Automating Camera Management for Lecture Room Environments

Qiong Liu, Yong Rui, Anoop Gupta and JJ Cadiz

September 21, 2000

Technical Report
MSR-TR-2000-90

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

Automating Camera Management for Lecture Room Environments

Qiong Liu, Yong Rui, Anoop Gupta and JJ Cadiz
Collaboration and Multimedia Systems Group, Microsoft Research
One Microsoft Way
Redmond, WA 98052-6399
q-liu2@ifp.uiuc.edu, {yongrui, anoop, jjcadiz}@microsoft.com

ABSTRACT
Given rapid improvements in network infrastructure and streaming-media technologies, a large number of corporations and universities are recording lectures and making them available online for anytime, anywhere access. However, producing high-quality lecture videos is still labor intensive and expensive. Fortunately, recent technology advances are making it feasible to build automated camera management systems to capture lectures. In this paper we report on our design, implementation and study of such a system. Compared to previous work—which has tended to be technology centric—we started with interviews with professional video producers and used their knowledge and expertise to create video production rules. We then targeted technology components that allowed us to implement a substantial portion of these rules, including the design of a virtual video director. The system's performance was compared to that of a human operator via a user study. Results suggest that our system's quality is close to that of a human-controlled system. In fact, most remote audience members could not tell if the video was produced by a computer or a person.

Keywords
Automated camera management, Video production rules, Virtual video director, Speaker tracking, Sound source localization.

1 INTRODUCTION

Given the rapid pace of technological change and accompanying emphasis on life-long learning, both universities and corporations are offering more lectures, seminars, and classes to teach and train students and employees. To accommodate audiences' time and/or space conflicts, many of these lectures are made available online, allowing people to attend remotely, either live or on-demand. For instance, at Stanford University, lectures from over 50 courses are made available online every quarter [20]. The Microsoft Technical Education Group (MSTE) has supported 367 on-line training lectures with more than 9000 viewers from 1998 to 1999 [12]. In fact, online audiences for many of the talks are now starting to exceed the number of people who attend the talks in person [12].

While online publishing of lectures is gaining momentum, clearly majority of lectures and talks that occur are not recorded and made available online. A key barrier is the cost of equipping lecture rooms with cameras and the cost of people recording and putting the talks online. The former is a one-time cost and is becoming lower every day, but the latter is a recurring cost and it dominates. In our own organization, $500+ are spent on each talk that is made available online. This is primarily people cost, as the disk storage used for a one hour talk streamed at 256Kbps is ~120Mbytes (~$1 cost).

Fortunately, recent progress in computer vision and signal processing technologies is making it feasible to start automating the camera-management task for capturing lectures. While there have been previous attempts at this task, as we will discuss in the related work section, we believe they have been more technology centric rather than people/audience centric. Although technology is an indispensable part of the system, people are the final consumer of the product. Therefore, we started with an people-centric approach to address this problem. Specifically, we explore answers to the following questions in this paper:

1. How does a human camera crew record lectures, i.e., what are the camera-management rules important for capturing lectures?
2. If we can gather and summarize the rules used by the camera crew, how can we design and implement a fully automatic camera management system to realize those rules?
3. What is the overall quality of the automatic camera management system compared with that of a reasonable human operator? Can the system pass the Turing test? Furthermore, what are remote audiences' reactions to the various rules implemented in the system?

To address the first question, we scheduled discussion sessions with five professional video producers from our corporate studios and two from our Research Lab's lecture production team. We collected the rules they used in their everyday production practice, ranging from camera setup to video editing.

To address the second question, based on the state-of-the-art techniques in computer vision and signal processing, we evaluate which rules are achievable today at reasonable cost. For example, one rule is "*give lead room of gaze direction or head orientation for the speaker*." However, this rule is quite challenging to implement, because real-time and robust gaze detection and head orientation estimation are still open research problems in the computer vision research community [21,23].

To address the third question, we incorporated the feasible rules into a fully functional, multi-camera, automated system for capturing lectures. We present results from user studies indicating that automated camera management systems are quickly approaching the performance of human operators.

The rest of the paper is organized as follows. In Section 2, we provide a brief review of related research on lecture room automation. In Section 3, we present various video production rules collected from professionals. In section 4, we present detailed descriptions of how we design our system and how we implement the video production rules. We present experimental method and results from user studies in Sections 5 and 6. We present concluding remarks in Section 7.

2 RELATED WORK

In this section, we provide a brief review of related work from two aspects: individual tracking techniques and existing automated lecture capture systems.

