Design of a Virtual Auditorium

Milton Chen
Computer Systems Laboratory
Stanford University
miltchen@graphics.stanford.edu

ABSTRACT

We built a videoconference system called the Virtual Auditorium to support dialog-based distance learning. The instructor can see dozens of students on a tiled wall-sized display and establish eye contact with any student. Telephone-quality audio and television-quality video can be streamed using commodity codecs such as wavelet and MPEG-4. Support for stream migration allows a seamless user interface to span the multiple computers driving the display wall.

We performed user studies on the auditorium parameters. We found that the optimal display wall size must balance two contradictory requirements: subjects prefer larger videos for seeing facial expressions and smaller videos for seeing everyone without head movement. Ideally, each video should have a field of view that spans 14 degrees, which corresponds to a slightly larger than life-size image. At the very least, each video should have a field of view of 6 degrees. We found that a video window should be less than 2.7 degrees horizontally and 9 degrees vertically from the camera in order to maintain the appearance of eye contact for the remote viewer. In addition, we describe a previously unreported gaze phenomenon: a person's expectation determines his perception of eye contact under ambiguous conditions.

Keywords

Distance learning, virtual auditorium, display wall, eye contact.

1. INTRODUCTION

Teaching is an inexact art. Since the days of Socratic dialogs, most educators have believed that learning is most efficient when the instruction is tailored to the students' current understandings [4]. In a classroom, students can display verbal and visual cues to indicate their state of comprehension. Teachers learn to alter the path of instruction when looks of puzzlement, boredom, or excitement are observed. As P.W. Jackson elegantly said, "Stray thoughts, sudden insights, meandering digressions and other unpredicted events constantly ruffle the smoothness of the instructional dialog. In most classrooms, as every teacher knows, the path of educational progress could be more easily traced by a butterfly than by a bullet" [16].

Currently, the most popular synchronous distance learning method is a televised lecture with an audio back channel [29]. These systems, such as the Stanford Instruction Television Network (SITN) [31], allow remote students to see and hear the instructor, but the instructor and other students can only hear the remote students. A major disadvantage of these systems is that they often do not support dialog-based teaching [29]. We observed four SITN courses offered to both local and remote students. The class topics were Object-Oriented Programming, Computer Systems, Computer Vision, and Human-Computer Interaction. We observed that instructors employed dialog-based teaching, often stopping the lecture until questions posed to the class were answered. In one course, the instructor asked an average of nine questions per class and the local students asked an average of three questions per class. However, only one remote student from the four observed classes ever asked or answered a question.

Some studies suggest that adding video to an audio link does not significantly alter the surface structure of communication or the task outcomes [23][25][30]; however, these studies did not include the task of teaching and learning. We hypothesize that dialog-based distance teaching is possible if we allow the instructor to see the remote students and the remote students to see each other. We designed a Virtual Auditorium to test this hypothesis.

The Virtual Auditorium allows dozens of students to take a class from different locations. Each student requires a web camera and a high-speed computer network connection. Students can see the instructor and other students in a video grid on their computers. The instructor can see the students projected near life-size on a tiled wall-size display. The instructor can also establish eye contact and direct his gestures to any one student, a group of students, or the entire class.

We begin by describing previous work in the next section. Section 3 describes the auditorium environment. Section 4 describes a software architecture that uses commodity codecs and parallel decompression to mitigate the communication and computation bottlenecks of streaming many high quality AV streams. Section 4 also describes an interface that allows a single pointing device to move videos anywhere on the display wall without regard to computer boundaries. Section 5 describes two user studies on the auditorium parameters: the optimal size to display students and the maximum allowable angle between the display and the camera to achieve the appearance of eye contact for the remote viewer. Section 5 also describes a previously unreported gaze phenomenon: a person's expectation determines his perception of eye contact under ambiguous conditions. Section 6 describes the practical details of the auditorium construction and our plans for future work.

2. PREVIOUS WORK

Despite enormous development efforts in videoconferencing, it remains a challenge to link dozens of people when each person is in a different location. Commercial systems typically can link four or five sites through picture-in-picture or voice-activated switching [6]. The picture-in-picture approach merges all videos into a single video at a multipoint control unit (MCU); thus, participants can see each other, but each person is transmitted at a reduced resolution. In voice-activated switching, all videos are streamed to the MCU; the MCU then transmits the videos such that the current speaker sees the previous speaker and other people see the current speaker. The inability to choose whom to see has been observed to be unpleasant [30]. An advantage of the MCU is that the bandwidth and processing required for each participant does not increase as the number of participants increases; however, the processing requirement of the MCU makes it difficult to build. To scale beyond the limitations of the MCU, we distribute the processing requirements of a MCU to many computers.

A system that did not use a MCU was the Mbone conferencing tools [19][33]. Audio and video were multicast; thus, in theory, large-scale conferences were possible. A recent application of the Mbone tools was the AccessGrid where each node was a room that could accommodate 3 to 20 people [2]. Each node had a wall-size display illuminated by up to six projectors; however, the single computer that drove the six projectors was a potential computational bottleneck. To avoid this bottleneck, we use multiple computers to drive a multi-projector display. Our AV capture hardware was selected from the AccessGrid specification, which greatly accelerated our effort.

Three projects, Forum, Flatland, and TELEP, studied the usage pattern of conference technology. Forum broadcasted the instructor's audio and video to all students, and a student's audio was broadcasted when he pressed a button [13]. They found that instructors preferred to see students and that the press-button-to-talk usage model did not support instantaneous feedback such as laughter and applause. To support spontaneous feedback, we use high-end echo cancellation hardware and microphone headsets so that all microphones can be open at all times.

The Flatland project studied how users adapted to alternative interaction models over time when the remote students could not send audio or video [34]. They presented encouraging data showing that people could adapt to non-face-to-face interaction models; however, like Forum, they reported that instructors missed the verbal and visual feedback of a face-to-face classroom.

The TELEP project studied the effect of allowing the instructor to see the remote students [17]. TELEP could gather up to 38 headshots of remote students and presented these videos on a large screen to the instructor. One drawback of TELEP was that its streaming engine introduced a 10 to 15 second delay before the audio and video were presented to the remote audience. Round-trip audio delays exceeding 200 milliseconds are noticeable [28] and excessive audio delay can make a conferencing system difficult to use [18]. Our system supports low latency audio and video streaming.

Videoconferencing systems typically did not allow eye contact. Eye contact could be supported using one of three approaches: merging the camera and display path optically [15][26], warping the video, or mounting the camera close to the display so they appear to share the same optical path. The Hydra system used the third approach [6][30]. Each Hydra node used three 12 cm diagonal monitors that were each paired with a camera, speaker, and microphone. This configuration allowed the user to establish eye contact with any of the three remote users. Like Hydra, we also minimize the distance between the display and the camera to support eye contact; however, we map many students to one display and dynamically steer the instructor's gaze to a student.

