: DM Bandwidth per stream type

The video latency from the meeting room to a nearby remote client is approximately 250ms. The CPU usage (for both CPUs) for the DM meeting room server for each major component is given in Table 2.

Component	% CPU
Video capture	1
Audio capture	1
Beamformer	3
MJPEG	5
Noise and Gain	4
Panorama	20
Person tracker	26
SSL	6
Virtual director	1

Table 2: DM Meeting Room Server CPU utilization

The microphone array provides a high quality recording and transmission of the audio. The noise reduction provides an additional 15dB signal-to-noise ratio (S/N) for speakers. The beamformer provides a 6dB S/N enhancement compared to sources 90º from the source direction. The beamformer also helps dereverberate the audio, which provides a significant perceptual enhancement for remote listeners.

For the user study in this paper, the beamformer was used in the toroid mode, which produces a donut shape radially symmetric beam to help eliminate noise coming from above and below the RingCam (e.g., projectors and meeting room PCs).

6. USER STUDY

This section describes the method of the study used to evaluate the system, as well as the results.

6.1 Method of Study

To evaluate the DM system, a user study was conducted. The DM system was set up in a meeting room controlled by the researchers and various teams around the company were asked to hold one of their regular meetings in the DM meeting room. In addition, to test the post-meeting viewing experience, at least one person was asked to miss the meeting.

Thus, meetings recorded for this study were "normal" meetings: they were meetings among people who knew each other, and meetings that would have happened even if the team not participated in the study. The only differences were that the meeting was held in our special meeting room, the meeting was recorded, and at least one person was not present. Note, however, that often the person who missed the meeting would have done so anyway—sometimes the person was on vacation, and sometimes the person had a conflicting appointment.

All the meetings were recorded using the DM system. At the conclusion of each meeting, the attendees were asked to fill out a

brief survey asking a variety of questions about how comfortable they were having the meeting recorded and how intrusive the system was.

A few days after the meeting, the people who missed the meeting came to a usability lab to view the meeting using the viewer client described in Section 4.2.2. (The usability lab is a room with a one-way mirror and a computer instrumented such that it is easy to monitor and record all the user's actions with the computer.) Participants were given a brief introduction to the user interface and then asked to view the meeting as they would if they were in their office. Participants were told they could watch as much or as little of the meeting as they wanted.

While participants viewed meetings, the researchers observed their use of the viewer client. Most of the button presses and other interactions with the client were automatically logged, and after people finished viewing the meeting, they were asked to complete a brief survey about their experience. One of the researchers also interviewed them briefly to chat about their experience.

Question N = 10 groups	Avg	Std dev
I was comfortable having this meeting recorded.	3.9	0.7
The system got in the way of us having a productive meeting.	1.7	0.4
I felt like I acted differently because the meeting was being recorded.	3.1	1.1
It was awkward having the camera sitting in the center of the table.	3.0	0.8

Table 3: Survey responses from people who participated in meetings that were recorded. All questions were answered using the following scale: 5 = strongly agree, 4 = agree, 3 = neither agree nor disagree, 2 = disagree, 1 = strongly disagree

Note that all the meetings were captured using the first version of the RingCam and an early version of the microphone array with only four microphones.

6.2 User Study Results

Ten meetings were recorded for the user study. 34 people participated in a meeting, and eleven people missed a meeting and watched it afterward. The participants were from a variety of divisions within our company (research, product development, human resources, facilities management, etc.), and the types of meetings ranged in variety from brainstorming meetings to weekly status meetings.

Two sets of results from the user study are presented: results from the people who participated in the meetings that were recorded, and results from the people who viewed the meeting after the fact.

6.2.1 Results from Meeting Participants

Table 3 shows survey results from people who participated in meetings that were recorded. People were generally comfortable having their meetings recorded, although this could be due to self-selection (people who would not have been comfortable may have chosen not to volunteer for the study). People did not think the system got in the way of their having a productive meeting, and people also did not think it was awkward to have a camera mounted in the top corner of the room.

Question N = 11	Avg	Std dev
It was important for me to view this meeting.	3.7	0.5
I was able to get the information I needed from the recorded session.	4.6	0.5
I would use this system again if I had to miss a meeting.	4.4	0.8
I would recommend the use of this system to my peers.	4.0	0.9
Being able to browse the meeting using the whiteboard was useful	3.2	1.2
Being able to browse the meeting using the timeline was useful	4.0	0.9
Being able to speed up the meeting using time compression was useful	4.1	1.3
Being able to see the panoramic (360º) view of the meeting room was useful	4.4	0.9
Being able to see the current speaker in the top-left corner was useful	4.1	1.2

Table 4: Survey responses from people who missed a meeting and watched it afterward. All questions were answered using the following scale: 5 = strongly agree, 4 = agree, 3 = neither, agree nor disagree, 2 = disagree, 1 = strongly disagree

However, people had mixed feeling about whether they felt as if they acted differently as a result of having the system in the room. This feeling may diminish over time as people became more comfortable with the system, or it may remain as people are constantly reminded that all their words and actions are being recorded. In one meeting, one participant remarked, "I probably shouldn't say this because this meeting is being recorded, but…."