2.1 Tracking techniques

Tracking technology is required both to keep the camera focused on the speaker and to display audience members when they talk. There are three general classes of tracking technology: sensor-based, motion-based, and microphone array-based. While all the three methods can be used for the speaker, only the last one is normally use for audience.

For sensor-based approaches, the speaker wears an IR or magnetic devices that emits electric or magnetic signals. A receiver unit uses the signal to locate the speaker. This technique has been used in both commercial products [17] and research prototypes [15]. Even though tracking is quite reliable using this technique, we consider wearing an extra device around the speaker's neck to be inconvenient and obtrusive.

Much literature also exists for tracking human object using vision-based techniques. Typical ones include skin-clor based tracking [21], motion-based tracking [10], and shape-based tracking [3]. Compare with sensor-based techniques, vision techniques are less obtrusive but are normally less accurate.

Microphone arrays are the best technology used to locate audience members who are talking. In general, these approaches can be separated into two categories: the general cross-correlation (GCC) approaches [6] and the blind de-convolution approaches [4]. There are also commercial products available for microphone array-based sound localization (e.g., PictureTel [18] and PolyCom [19]).

To summarize, different techniques exist for tracking objects. Sensor-based solutions are more reliable but less convenient. Vision and microphone array based techniques are unobtrusive and their quality is quickly approaching that of the sensor-based techniques.

2.2 Related systems

Several projects exist for lecture room automation [5,15,22]. In [22], Wang and Brandstein report a real-time talking head tracker that targets automated video conferencing. Their algorithms are computationally efficient to locate and track a talking head. However, their focus is a single-camera system, which is different from our goal.

In [15], Mukhopadhyay and Smith present an interesting system that captures audio/video information in a lecture room environment. They use a moving camera to track the lecturer and a static camera to capture the entire lecture dais. Though there are overlaps between this system and ours, the focus differs significantly. For example, "slides and video synchronization" is one of their major focuses. Our system further differs in the following important aspects. First, their system is designed for off-line, on-demand lecture watching. Ours, on the other hand, simultaneously edits a lecture while it is being recorded, which is suitable for both live broadcasting and on-demand viewing. Second, they track a speaker using a sensor-based technique, while in our system no extra equipment is needed for the speaker. In fact, the speakers are almost never aware that a camera is tracking them. Third, their system's editing rules are based almost entirely on the timing of slide transitions, while we take a systematic approach to collect and implement the video production rules used by human professionals.

Bellcore's AutoAuditorium [5,9] is one of the pioneers in lecture room automation. Among existing systems, it is closest to ours and has influenced our system design. The AutoAuditorium system uses four cameras to capture a lecture. Three of the cameras are fixed: one looks at the stage, one looks at the screen, and one looks at the lectern from the side. The fourth camera is a pan/tilt/zoom camera that tracks the speaker automatically by using computer vision techniques. The four video streams are then connected to a video mixer. An AutoAuditorium director (a software module) selects which video to show based on heuristics. For example, if the screen is projected by slides, the AutoAuditorium director will construct a "combination shot" where the speaker is placed in a picture-in-picture box in the lower corner of the screen camera image [5].

Our system differs from AutoAuditorium in several important aspects. First, no audience camera is used in the AutoAuditorium system. However, professional video producers suggest that an audience camera is important for lecture capture. It can focus on an audience member who asks questions and can provide random audience shots to make the lecture more enjoyable to watch. Second, picture-in-picture causes video resolution loss, which should be avoided in lecture videos. Third, there is no user study reported to compare the quality of their system against that of a human operator. We will report two user studies on our system at a later section in this paper.

Additional research projects exist for exploring other aspects of lecture automation, such as Classroom2000's effort on notes-capturing [7] and STREAM's effort on cross-media indexing [9]. Furthermore, several researchers have examined video mediated communication (e.g. Hydra, LiveWire, Montage, Poletholes, and Brandy Bunch) in the field of teleconferencing [8]. However, given its loose relation to this work, we do not elaborate on it here.

To summarize, significant progress has been made in tracking techniques and system architecture during the past few years. This paper contributes to the field by explicitly summarizing

video production rules, presenting a system realizing those rules, and giving detailed user study results.

3 VIDEO PRODUCTION RULES

As reviewed in the previous section, a common drawback in the existing systems is the lack of systematic study on professional video production rules. To ensure a successful system, we consider it imperative to collect, understand and implement those rules. We scheduled two formal discussions sessions and several informal sessions with seven professional video producers. We summarize the rules by category as follows.