Some research had also focused on lecture recording [1][22] and browsing [11][12]. Asynchronous learning tools derived from these research could combine with the Virtual Auditorium to provide a comprehensive distance learning environment. Distance learning efficiency was also an area of active research. The widely repeated Stanford Tutored Videotape Instruction study had shown that students who watched and discussed a lecture in small groups performed better than students who watched the lectures in class [9].

3. AUDITORIUM ENVIRONMENT

A Virtual Auditorium consists of an instructor node and up to a few dozen student nodes. The instructor node consists of a wall-sized display powered by a cluster of computers. Each student node consists of a Pentium III class PC. All nodes are connected by high-speed computer networks such as Ethernet or Internet 2.

The conceptual usage model of the Virtual Auditorium is that all participants can be seen and heard with minimal latency at all times. Unlike voice-activated switching, the Virtual Auditorium lets the user decide at whom to look. Unlike SITN [31] and FORUM [13], the Virtual Auditorium does not require a student to explicitly request the audio channel before he can be heard. Our experience with SITN as well as the findings of [13][30] suggests that keeping all channels open all the time is essential in creating spontaneous and lively dialogs.

The instructor node can also accommodate local students. The local students would be seated in front of the display wall such that the remote students appear as an extension of the local students. A complication of having local students is that the conceptual usage model of one camera capturing one person may be broken, thus potentially causing difficulties in interaction between the remote and local students.

3.1. Display Wall

The instructor can see the remote students on the display wall shown in Figure 1. The instructor can sit behind the control panel shown in Figure 2 or walk around in front of the display wall. Figure 4 shows the layout of the display wall and control panel in the auditorium.

The display wall allows the instructor to see the remote students at roughly life size. Three rear-projected displays are tiled to form the 15 by 4 foot display wall. The wall is divided into a grid of seats where students can appear. The instructor can alter the seating arrangement by dragging the student's video to any empty seat using a wireless mouse.

The display wall audio system allows the instructor to easily locate the speaking student. Each of the three sections of the display wall has a loudspeaker and students displayed on the same section share the same loudspeaker. Since students displayed on different sections of the display wall use different loudspeakers, the instructor can locate the general direction of the speaking student from the location of the loudspeaker. Each student's instantaneous audio volume is displayed next to his name to

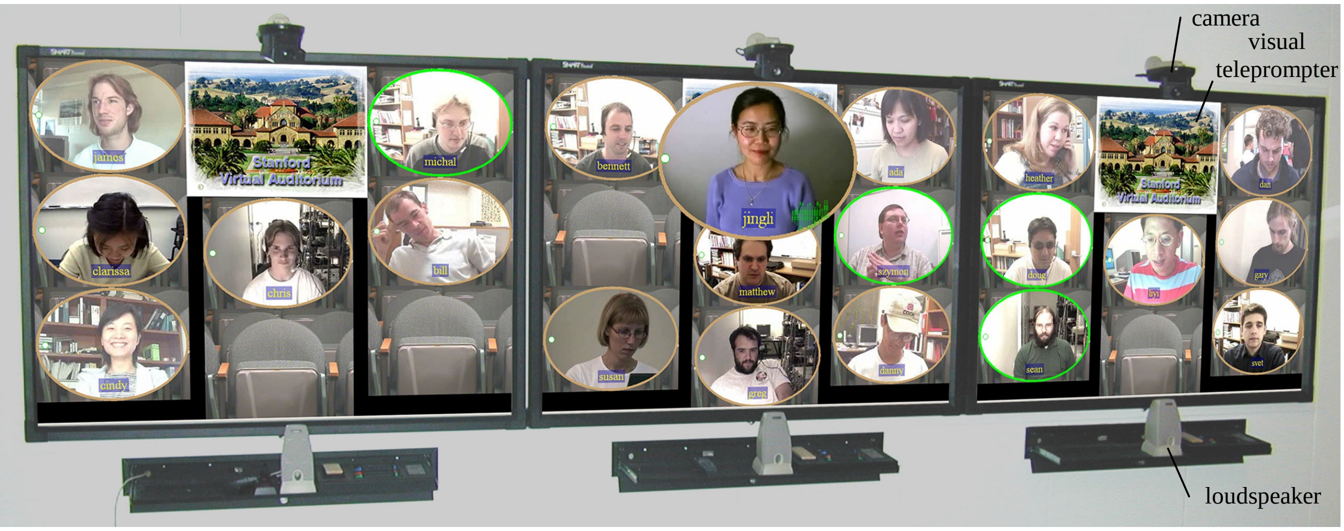

Figure 1. The Virtual Auditorium display wall. This wall can display 24 students and the instructor can move students to different seats. Videos are elliptically shaped to provide a common background for all students. A student's voice is rendered from the loudspeaker closest to his image and his instantaneous audio amplitude is displayed next to his name. The audio localization and amplitude display allow the instructor to easily identify the speaker. Directly below the cameras are regions called the visual teleprompters that show visual aids or students in directed gaze mode.

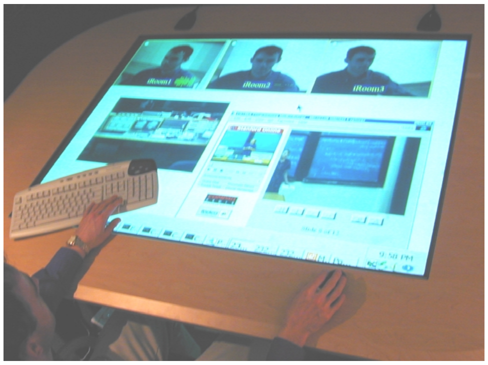

Figure 2. The Virtual Auditorium control panel. The control panel is used to display and manipulate visual aids. It can also show the different views of the instructor.

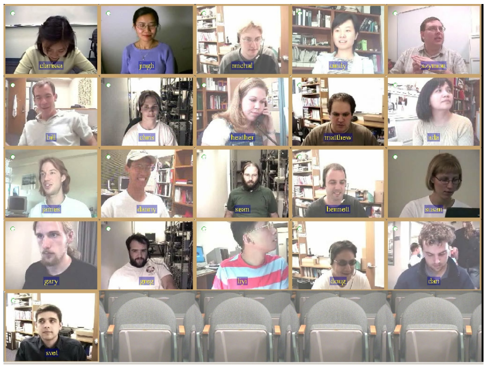

Figure 3. Screen shot of a student's monitor. The instructor is not framed differently from the students to encourage student discussions.