Furthermore, people were divided on whether having the RingCam sit in the middle of the table was awkward. Some participants wrote: *"Very inconspicuous"*, and *"System seemed pretty low profile – I was a little self-conscious but lost that sense quickly."* Others wrote: *"Camera head obscured other people's faces", and "The center camera was distracting"*.

However, the prototype used for the study was 14" tall while the newer prototype shown in Figure 5 is only 9". Further studies are required to determine if the newer camera is less obtrusive.

6.2.2 Results from Meeting Viewers

Survey results from people who missed a meeting and watched it afterward using the DM system are shown in Table 4. One question participants were asked is what role they would have played had they been at the meeting. One participant said he would have been the meeting leader, two said they would have been primary contributors, five said they would have been secondary contributors, and three said they would have been mostly observers (one person did not answer this question). However, of the twelve people who viewed meetings, eight "agreed" that it was important for them to view the meeting, while four "neither agreed nor disagreed."

One important note about the meetings participants watched is that the meetings often had synchronization issues between the audio, video, and whiteboard. Due to some bugs in the system at

the time of the study, it was common to have the audio and video out of sync by as much as three seconds, and the whiteboard out of sync with the audio/video by as much as a minute.

However, despite these issues, feedback from participants was quite positive. Participants said they were generally able to get the information they needed from the recorded session and that they would recommend use of the system to their peers.

Participants were also asked about specific parts of the interface to try to determine their relative usefulness. The survey data indicate that the panoramic video from the meetings was the most useful while the whiteboard browsing feature was the least useful, although this is likely because few of the meetings used the whiteboard extensively.

In addition, not all people used time compression and the timeline to browse the meeting quickly (instead of watching it beginning to end). Out of twelve participants, only three used time compression regularly to watch the meeting; however, the reasons for not using time compression varied. In once instance, time compression was not working. In another, time compression was causing the audio and video to get out of sync. In another, one person in the meeting had a heavy accent, and it was difficult to understand him when using time compression.

However, the participants who did use the timeline and time compression were very enthusiastic about these features. One participant said the timeline was, "extremely useful…the most useful part of the UI."

From observing people and talking to them afterward, several areas for improvement were discovered. First, people often requested person-specific time compression settings. In fact, one participant remarked that this was his "biggest feature" he recommended adding. Currently the time compression buttons only adjust the overall speed of the meeting, thus if one person speaks in a manner that makes it difficult to hear them with time compression (for example, if they speak very quickly or very softly, or if they speak with an accent), then the user must constantly adjust the time compression or stop using it altogether.

Second, a related feature would be a "what was that?" button that would jump the audio and video backwards 5 seconds and slow the speed down to 1.0x. As participants listened to the meeting, they often encountered points that wanted to hear again. This was especially true for participants who liked to skim the meeting using time compression.

Third, participants remarked that the audio quality of people who were speaking while standing at the whiteboard was very poor, sometimes so much so that they could not understand what the person was trying to say. However, this issue may have occurred because the study used an early version of the microphone array with only 4 microphones, as well as an earlier version of the beamforming algorithms.

Fourth, participants had ideas for a variety of features to add to the timeline. One participant wanted a way to mark important parts of the meeting by "putting gold stars on the timeline". The same participant wanted a way to highlight a portion of the meeting and e-mail a link to the portion to a colleague. Another participant wanted the system to keep track of what parts of the meeting he had already seen and show it on the timeline; with such a feature, he could skip around the meeting and not lose track of what he still needed to watch.

Interestingly, one feature participants were not enthusiastic about adding was automatic name recognition of speakers on the timeline. For some of the meetings names were entered by hand by the researchers, but to test the usefulness of the names, no names were entered for several meetings (the lines were simply labeled, "speaker 1", "speaker 2", etc.). When researchers asked participants afterward if having no names affected the usefulness of the timeline, the consensus was that because the participant knew all the people in the meeting and their voices, names were not necessary. However, some participants remarked that if the meeting was filled with strangers, names would be more important (although this scenario was not tested).

However, even though the meetings often had issues with audio and video synchronization, an older camera and microphone array were used to capture the meeting, and a version 1 interface was being used, participants' reaction to the experience was very positive. As Table 4 shows, there was a strong consensus that "I would use the system again if I had to miss a meeting" (average of 4.4 out of 5).