3.1 How to set up the cameras

In video production, especially in filmmaking, there is a "line of interest" [1,11]. This line can be the line linking two people, the line a person is moving along, or the line a person is facing. An important rule is "*don't cross the line*." For example, if an initial shot is taken from the left side of the line, subsequent shots should all be taken from that side. This rule will ensure a moving person maintains the direction of apparent motion [11]. This rule can only be violated when a neutral shot is used to make the transition from one side of the line to the other.

To ensure the above rule, the cameras need to be set up properly. Figure 1 shows a top view of one of our organization's lecture rooms, where our system is installed. The lecturer normally moves behind the podium and in front of the screen. The audience area is in the right-hand side in the figure and includes 60 seats. There are four cameras in the room: a speaker-tracking camera, an audience-tracking camera, a static overview camera that gives an overview shot of the dais area, and a scan-converter camera that captures whatever is being displayed on the screen from the projector (typically PowerPoint slides).

In this lecture room environment, when the object of interest is the speaker, the "line of interest" is the line that the speaker is moving along: a line behind the podium and in front of the screen. It is easy to verify that our camera setup satisfies the rule of not crossing this line. When the object of interest is the audience, the line of interest is the line linking the speaker and the audience. Our camera setup satisfies the rule in this case as well.

3.2 How to frame the speaker

The speaker is the most important object in a lecture. Thus, correctly framing the speaker is of great importance. Rules from the professionals state:

1. Give lead room of gaze direction or head orientation for the speaker.
2. Don't move the speaker-tracking camera too often—only move when the speaker moves outside a specified zone.
3. Frame the speaker so that there is half-a-head of room above the speaker's head.

The professionals all agreed on the first two rules, but some did not agreed with the third rule. We decided to try to implement all the three rules nonetheless. After evaluating state-of-the-art techniques, we dropped the first rule. Eye gaze detection or even head orientation estimation are still open research problems in the computer vision research community. For eye gaze detection, one of the best techniques is developed in the IBM BlueEye project [23]. In their system, two near infrared (IR) light sources and a camera are used. By thresholding the frame difference from the two cameras, eye gaze direction can be estimated. Unfortunately, such a technique is not suitable in the lecture room environment. Head orientation estimation is an easier problem than eye gaze detection. However, achieving real-time and reliable results is still far from reach. One of the most recent systems is reported by Stiefelhagen *et. al*. [21]. They achieve good results by using a neural network, but when the testing head is in a different environment than the training cases, accuracy degrades significantly [21]. Because of imperfection of existing technologies, we decided to drop the first rule for the current version of our system. However, our system is flexible enough to incorporate this rule in the future when necessary techniques become available.

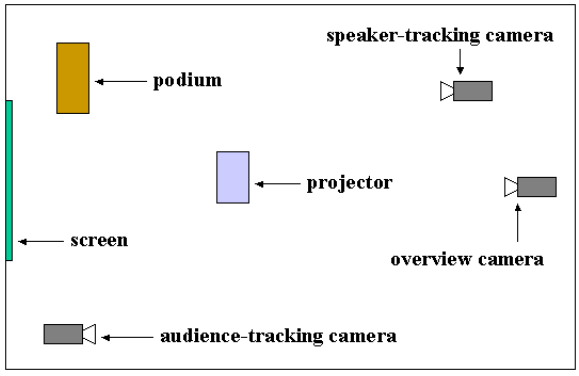


Figure 1. The top view of the lecture room layout

3.3 How to edit

The previous rules concern what an individual cinematographer should do. The following rules govern what a director should do with multiple videos sent from multiple cinematographers.

1. Establishing the shot first. In lecture filming, it is always good to start with an overview shot such that remote audiences get a global context of the environment.
2. Don't make jump cuts—when transitioning from one shot to another, the view and number of people should be significantly different. Failing to do so will generate a jerky and sloppy effect.
3. Don't cut to a camera that is too dark. This will ensure better final video quality.
4. Each shot should be longer than a minimum duration D_{min} (normally four seconds). Violating this rule is distracting for the remote audience.
5. Each shot should be shorter than a maximum duration D_{max}. Violating this rule makes the video boring to watch. The value of D_{max} is different depending on which camera is used.
6. When all other cameras fail, switch to safe back-up cameras (the overview camera in our case).
7. When a person in the audience asks a question, promptly show that person. This is important for remote audience members to follow the lecture.