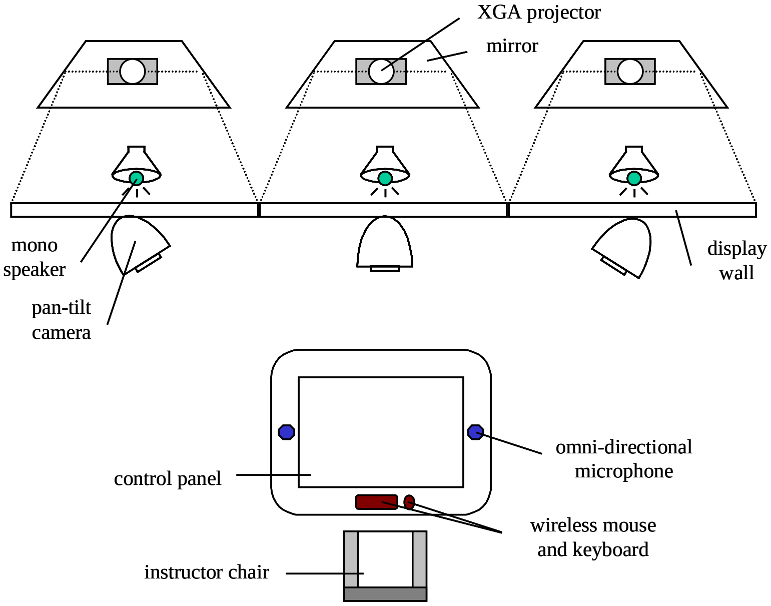

Figure 4. The Virtual Auditorium top view diagram. The display wall consists of three rear-projected displays spanning a total of 15 by 4 feet. The projectors point upward and mirrors are used to fold the 7-foot throw distance into the 4-foot deep space behind the display wall. The control panel is a 4 by 3 feet table illuminated from below. The instructor chair is 10 feet from the display wall. Mono loudspeakers are used to enhance sound localization. An echo cancellation mixer frees the instructor from wearing a headset during class. The auditorium walls are soundproofed and the ventilation system is tuned to decrease the ambient sound level.

enhance the visual signal of lip movement; thus, allowing the instructor to easily locate the speaking student from within each sections of the display wall.

3.2. Eye Contact with Directed Gaze

The instructor can establish eye contact with any one student, a group of students, or the entire class using a technique called directed gaze. Three cameras with pan, tilt, and zoom capability are mounted above the display wall. Figure 5 shows the three views of an instructor from these cameras. Below each camera is a region of the display wall called the visual teleprompter. The angle between a camera and its visual teleprompter is minimized such that the instructor can establish eye contact with the student displayed at the visual teleprompter.

When the instructor is lecturing, only visual aids are shown at the visual teleprompter. Each student sees the instructor from the camera closest to the location that he occupies on the display wall; therefore, when the instructor looks at the visual teleprompter, the instructor is making eye contact with all of the students rendered on that section of the display wall.

When a student is speaking, his video is enlarged and displayed at the closest visual teleprompter. At the same time, all the other students sharing that display begin viewing the instructor from one of the other two cameras; therefore, when the instructor looks at the student displayed at the visual teleprompter, the instructor is making eye contact with only this student. The instructor can also manually place a student at the visual teleprompter by double clicking on that student's video; thus, allowing him to establish eye contact with a currently silent student. Directed gaze can also be used to direct gestures to a target student.

A disadvantage of directed gaze is that the conceptual usage model is different from a face-to-face environment. In a face-to-face classroom, the instructor can establish eye contact with any student by looking at that student; however, in the Virtual Auditorium, the instructor must select a student first before eye contact can be established with a silent student. An advantage of directed gaze is that only two cameras are required to allow eye contact with any student independent of the class size.

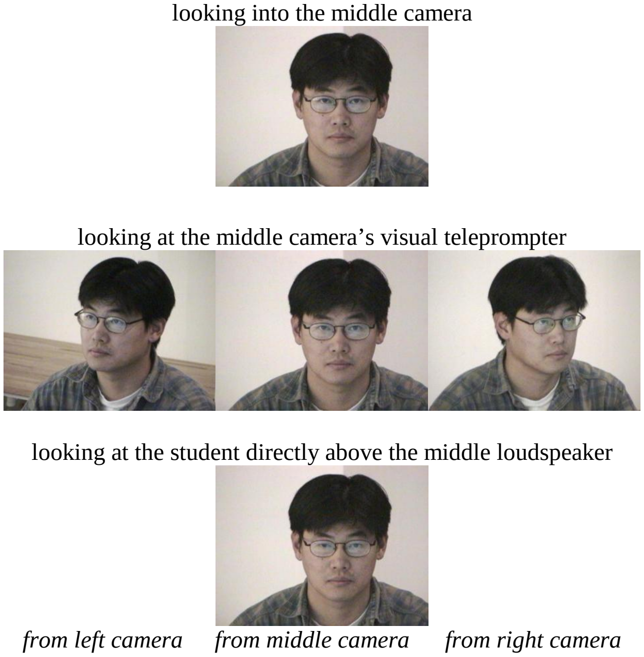
looking into the middle camera
looking at the middle camera's visual teleprompter
looking at the student directly above the middle loudspeaker
from left camera
from middle camera
from right camera

Figure 5. Views of the instructor from the display wall cameras. The pictures are laid out on a grid where the horizontal axis indicates the camera used to take the picture and the vertical axis indicates where the instructor was looking. Notice that from the middle camera, looking into the camera is indistinguishable from looking at the visual teleprompter. The figure also shows that students looking from the left and right cameras can see that the instructor is looking at someone else.

3.3. Floor Control

Students attend the class in front of their computers equipped with Universal Serial Bus or other inexpensive cameras. Students are also required to wear microphone headsets unless local echo cancellation devices are available. Figure 3 shows a screen shot of a student's monitor. Notice that the instructor is not framed differently from the students to encourage student discussions. If communication or computation bottlenecks exist, only the instructor and self views are rendered at full frame rate and the other students are updated every few seconds like Xerox's Porthole [8]. The instructor can also choose a lecture centric layout for the students' monitors, where only him and the visual aids are shown.

A student can request to speak by either raising his hand or pressing the space bar on the keyboard. Pressing the space bar causes all video windows showing him to have green borders. Our original design did not allow the students to request the floor by pressing the space bar key since we wanted to design an easy-to-learn system by minimizing the introduction of new conceptual usage models. Our user studies found that a raised hand among up to 36 video streams was easily detected; however, during live trials, we observed that while the instructor always saw a student's raised hand immediately, other students sometimes saw that same hand a few seconds later. When there are many students in the auditorium, the system will decrease the frame rate to reduce the communication and computation requirements, thus video frames are not always immediately delivered to other students. The uncertainty of when other students will also see a raised hand can cost confusion; thus, students can request to speak by pressing the space bar key.

From the control panel, Microsoft NetMeeting is used to share visual aids with the students. A down-sampled version of the visual aids also appears at the visual teleprompters. A mouse and keyboard middleware links the control panel and the three sections of the display wall into a single sheet, thus allowing a single mouse and keyboard to move seamlessly across the displays.