7. CONCLUSIONS AND FUTURE WORK

We have described a system to broadcast, record, and remotely view meetings. The system uses a variety of capture devices (360º RingCam, whiteboard camera, overview camera, microphone array), to give a rich experience to the remote participant. Archived meetings can be quickly viewed using speaker filtering, spatial indexing, and time compression. The user study of the recorded meeting scenario shows that users found the system captured the meeting effectively, and liked the panoramic video, timeline, speaker window, and time compression parts of the system.

We plan to greatly extend the capabilities of DM in many ways. First, we want to add duplex audio and video real-time communication over the intranet, so that the telephone network is not required. This is a challenging task, as it involves significantly lowering the audio/video latency, lowering the network bandwidth requirements, and adding echo cancellation suitable for microphone arrays.

For recorded meetings, we plan to enhance the meeting analysis to include human activities (such as when people enter/exit a room) and detect whiteboard pointing events (e.g., show not only when at equation was written on the whiteboard, but also when it was pointed to). The virtual director can be improved to include other video sources, such as the overview window (e.g., show the overview window when someone enters/exist the room), and to show multiple people in the speaker window (e.g., when two people are talking quickly back and forth). Speech recognition can be used to generate searchable transcripts of the meetings. Finally, we plan to use digital rights management (DRM) to provide security on the recorded meetings (e.g., this meeting can only be viewed by the participants, but cannot be copied), and to setup automatic retention policies for recorded meetings.

An open question is how a system like DM will change the attendance of meetings. For example, at Microsoft Research, live attendance of research lectures dropped significantly once they were broadcasted and recorded for on-demand viewing. If meetings are similarly broadcasted and recorded, will fewer people attend them, and if so, how would this affect the employee's and the company's overall performance?

8. ACKNOWLEDGMENTS

We thank Gavin Jancke and Lon-Chan Chu for helping develop the software, Mike Holm, Mehrdad Basseri, Dawson Yee, Vince Jesus and Ben An for helping develop the hardware components, and thank Michael Cohen and Rico Malvar for valuable discussions.

9. REFERENCES

[1] BeHere, http://www.behere.com

[2] Brother CopyPoint Whiteboard, http://www.electronicgadgetdepot.com/w/Whiteboards/

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

[4] P. Chiu, A. Kapuskar, and L. Wilcox, Meeting capture in a media enriched conference room, *Multimedia Magazine*, vol. 7, no. 4, Oct-Dec 2000, pp. 48-54.

[5] Daniilidis, K., Preface, *Proc. of IEEE Workshop on Omnidirectional Vision*, June, 12, 2000.

[6] J. Foote and D. Kimber: "FlyCam: Practical Panoramic Video," in *Proc. IEEE International Conference on Multimedia and Expo*, August 2000

[7] L. He and A. Gupta. Exploring Benefits of Non-Linear Time Compression. ACM Multimedia 2001.

[8] L. He, Z. Liu and Z. Zhang. Why Take Notes, Use the Whiteboard Capture System. Technical Report MSR-TR-2002-89, September 2002.

[9] W. Jiang and H. Malvar. Adaptive Noise Reduction of Speech Signals. Microsoft Technical Report MSR-TR-2000-86, July 2000.

[10] S.B. Kang. Radial Distortion Snakes. MVA 2000

[11] Mimo, http://www.mimio.com/index.shtml

[12] H. Nanda and R. Cutler, Practical calibrations for a real-time digital omnidirectional camera. CVPR Technical Sketch, December 2001.

[13] PictureTel, http://www.picturetel.com/

[14] PolyCom, http://www.polycom.com/

[15] PolyCom StreamStation, www.resourcemanagement.net/streamstation.htm

[16] H. Richter, W. Geyer, L. Fuchs, S. Daijavad, S. Poltrok, Integrating Meeting Capture within a Collaborative Team Environment, Proc. of the *International Conference on Ubiquitous Computing, Ubicomp 2001*, Atlanta, GA, September 2001.

[17] RTP: A Transport Protocol for Real-Time Applications. RFC1889. http://www.faqs.org/rfcs/rfc1889.html.

[18] Y. Rui and Y. Chen, Automatic Detection and Tracking of Multiple Individuals Using Multiple Cues, US patent pending, 2001.

[19] Y. Rui, A. Gupta and J.J. Cadiz, Viewing meetings captured by an omni-directional camera, *Proc. ACM CHI'2001*, Seattle, WA, April, 2001

[20] E. Saunder, Image Mosaicing and a Diagrammatic User Interface for an Office Whiteboard Scanner, *Proc. of CoBuild'99*, Pittsburgh. LNCS 1670. Springer: Heidelberg

[21] SmartTech, http://www.smarttech.com/products/shared/pdf/broch_mktg-015.pdf

[22] R. Szeliski and H.-Y. Shum. Creating full view panoramic image mosaics and texture-mapped models. Computer Graphics (SIGGRAPH'97), pages 251-258, August 1997.

[23] Tandberg, http://www.tandberg.com

[24] WebEx, http://www.webex.com