8. Occasionally, show local audience members for a period of time (e.g., 5 seconds) even if no one asks a question. This will make the final video more interesting to watch.

The first six rules are generic while the last two specifically deal with how to properly select audience shots. For rule 1, our system always starts with an overview camera shot to establish the lecture context. For rule 2, our camera setup (Figure 1) ensures there are no "jump cuts" in our system because all cameras' views are significantly different from each other. For rule 3, the camera's gain control has been carefully calibrated such that shots meet the brightness requirement. As for rules 4 through 8, the following sections provide a detailed discussion of how we implemented them.

4 AUTOMATIC CAMERA MANAGEMENT SYSTEM

This section describes the system we built based on the rules described in the previous section.

4.1 User interface for remote audience

Before going into the details of the entire automatic camera management system, we first describe the user interface for the remote audience (Figure 2).

The left portion of the interface is a standard Microsoft MediaPlayer window, in which the director-edited video is shown. The right portion of the interface displays lecture slides that are synchronized with the video. The outputs of the speaker-tracking camera, the audience-tracking camera and the overview camera are first edited and then displayed in the MediaPlayer window. The output of the slide scan-converter camera is displayed directly on the right-hand side of the window. An obvious alternative to this interface is to eliminate the right window and integrate the output of the slide camera into the MediaPlayer window. However, the interface shown in Figure 2 is the interface already in use by our organization's lecture capture team. Thus, to conduct a controlled study, we used the same interface for our system. Note that because the slides are always shown in the interface, it simplifies editing rules used by the virtual video director, even though our system can also handle the case when slides are not displayed separately.

4.2 System description

In this section, we first describe how a human camera crew films a lecture and then present how we design our system to achieve a similar goal.

Figure 2. The user interface for remote audience

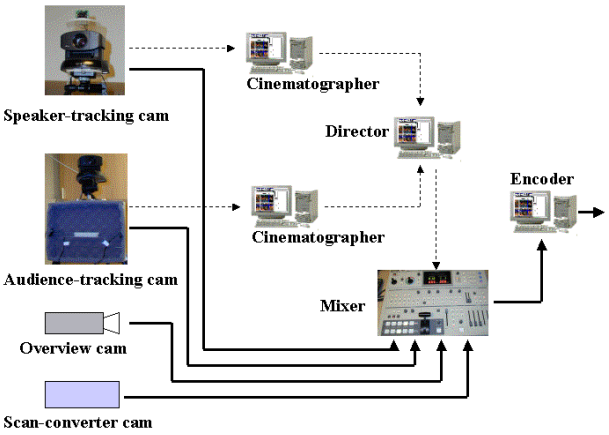


Figure 3. System block diagram. Dashed lines indicate control signals and status signals. Solid lines indicate video data.

To produce a high-quality lecture video, human operators need to perform many tasks, including tracking a moving lecturer, locating a talking audience member, or showing presentation slides. It takes many years of training and experience for a human operator to perform all these tasks. Consequently, high-quality videos are usually produced by a video production team that includes a director and multiple cinematographers. Distributing the video production tasks to different crewmembers and creating final video products through collaboration make the video production process smooth. This strategy is a good reference for a computer-based video production system. Inspired by this idea, we organized our camera management system according to the structure of a video-production team. A block diagram is shown in Figure 3.

Considering different roles taken by the virtual cinematographers and the virtual director, we designed a two-level structure in our system. At the lower level, cinematographers are assigned to different cameras for basic video shooting tasks, such as tracking a lecturer or locating a talking audience. Each cinematographer decides its camera's status, (e.g., "ready" or "not-ready") and reports the status up to the virtual director. At the upper level, the virtual director collects status and events information from all the cinematographers and controls the video mixer to decide which camera is the output camera (Figure 3). The edited lecture video is then encoded for both live broadcasting and on-demand viewing.

4.2.1 Speaker-tracking camera

The speaker-tracking camera follows a lecturer's movement and gestures for a close-up shot. As detailed in Section 2, there are various tracking techniques available. Some ask the lecturer to wear light or electronic wave transmitters to assist the tracking [15,17], which we consider to be obtrusive and inconvenient for the lecturer. Others require manual initialization of color, snakes, or blob for the tracking algorithm [3]. While perfectly valid in their targeted applications, these approaches don't satisfy our goal of building a fully automatic

system. To avoid those issues, in our system we use motion information as our cue to track the speaker. Specifically, we mounted a static wide-angle camera (Figure 4 (a)) on top of the speaker-tracking camera and use the video frame difference from the wide-angle camera to guide the active camera to pan, tilt and zoom. Our tracking scheme does not require the speaker to wear any extra equipment, nor does it require any human assistance. Knowing the exact field of view of the wide-angle camera and the tracking cameras, we can try to maintain a half-head of space above the speaker's head, as stated in rule 3. To comply with rule 2, when the speaker moves too frequently, the camera tries to zoom out.