4. SOFTWARE IMPLEMENTATION

Videoconferencing with a large number of students is difficult due to the communication and computation requirements. Linking 20 students using a NetMeeting-grade compression scheme could require the network to sustain up to 200Mbps, a requirement that would challenge even Internet 2. One approach to lowering the bandwidth requirement is to use a more efficient codec. Section 4.1 describes one such system based on MPEG-4.

A single PC currently cannot decompress a large number of high quality video streams. One solution is to use multiple

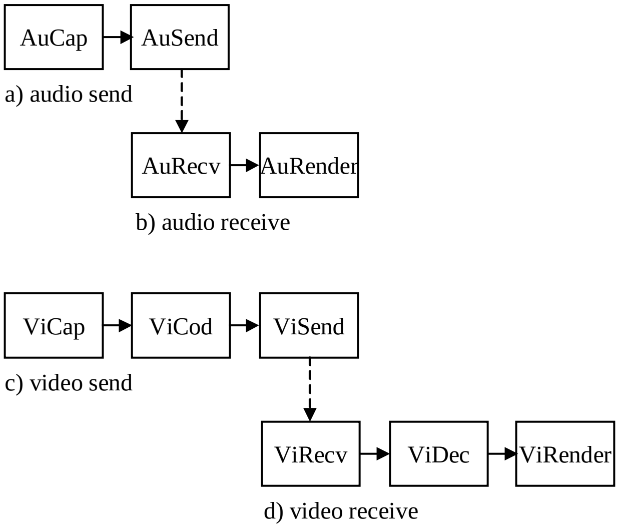

Figure 6. DirectShow AV filter graphs. A solid arrow indicates that the filter uses shared memory to pass data to the downstream filter. A dashed arrow indicates that the filter uses network packets to pass data to the downstream filter. Each AuSend and ViSend filter can broadcast data to many computers.

computers and piece together the computer outputs into a single display. Such a parallel-decoding system is easier to use if a seamless user interface can span all the computers driving the display. The interface should allow a single pointing device to move videos to anywhere on the display without regard to computer boundaries. Section 4.2 describes one such interface based on stream migration.

Significant effort is usually required to retrofit an existing conference system to use a newer codec or better transport mechanism. Section 4.1 describes a modular architecture based on Microsoft's DirectShow that allows streaming components to be upgraded with minimal programming effort. This architecture also allows for rapid prototyping.

Noticeable audio delay can make spontaneous and lively communication difficult; thus, the total system delay must be comparable to that of the telephone. Current commercial systems typically cannot stream television quality video and it is unclear what level of video quality is required for a remote classroom. Nevertheless, our implementation should allow streaming of television quality video to support user studies on video quality.

In order of importance, our design goals are:

- Telephone quality audio and television quality video
- Lower bandwidth requirement than the current commercial conferencing systems
- Seamless user interface that hides the machine boundaries of a multi-computer display wall
- Modular architecture for component upgrade and rapid prototyping

4.1. Modular AV Streaming

Our implementation uses Microsoft DirectShow. DirectShow specifies language-independent interfaces for multimedia software and hardware components, also known as filters. DirectShow also provides a framework for controlling filters, which form directed graphs. Data originates from source filters such as video capture, flows through transform filters such as compression codecs, and is consumed by sink filters such as video renderers. Filters negotiate with each other on a common media format and the DirectShow framework automatically inserts format converters if required. A disadvantage of DirectShow is that unlike previous systems with similar characteristics [21], it requires the Microsoft Windows operating system. An advantage of DirectShow is that numerous commodity filters are available.

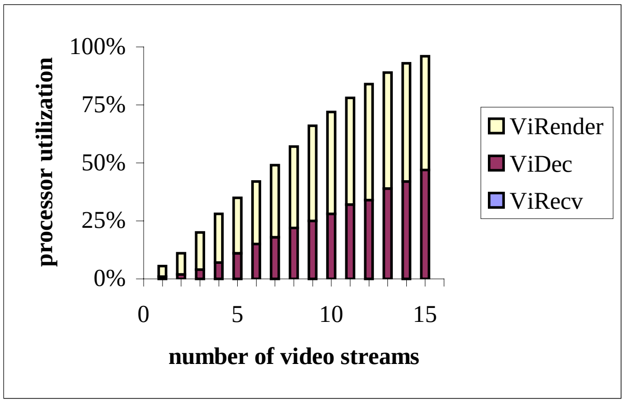

Figure 7. Measured processor utilization for the video rendering graph in Figure 6d on a dual 550 MHz Pentium III Xeon machine. Videos are 320 by 240 pixels, 15 fps, and compressed with Microsoft MPEG-4. Receiving from network takes less than 1% utilization and is not visible on the chart. The utilization for one stream is 5.5 percent.

Figure 6 shows the Virtual Auditorium filter graphs. We implemented all the filters shown except the video compression and decompression filters. Commodity filters were used for rapid prototyping; however, custom implementation was required due to latency and compatibility issues.

4.1.1. Audio Streaming

Figure 6a and 6b show the audio streaming modules. The AuCap filter uses the Microsoft DirectSoundCapture interface to capture audio from the microphone. The AuRender filter uses the Microsoft DirectSound interface to write audio to the loudspeaker. Almost all PC sound cards support DirectSound and DirectSoundCapture. AuRender also computes the instantaneous audio volume.

The AuCap filter retrieves data from the sound card in roughly 30 millisecond chunks. The AuSend filter sends each chunk of data using UDP unicast or multicast to multiple computers. Data is sent without per packet descriptive information. The AuRecv filter performs a blocking read on a UDP port or a multicast address and passes the received data immediately to the AuRender filter. The AuRender filter maintains a playback buffer that stores the received but not yet played audio to offset capture and network jitters. The DirectSoundCapture clock typically runs slightly faster than the DirectSound clock. This difference causes the playback buffer to accumulate, thus gradually increasing the overall latency. When the playback buffer has accumulated 200 milliseconds of audio, we clip the playback buffer to 60 milliseconds. These two parameters were empirically tested to yield good sound quality.

Packet replication is used to fill in for lost packets. In the case of severe packet loss, playback is stopped until new data is received.

The overall audio latency is comparable to that of the telephone. AuCap incurs 60 milliseconds of latency, a limitation of DirectSoundCapture. AuRender incurs another 60 milliseconds of latency in the playback buffer. Network delay is typically 10 to 20 milliseconds. Audio and video are not synchronized during playback to minimize audio latency as recommended by [14]. The processor utilization for audio processing is negligible on a modern PC.