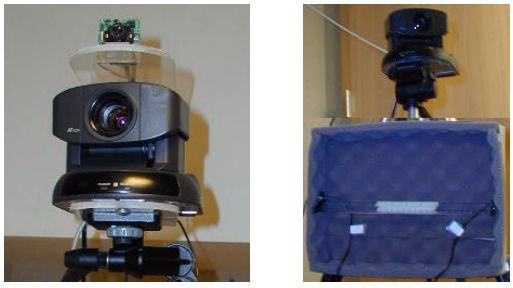
Because of the imperfections of the computer vision techniques, as much as we try to comply with the two framing rules, the camera still sometimes loses track of the speaker or provides a bad shot. The virtual cinematographer for this camera is responsible for deciding the camera's status and reporting the status up to the virtual director. If the speaker is properly framed, the camera reports "ready", otherwise it reports "not ready".

4.2.2 Audience-tracking Camera

As detailed in Section 2, using a microphone array, various approaches exist for locating talking audience members [4,6]. Because the de-convolution approach requires high signal to noise ratio (SNR), we adopt the GCC approach in our system. It uses correlation techniques to find the time difference that an audio signal reaches two microphones. From the time difference and microphone array's geometry, the sound source location can be estimated. Our audience-tracking camera and the microphone array is shown in Figure 4(b).

Although elegant and simple in theory, many practical issues need to be taken into account for microphone array based techniques. For example, a typical lecture room is filled with different sounds, including the lecturer's voice, the projector's fan noise, the computer's noise, and most importantly, reverberations and reflections of sounds. All these issues affect the accuracy of sound-source localization. To improve accuracy, we add an adaptive Wiener filter to suppress stationary noise before the signal is sent to the microphones [14].

There are three statuses for the audience-tracking camera: "ready", "not-ready", and "general". The "ready" status indicates that the audience-tracking camera has correctly located the talking audience member. The "not-ready" status indicates that the camera is still trying to focus on the talking audience member and the shot is not ready for broadcast. The "general" status indicates that no sound is detected from the audience. The "ready" status supports the rule "when a local audience is asking a question, promptly show the audience." In addition, it is important to have a "general" status for this audience camera so that it can support the rule "show the audience occasionally for a shot period even if there is no question".



(a) (b)

Figure 4. (a) Speaker-tracking camera: the top portion is a static wide-angle camera; (b) Audience-tracking camera: the lower portion is a two-microphone array used to estimate sound source location

4.2.3 Virtual video director module

The responsibility of the director is to gather and analyze reports from different cinematographers and to control the video mixer to generate the final video based on video editing rules. The virtual director uses two important components to achieve the goal: a status vector to maintain each cinematographer's status and a finite state machine (FSM) to decide which camera should be chosen.

Because there are three cinematographers in the system, the status vector has three elements, representing the current statuses of the speaker-tracking cinematographer, the audience-tracking cinematographer, and the overview cinematographer, in that order. The first vector element can take two values, i.e., "ready" and "not-ready". The second element can take three values, i.e. "ready", "not-ready", and "general". Because the static overview camera is always ready, the third vector element takes only one value: "ready". Together, they represent a combination of 2x3x1=6 overall statuses for the whole system.

The other component maintained by the virtual director is an FSM. In [11], He *et. al*. proposed a hierarchical FSM structure to simulate a virtual cinematographer in a virtual graphics environment. This work influenced our design of the virtual cinematographer and the virtual video director. Compared with their system, our system works in the real world instead of a virtual world, which imposes many physical constrains on the way we can manipulate cameras and people. For example, it was not possible in our system to obtain a shot from an arbitrary angle.