4.1.2. Video Streaming

Figure 6c and 6d show the video streaming modules. The ViCap filter can capture video from any video capture card or camera that supports the Video-For-Windows or the Windows-Driver-Model interface. The ViRender filter can use Microsoft's GDI to render video to an arbitrarily shaped window. It also exposes an interface for annotating video with the student name and audio volume. A disadvantage of GDI is that it is less efficient than DirectDraw. An advantage of GDI is that it supports transparency and text output.

The ViSend filter can use UDP unicast or multicast to stream raw video data to nodes with different bandwidth requirements. Compressed video frames larger than the maximum UDP packet size are divided into multiple packets. A packet descriptor is attached to the end of each network packet. Attaching the descriptor to the tail, rather than the head, of each packet allows the buffer allocated for compression to be used to construct the network packet, thus saving a memory copy. The descriptor contains the media sample time, sequence number, and DirectShow specific sample information. To produce a lower bandwidth stream, ViSend can extract intra-frames, compressed frames that do not require other frames to decompress.

The filter graph in Figure 6c and 6d can use a different compression and decompression scheme by changing the ViCod and ViDec filters. This is the only change necessary since the ViSend and ViRecv filters can accept any compression format and the ViRender filter accepts uncompressed video frames. We have evaluated Microsoft MPEG-4, Intel H263, Intel wavelet, PICVideo Motion-JPEG, and PICVideo Lossless-JPEG. At approximately the same visual quality, Microsoft MPEG-4 has the lowest data rate at roughly 100Kbps for a 320x240x15fps video; this is roughly half the bandwidth requirement of Microsoft NetMeeting.

The processor utilization for the video capture graph, Figure 6c, is 9 percent on a dual 550 MHz Pentium III Xeon using a Hauppauge WinTV-GO PCI video capture card. The actual video capture takes less than 1 percent processor utilization using PCI capture cards but about 20 percent for USB capture solutions. Since the network send takes negligible processing, one computer can provide video to a large number of students. Figure 7 shows the processor utilization for the filter graph in Figure 6d. Notice that a modern PC can process a dozen video streams before reaching maximum utilization. Television quality video, 640 by 480 pixel at 30 fps, can also be processed on a high-end PC.

4.1.3. Conference Session Startup

The Virtual Auditorium software consists of four applications: AudioServer, VideoServer, CameraServer, and AVC_Client. The AudioServer creates the filter graph in Figure 6a. It listens for network requests to stream out audio, and if the requesting address matches an approved address in a database, it adds the requesting address into an array of destination addresses in the AuSend filter. The VideoServer parallels the function of the AudioServer, and builds the filter graph in Figure 6c. The CameraServer is launched if the camera has a control interface for the camera's pan, tilt, or zoom. The AVC_Client creates the filter graphs in Figure 6b and 6d and establishes TCP connections with the servers to request streaming.

The Virtual Auditorium software can be started from Internet Explorer. We created an ActiveX web page that contains a text box and connect button. After entering the name of the computer to be connected in the text box and pressing the connect button, ActiveX downloads the required DirectShow filters and applications, registers the filters with the operating system, and launches the client and server applications. This process can be repeated to connect additional people, or alternatively, a class name can be entered to connect to a group of people. When the last Virtual Auditorium application is closed, all traces of the Virtual Auditorium are removed from the user's system. An advantage of this startup procedure is that an explicit software install is not required to use the Virtual Auditorium.

4.2. Hiding Machine Boundaries

Figure 7 shows that it is not possible to decode a large number of video streams using a single computer, thus parallel decoding is necessary to show a large number of videos. Such a parallel-decoding system is more usable if a seamless user interface can span all the computers driving the display; specifically, the user should be allowed to use a single pointing device to drag video windows across computer boundaries.

To allow a single pointing device, a mouse in our case, to control the multiple computers driving a display wall, all computers run a mouse server. The mouse is physically connected to another computer that intercepts all mouse events, maps the mouse coordinates to the corresponding point on the display wall, and passes the events to the computer driving that section of the display wall. The mouse servers listen for mouse events and insert the received events into the Windows Message Queue.

To allow dragging of AVC_Client video windows between computers, all computers driving the display wall run a remote execute server [20]. When more than half of an AVC_Client window crosses a screen boundary, it calls the remote execution server to launch an AVC_Client on the next screen and closes itself. The new AVC_Client uses the arguments of the parent AVC_Client to reestablish connections with the audio and video servers; thus, the instructor will see and hear the moved student from the new screen. The final step in the migration process is to have the moved student view the instructor from the camera associated with the new screen. This is accomplished by creating a temporary connection to the student's AVC_Client and requesting this AVC_Client to connect to the VideoServer associated with the new camera. The physical relationship between the computers driving the different sections of the display wall is stored in a database.

The migration process takes about one second to complete. The AVC_Client that started the migration process waits until the new AVC_Client is running before exiting; this delay prevents the moved student from disappearing from the display wall during the migration process.

5. USER STUDY

Video size and fidelity influence people's psychological response to the content of the video [27]. When the signal to noise ratio of videoconference is low, the user must concentrate to parse out valuable information from the noise. Cinematographers have long applied techniques to exaggerate or suppress visual signals to better communicate with the viewer. People have often assumed that a life-size rendering of a remote participant is ideal for videoconferencing [5][26]; it is unclear if that still applies when communicating with a group. Section 5.1 presents the findings of a user study on the optimal display size for communicating with a group.

The directed gaze technique assumes that a remote viewer cannot distinguish between a person looking into a camera or at a display if the angle between the camera and the display is small. Although the perception of gaze has been extensively studied [3][7][10][32], it is still unclear how small this camera-to-display angle must be to achieve eye contact in an auditorium environment. Section 5.2 presents the findings of a user study on the required camera-to-display angle to achieve eye contact.

5.1. Optimal Display Wall Size

This study focused on the effect of video size on the ease of detecting facial expressions. Facial signals can be classified into five message categories: emotions, emblems, manipulators, illustrators, and regulators [24]. Emblems are culture specific symbolic communicators such as a wink. Manipulators are self-manipulative associated movements such as lip biting. Illustrators are actions accompanying and highlighting speech such as a raised eyebrow. Regulators are nonverbal conversational mediators such as nods or smiles.

We limited this study to detecting smiles. Smiles are well defined and easy for people to act out, thus allowing us to generate a large, unambiguous dataset. Smiles can also be subtle, thus a system that allows smiles to be detected with ease may also be acceptable for detecting facial expressions in general. We conducted a series of four experiments on the ideal and minimum size to converse with a single person, the recognition time for smiles as a function of video size, preference of display size for viewing a group of people, and the recognition time for spotting a smile from a group of people.

Twelve subjects, chosen from graduate students, university staff, and working professionals, participated in this study. Eight subjects had used videoconferencing less than two times and the other four subjects had extensive experience with videoconferencing. No subject had visual impairments that prevented him from seeing the displays clearly.