The virtual director's FSM determines at any given moment which camera is selected as the output camera. Figure 5 shows a three-state FSM where the speaker-tracking camera,

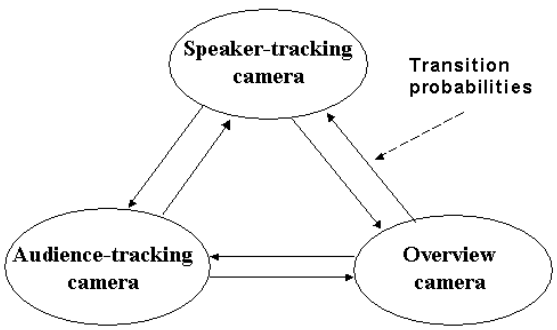


Figure 5. A three-state FSM

audience-tracking camera, and overview camera are each represented by a state. The three states are fully connected to allow any transition from one state to another.

Transiting from one state to another is triggered by events and governed by transition probabilities. To encode video editing rules into the FSM, the following two events are defined:

1. STATUS_CHANGE events: if any of the three cinematographers reports a status change (for example, from "ready" to "not-ready") a STATE-CHANGE event is generated and sent to the virtual director.
2. TIME_EXPIRE events: these events encode Rule 5 in the video editing rules. If a particular camera has been used for too long, a TIME-EXPIRE event is generated. The value for D_{max} depends on the camera in question (Table 1).

Table 1. The values of D_{max} for different cameras and status (in seconds).

Speaker-tracking cam		Audience-tracking cam			Overview cam
Ready	not-ready	ready	not-ready	general	ready
60	0	10	0	5	40

When to transit in the FSM is triggered by the above two event types. Where to transit is determined by the transition probabilities (Figure 5). The set of transition probabilities encode professional rules. For example, the transition probabilities easily allow us to encode the video editing Rules 6, 7 and 8 into the FSM:

- Rule 6: If the statuses for both the speaker-tracking camera and the audience-tracking camera is "not-ready", with probability 1.0 switch to the overview camera (safe back-up).
- Rule 7: If the status of the audience-tracking camera is "ready", i.e., properly framed the audience member who is asking a question, regardless of the statuses of other cameras, with probability 1.0 switch to the audience-tracking camera.
- Rule 8: If the status of the audience-tracking camera is "general", with a small probability, e.g., 0.1, switch to the audience camera to ensure an occasional random audience shot.

The combination of the status vector and the FSM allowed us to easily encode the professional editing rules into the virtual director's knowledge. We have discussed video editing Rules 1, 2 and 3 in the previous section and have discussed Rules 5, 6, 7 and 8 in this section. To ensure Rule 4 (*each shot should be longer than a minimum duration* D_{min}), the virtual director maintains a timer to keep track of how long each shot has been. If the shot length is less than D_{min}, no state transition is made, regardless of the three cameras' statuses.

5 USER STUDY METHODOLOGY

Our user study has two goals. First, we want to evaluate how much each individual video production rule affected the remote audience's viewing experience. Second, we want to compare the overall video quality of our automated system to that of a human operator. The human operator that we use in the study is our organization's regular camera operator, who has many years of experience in photo and video editing.

Our system is deployed in one of our organization's lecture rooms. Originally, there are four cameras in the room, as shown in Figure 1. The camera operator uses those four cameras to record regular lectures. The lectures are broadcast live to employees at their desktops and archived for on-demand viewing.

To make a fair comparison between our system and the human operator, we have restructured the lecture room such that both the human operator and our system have four cameras: they share the same static overview camera and slide projector camera, while both of them have separate speaker-tracking cameras and separate audience-tracking cameras that are placed at close-by locations. They also use independent video mixers.

For user testing, two studies were conducted. The first study is a field study with our organization's employees while the second is a lab study with participants recruited from near by colleges. For the field study, four lectures are used: three are regular technical lectures and the fourth is a general-topic lecture on skydiving held specifically for this study. This skydiving lecture is also used for the lab study.

For the fist study, a total of 24 employees watched one of the four lectures live from their desktops in the same way they would watch any other lectures. While providing a realistic test of the system, this study lacks a controlled environment: remote audience members may have watched the lecture while doing other tasks like reading e-mail or surfing the web. For a more controlled study, we conducted a lab study with eight college students who are not affiliated with our organization. College students are recruited because of their likelihood of watching lectures in their day-to-day life.

The user interface for both studies is shown in Figure 2. All four lectures for the study are captured simultaneously by the human camera operator and our camera management system. When participants watch a lecture, the human operator captured version and our system captured version alternate in the MediaPlayer window (Figure 2). For the three 1.5-hour regular lectures, the two versions alternate every 15 minutes. For the half-hour skydiving lecture, the two versions alternate every 5 minutes. Which version was shown first was randomized. After watching the lecture, participants provided feedback using a survey. Results are reported in the following section.