5.1.1. Ideal and Minimum Display Size for Dyads

This experiment is on the ideal and minimum video size for dyadic videoconferencing. We recorded a student during a videoconference at 320 by 240 pixels and 15 fps for one minute. We played back this video on a 5 by 4 foot display, the middle screen of the Virtual Auditorium display wall, with the subject seated 10 feet away. The subject could adjust the video from covering the entire screen to a single pixel by using the arrow keys on the keyboard. The subject was asked to select the ideal and minimum size to comfortably videoconference with the recorded person.

Figure 8 shows the result of this experiment. The average minimum visual angle is 6 degrees. The ideal size is 14 degrees, slightly larger than the life size of 12 degrees. When asked why they chose that particular size as ideal, subjects mentioned the tradeoff between the ease of seeing facial expressions and invasion of personal space. For example, most people found the full screen video the easiest for judging facial expressions, but they also found that the full screen face seemed to invade their personal space.

5.1.2. Recognition Time of Smiles for Dyads

The second experiment measured the subjects' recognition time for smiles as the video size was changed. We recorded five 30-second videos of a student at 320 by 240 pixels and 15fps. We asked the student to smile five times for each video. We viewed the five videos frame by frame and recorded the time when the smile began and ended. A smile beginning was defined as the first video frame when viewed as a static image that the person can be judged as unambiguously smiling. We played these recordings at 1024x768 (full screen), 640x480, 320x240, 160x120, and 80x60 pixels on the same display as in the previous experiment. These sizes corresponded to a visual angle of 27, 17, 8, 4, and 2 degrees when viewed from the subject's seat at 10 feet away from the screen. The subjects were instructed to press the space bar key on the keyboard whenever they saw the recorded person smile.

Figure 9 shows the result of this experiment. The *p* values of one-tailed T-test between the 17, 8, 4, and 2-degree data to the 27-degree data were 3.7%, 5.4%, 0.2% and 0.2%. This suggested that the knee of the recognition time curve was between 4 and 8 degrees, roughly the same as the minimum comfortable viewing angle from the previous experiment. Subjects' spontaneous comments indicated that the 2-degree viewing case was worse than the recognition time curve indicated. Many subjects said that the task was impossible or that it took all of their concentration. Also, notice that when the display was larger than what people found to be ideal, their recognition times did not noticeably improve.

5.1.3. Preference of Display Size for Groups

The third experiment asked subjects whether they preferred a large or immersive display to see a class of 9 and 36 students. The large display spanned a 27-degree field of view, a rough approximation to the Society of Motion Picture and Television Engineers 30-degree recommendation for big screen home theater. The immersive display, based on immersive movie theaters, spanned a 64-degree field of view. The key difference between the two was that the immersive display required head movements to see different parts of the screen. These two viewing options corresponded to the middle screen and to all three screens of the Virtual Auditorium display wall respectively. Each student in the 9-person class spanned 9 degrees on the large display and 14 degrees on the immersive display. Each student in the 36-person class spanned 4 degrees on the large display and 7 degrees on the immersive display. Each of the 45 people displayed was prerecorded for one minute watching a SITN course. The subjects were seated 10 feet from the screen and could press a button to toggle between the viewing conditions before stating their preference.

Figure 10 shows the result of this experiment. On both the large and immersive displays, each video in the 9-student class

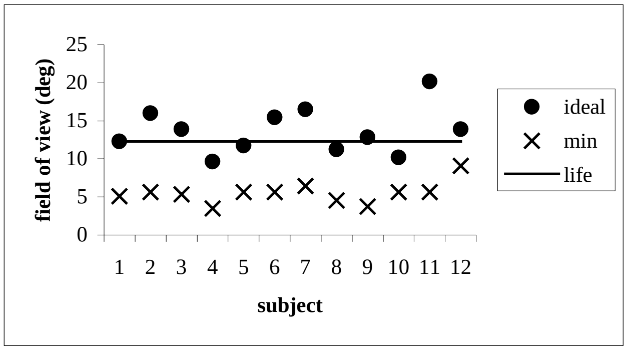

Figure 8. Ideal and minimum field of view for comfortably videoconferencing with a single person. The solid line indicates life size at 12 degrees. The average ideal and minimum are 14 and 6 degrees respectively.

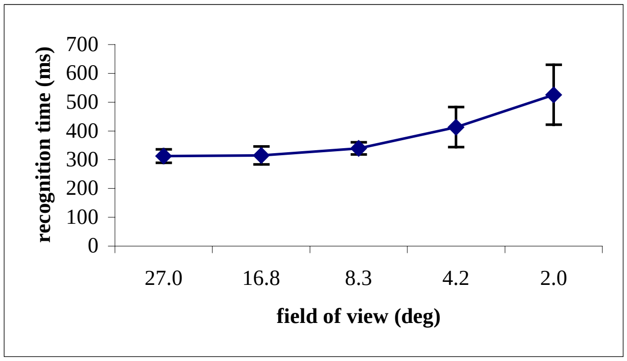

Figure 9. Average and standard deviation of smile recognition time in a video as a function of the video display size.

was larger than the minimum video size found in the first experiment, thus most subjects preferred the smaller display since it allowed them to see everyone without head movements. Showing 36 people on the large display forced each video to be below the minimum size for comfortable viewing, thus most people preferred the immersive display even though that required head movements to see everyone.

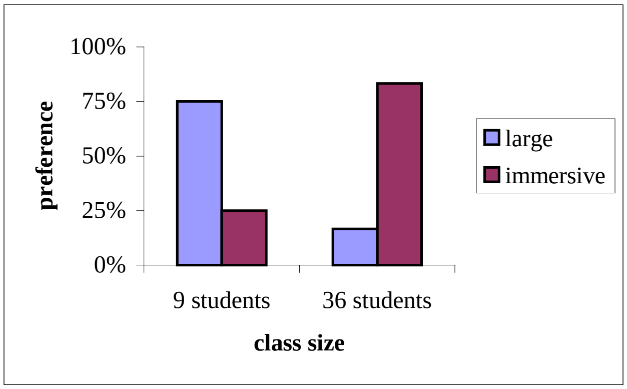

Figure 10. Preference for seeing 9 and 36 students on a large or immersive display. 75% of the 12 subjects preferred seeing 9 students on the large display while only 16% of the subjects preferred seeing 36 students on that same display.