6 USER STUDY RESULTS

The user studies are intended to test how well the computer performed compared to the human operator. We measure performance using questions based on each of the rules outlined in section 3, as well as two Turing test questions to see if people could determine which video is produced by a person as opposed to our camera management system.

6.1 Tracking the speaker

Two survey questions were asked corresponding to the two speaker-tracking rules (Table 2).

Table 2. Survey results for speaker-tracking quality

(1 = strongly disagree, 5 = strongly agree)	Study session	Human operator			Our system		
		Mean	Median	St. dv.	Mean	Median	St. dv.
The operator followed the speaker smoothly	Field	3.19	3.00	0.83	2.65	2.50	0.88
	Lab	3.50	3.50	0.53	2.87	3.00	0.83
The operator zoomed and centered the camera appropriately	Field	3.11	3.00	0.88	2.67	3.00	1.02
	Lab	4.00	4.00	0.53	3.00	3.50	1.20

A Wilcoxon Signed Ranks Test is used to test the significance of scores between the human operated system and our system. In all cases, the differences are not significant, but there was a clear trend that the human operator is rated higher than the automated system. For the first question, the test yields $z = 1.87$, $p = 0.06$ for the field study and $z = 1.52$, $p = 0.13$ for the lab study. For the second question, the results are $z = 1.81$, $p = 0.07$ and $z = 1.63$, $p = 0.10$ for the two studies, respectively.

6.2 Showing the audience

There are two rules on when to show the audience. We summarize the survey results in Table 3.

Table 3. Survey results for showing the audience

(1 = strongly disagree, 5 = strongly agree)	Study session	Human operator			Our system		
		Mean	Median	St. dv.	Mean	Median	St. dv.
The operator did a good job of showing audience when they asked questions	Field	2.53	2.00	1.01	2.22	2.00	0.94
	Lab	3.25	3.50	0.89	2.87	3.00	0.83
The operator did a good job of showing audience reactions to the speaker	Field	2.83	3.00	0.71	2.55	3.00	0.69
	Lab	3.25	3.00	1.04	2.50	2.50	0.93

Again, Wilcoxon Signed Ranks Tests are used to determine the significance of the difference between the operator and our system. None of the differences are found to be significant. For the first question, the test yields $z = 1.08$, $p = 0.28$ for the field study and $z = 0.76$, $p = 0.45$ for the lab study. For the second question, the test yields $z = 1.40$, $p = 0.16$ and $z = 1.66$, $p = 0.099$, respectively.

The fact that none of the ratings are significantly different is somewhat surprising to us. Because of the imperfections of our microphone array-based audience tracking technique and the noisy lecture room environment, our audience-tracking camera did not find the correct audience member on several occasions. One lab study subject wrote "*when one audience member asked a question, it took them a long time to zoom in*." The study data seems to suggest that people are quite forgiving of the system's audience tracking ability.

6.3 Lighting

The video editing rule 3 tells us "not to cut to a camera that is too dark". We therefore asked the question shown in Table 4. This question is the only one where ratings are higher for our system, although none of the differences are significant. Tests yield $z = 1.83$, $p = .067$ for the field study and $z = 0.38$, $p = 0.71$ for the lab study.

Table 4: Results from the question asking about whether camera shots were sufficiently well lit.

(1 = strongly disagree, 5 = strongly agree)	Study session	Human operator			Our system		
		Mean	Median	St. dv.	Mean	Median	St. dv.
The operator showed camera shots that had sufficient amounts of light.	Field	2.63	2.00	1.07	3.24	4.00	0.94
	Lab	2.63	2.50	0.74	2.75	2.50	1.16

6.4 Overall Perception of the systems

One issue with data from the previous questions is that it may be unreasonable to expect that audience members pay specific attention to individual video production rules. Thus, we also ask overall quality questions. The results are summarized in Table 5, with Wilcoxon test results shown in Table 6. None of the ratings are found to be significantly different except for the question, "the operator did a good job of showing me what I wanted to watch" with the field study subjects. However, there is a general trend in all cases that the human is rated higher than the automated system. We believe this trend can be explained by the imperfect performance of our tracking techniques.

The last question on the survey is a simple Turing test: "do you think each camera operator is a human or computer?" The results are summarized in Table 7.

The data clearly show that participants could not determine which system is the computer and which system is the human at any rate better than chance. For these particular lectures and participants, our system passed the Turing test.