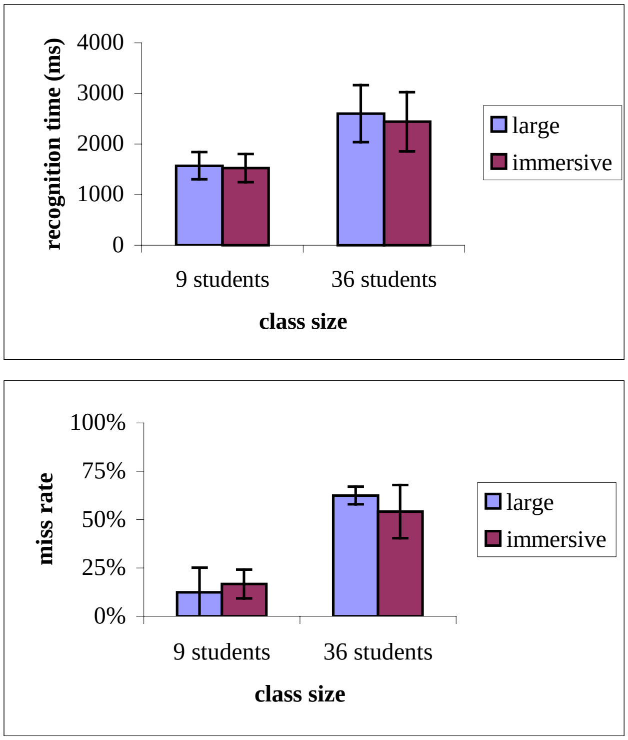

Figure 11. Average and standard deviation of smile recognition time and miss rate from 9 and 36 students on the large and immersive displays.

5.1.4. *Recognition Time of Smiles for Groups*

In a classroom, there are two modes of sensing the students' facial signals. In one case, the instructor pauses and actively scans the class for feedback, such as after asking a question or telling a joke. In the other case, while the instructor is lecturing, he passively senses the mood of the class. The difference between the two cases is that monitoring facial signals is the primary task in the first case and a background task in the second case. The fourth experiment is designed to test the ability of people to actively scan for facial signals.

The fourth experiment measured the subjects' recognition times for smiles for the two class sizes using the same two viewing options as in the third experiment. We recorded 45 people watching SITN courses for one minute. Some of the people were asked to smile at a particular time. The smile times were pre-calculated to give 12 smiles randomly distributed over the one-minute interval for each class. These videos were viewed frame by frame to determine the smile timings as in the second experiment. The subjects were instructed to press the space bar key on the keyboard if they saw any one of the recorded students smile.

Figure 11 shows the result of the fourth experiment. The smiles in our test data lasted an average of three seconds. When the subject could not process all of the videos within that time, some smiles were missed. This is not unlike a real classroom where sometimes the instructor may not notice a raised hand for a long time. This trend was observed when comparing the results for the 9 people class to the 36 people class. While only 12 percent of the smiles were missed when there were 9 students, 62 percent of the smiles were missed when the class size was increased to 36 students. This finding parallels what most teachers already know: large class size is detrimental to teacher-student interaction.

Given the strong preference in experiment 3, we were surprised to find very little difference in how the subjects actually performed when the display size was changed for each class size. This was probably due to the subjects' ability to adapt to the viewing condition. Since we were measuring the performance of active scanning, the subjects could devote more cognitive effort to balance the difference in display. Many subjects commented that this task was extremely tiring.

5.2. Maximum Angle for Eye Contact

This experiment measured the maximum allowable angle between the display and the camera to give the appearance of eye contact for the remote viewer. The Virtual Auditorium was used to connect the experimenter and the subject. The experimenter saw the test subject on the display wall from 10 feet away and the test subject saw the experimenter on a computer monitor through the middle camera above the display wall. The experimenter moved the video window showing the subject to various locations on the display wall. While looking at the test subject the experimenter asked if the subject thought the experimenter was looking at him. 320x240x15 fps video streams were used.

Eight subjects participated in this study and Figure 12 shows our findings. When the video window was directly below the camera, corresponding to a 3-degree deviation from the camera, all subjects said that the experimenter was looking directly at them. The angle for the most sensitive subject was +/- 2.7 degrees horizontally and 9 degrees vertically; however, there was a large variation in subjects' sensitivity. Our result is similar to the Mod II PicturePhone finding where their maximum allowable camera-to-display angle for eye contact was 4.5 degrees horizontally and 5.5 degrees vertically [32]. Unfortunately, we were unable to find the description of their experiment for further analysis.

All subjects reported that this task was very difficult to perform. This task was challenging because our experiment assumed that a single number could describe the required angle to achieve eye contact, in fact, this angle could be better characterized as a range of angles. Within a range of angles, the ring of confusion, each subject could not tell from the video if the experimenter was looking directly at him and his expectation determined his perception of eye contact. The classic gaze experiments used a few points as the target of gaze [7][10], while our experiment allowed an essentially continuous range of gaze targets; this may be the reason why the ring of confusion gaze phenomenon has not been previously reported.

6. DISCUSSION AND CONCLUSION

The Virtual Auditorium is routinely used to link small groups of people. Both the audio and video are of sufficient quality for people to talk freely. People also report a strong sense of awareness for the remote participants.

Traditionally it takes many people to build a videoconferencing system. With the exceptions noted in this paper, the Virtual Auditorium was built by the author over an 18-month period. We were able to accelerate our effort by using a modular design that leveraged commodity videoconferencing building blocks. The AV equipment was acquired and installed in six months, during which over 1500 feet of AV cables were laid. The software was written in twelve months. While less than one month was required to demonstrate streaming video, significant effort was required to craft the user interface and to make the software robust. The user study was conducted in three months and the SITN classroom observation was performed in six months.

The Virtual Auditorium is designed to test the hypothesis that dialog-based distance teaching is possible if we allow the instructor to see the remote students and the remote students to see each other. We have constructed an auditorium, written a videoconferencing software, and measured design parameters. For future work, we would like to formally test our hypothesis. Trials using the Virtual Auditorium to link Stanford to UC Berkeley, UC Santa Barbara, Sweden, Germany, and Japan are underway.

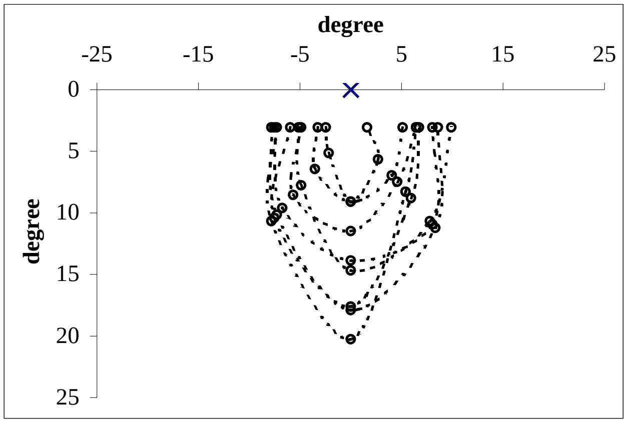

Figure 12. Maximum allowable angle between the camera and the display to maintain eye contact. The "x" at 0,0 degree indicates the location of the camera. Each curve plots the expected response of a subject and each data point shows the measured angle that eye contact is no longer perceived. The most sensitive subject required the display to be less than 2.7 degrees horizontally and 9 degrees vertically from the camera.

7. ACKNOWLEDGMENTS

This project is sponsored by the Stanford Interactive Workspace Project, the Stanford Immersive Television Project, the Stanford Learning Lab, and a Department of Defense Graduate Fellowship. We wish to thank Francois Guimbretiere and Maureen Stone for designing and building the display wall projection system, Brad Johanson for designing and implementing the keyboard, mouse, and remote execution servers, Dan Nelson for tuning the audio equipment, Susan Shepard for installing the soundproofing, Jingli Wang for participating in the pilot user studies, and Matthew Eldridge, Brad Johanson, Yung-Hsiang Lu, James Davis, Diane Tang, and Pat Hanrahan for critiquing an early draft.

8. REFERENCES

[1] G. Abowd, C. Atkeson, A. Feinstein, C. Hmelo, R. Kooper, S. Long, N. Sawhney, and M. Tani. Teaching and Learning as Multimedia Authoring: The Classroom 2000 Project. *Proceedings of ACM Multimedia*, pages 187-198, 1996.

[2] AccessGrid. http://www-fp.mcs.anl.gov/fl/accessgrid

[3] M. Argyle and M. Cook. Gaze and Mutual Gaze. Cambridge University Press, 1976.

[4] J. Bransford, A. Brown, and R. Cocking. How People Learn: Brain, Mind, Experience and School. National Academy Press, 2000.

[5] W. Buxton. Living in Augmented Reality: Ubiquitous Media and Reactive Environments. Video-Mediated Communication (edited by K. Finn, A. Sellen, and S. Wilbur), Lawrence Erlbaum Associates, pages 363-384, 1997.

[6] W. Buxton, A. Sellen, M. Sheasby. Interfaces for Multiparty Videoconferences. Video-Mediated Communication (edited by K. Finn, A. Sellen, and S. Wilbur), Lawrence Erlbaum Associates, pages 385-400, 1997.

[7] M. Cline. The Perception of Where a Person is Looking. *American Journal of Psychology*, pages 41-50, 1967.

[8] P. Dourish and S. Bly. Portholes: Supporting Awareness in a Distributed Work Group. *Proceedings of CHI*, pages 541-547, 1992.

[9] J. Gibbons, W. Kincheloe, and K. Down. Tutored Videotape Instruction: a New Use of Electronics Media in Education. *Science*, pages 1139-1146, 1977.

[10] J. Gibson and A. Pick. Perception of Another Person's Looking Behavior. *American Journal of Psychology*, pages 386-394, 1963.

[11] L. He, E. Sanocki, A. Gupta, and J. Grudin. Auto-Summarization of Audio-Video Presentations. P*roceedings of ACM Multimedia*, pages 489-498, 1999.

[12] L. He, E. Sanocki, A. Gupta, and J. Grudin. Comparing Presentation Summaries: Slides vs. Reading vs. Listening. *Proceedings of CHI*, pages 177-184, 2000.

[13] E. Isaacs, T. Morris, T. Rodriguez, and J. Tang. A Comparison of Face-to-face and Distributed Presentations. *Proceedings of CHI*, pages 354-361, 1995.

[14] E. Isaacs and J. Tang. Studying Video-Based Collaboration in Context: from Small Workgroups to Large Organizations. Video-Mediated Communication, Lawrence Erlbaum Associates, pages 173-197, 1997.

[15] H. Ishii and M. Kobayashi. ClearBoard: a Seamless Medium for Shared Drawing and Conversation with Eye Contact. *Proceedings of CHI*, pages 525-532, 1992.

[16] P. Jackson. The Teacher and The Machine. Horace Mann Lecture, 1967.

[17] G. Jancke, J. Grudin, and A. Gupta. Presenting to Local and Remote Audiences: Design and Use of the TELEP System. *Proceedings of CHI*, pages 384-391, 2000.

[18] R. Kraut and R. Fish. Prospects for Videotelephony. Video-Mediated Communication (edited by K. Finn, A. Sellen, and S. Wilbur), Lawrence Erlbaum Associates, pages 541-561, 1997.

[19] Lawrence Berkeley National Laboratory Mbone tools. http://www-nrg.ee.lbl.gov/nrg.html

[20] B. Johanson, S. Ponnekanti, C. Sengupta, and A. Fox. Multibrowsing: Moving Web Content across Multiple Displays. *Proceedings of Ubiquitous Computing Conference*, 2001.

[21] S. McCanne, E. Brewer, R. Katz, L. Rowe, E. Amir, Y. Chawathe, A. Coopersmith, K. Patel, S. Raman, A. Schuett, D. Simpson, A. Swan, T. Tung, D. Wu, and B. Smith. Toward a Common Infrastructure for Multimedia-Networking Middleware. *Proceedings of International Workshop on Network and Operating System Support for Digital Audio and Video*, 1997.

[22] S. Mukhopadhyay and B. Smith. Passive Capture and Structuring of Lectures. *Proceedings of ACM Multimedia*, pages 477-487, 1999.

[23] A. Noll. Anatomy of a Failure: PicturePhone Revisited. *Telecommunications Policy*, pages 307-316, 1992.

[24] NSF Workshop on Facial Expression Understanding, 1992. http://mambo.ucsc.edu/psl/nsf.txt

[25] R. Ochsman and A. Chapanis. The Effects of 10 Communication Modes on the Behavior of Teams During Co-operative Problem-Solving. *International Journal of Man-Machine Studies*, pages 579-619, 1974.

[26] K. Okada, F. Maeda, Y. Ichikawaa, and Y. Matsushita. Multiparty Videoconferencing at Virtual Social Distance: MAJIC Design. *Proceedings of CSCW*, pages 385-393, 1994.

[27] B. Reeves and C. Nass. The Media Equation. Cambridge University Press, 1996.

[28] R. Riez and E. Klemmer. Subjective Evaluation of Delay and Echo Suppressers in Telephone Communication. *Bell System Technical Journal*, pages 2919-2942, 1963.

[29] L. Rowe. ACM Multimedia Tutorial on Distance Learning, 2000.

[30] A. Sellen. Remote Conversations: The Effects of Mediating Talk with Technology. *Human-Computer Interaction*, pages 401-444, 1995.

[31] Stanford Instructional Television Network. http://www-sitn.stanford.edu

[32] R. Stokes. Human Factors and Appearance Design Considerations of the Mod II PicturePhone Station Set. *IEEE Transactions on Communication Technology*, pages 318-323, 1969

[33] University College London Mbone tools. http://www-mice.cs.ucl.ac.uk/multimedia/software

[34] S. White, A. Gupta, J. Grudin, H. Chesley, G. Kimberly, and E. Sanocki. Evolving Use of A System for Education at a Distance. *Proceedings of Hawaii International Conference on System Sciences*, 2000.