There are two implications. First, the computer-based operator appears not to be making any obvious mistakes repeatedly that the participants can notice. Second, many participants probably realize that even human operators make mistakes – they may sometimes be tired, or distracted, or plain bored by the speaker/content. The latter is not so unusual in practice. For example, many universities use student-hires to manage the cameras and the results can often be quite disastrous (from author's personal experience at Stanford).

7 CONCLUDING REMARKS AND FUTURE WORK

Table 5. Survey results for overall quality

(1 = strongly disagree, 5 = strongly agree)	Study session	Human operator			Our system		
		Mean	Median	St. dv.	Mean	Median	St. dv.
Overall, I liked the way this operator controlled the camera	Field	3.55	4.00	0.83	2.82	3.00	1.18
	Lab	4.00	4.00	0.53	3.00	2.50	1.31
The operator did a good job of showing me what I wanted to watch	Field	3.40	3.00	0.75	2.86	3.00	1.17
	Lab	4.00	4.00	0.53	2.88	2.50	1.13
I liked the frequency with which camera shots changed	Field	3.40	4.00	0.75	2.91	3.00	1.11
	Lab	3.50	3.50	1.20	2.75	2.00	1.39

Table 6. Wilcoxon test results for overall video quality

	Study session	Z	Asymptotic Significance
Overall, I liked the way this operator controlled the camera	Field	-1.847	0.065
	Lab	-1.613	0.107
The operator did a good job of showing me what I wanted to watch	Field	-1.517	0.129
	Lab	-2.081	0.037
I liked the frequency with which camera shots changed	Field	-1.399	0.162
	Lab	-1.633	0.102

We have reported the design, implementation, and user study results of a fully automated camera management system in a lecture room environment. Specifically, we collected and summarized video production rules from professional video producers. We evaluated and implemented the rules by using state-of-the-art techniques.

Our user studies revealed that there is a general trend that participants like the human operator's video better than the automated system's, even though the difference is not significant at the $p < .05$ levels in most cases. However, when it comes to the overall quality, e.g., the Turing test, the participants cannot distinguish one system from the other.

Although the current system has performed well, there are many aspects that can be improved. These include more robust and smoother speaker tracking, quicker and more accurate audience tracking, inclusion of more comprehensive video-production rules (e.g., when a speaker is showing a live demo), and ease of configuration of system. Given the increased availability of bandwidth on Intranets, there is also the possibility of providing new interfaces to viewers. For example, we can provide viewers the flexibility of choosing their own camera view; interfaces that support interaction between remote viewers, speakers, and local audience can also be very valuable [13]. We are continuing to explore these improvements and enhanced interface features.

Successful automatic lecture capture systems will make a huge impact on how people attend and learn from lectures. The cost of the hardware for such automated systems is already very reasonable (< $15K) and is coming down rapidly. By further eliminating the recurring production cost, the primary cost will be disk storage (< $5 for a one hour lecture stored at high quality at 512 Kbps), which is negligible. We will increasingly see a much larger fraction of presentations made accessible online – making a presentation available online will be like turning on a light switch. Techniques for browsing, annotating, and collaborating around online presentations [2,13,16] will allow people to save time and can lead to new models for scaling-up our education system.

Table 7. Turing test results

Study session	Correct	Incorrect	No opinion
Field	17	16	15
Lab	7	7	2

ACKNOWLEDGEMENT

The authors would like to thank Jim Crawford and Dave Crosier for helping us deploy the system in the MSR lecture room and set up the study environment, thanks Li-wei He and Jonathan Grudin for their valuable discussions in system design and evaluation, thanks Barry Brumitt for presenting the "Skydiving" lecture, and John McGrath, John Conrad, Travis Petershagen, Greg Small and Jay Deffinbaugh for sharing their video production rules with us.

REFERENCES

1. Arijon, D. *Grammar of the film language*, New York: Communication arts books, Hastings House Publishers, 1976.
2. Bargeron, D., Gupta, A., Grudin, J., and Sanocki, E., 1999. Annotations for Streaming Video on the Web. *Proc. WWW8.*
3. Baumberg, A. & Hogg, D., An efficient method for contour tracking using active shape models, *TR 94.11*, University of Leeds.
4. Benesty, J., Adaptive eigenvalue decomposition algorithm for passive acoustic source localization, *Journal of Acoustics of America,* vol. 107, January 2000, 384-391.
5. Bianchi, M., AutoAuditorium: a fully automatic, multi-camera system to televise auditorium presentations, *Proc. of Joint DARPA/NIST Smart Spaces Technology Workshop*, July 1998.
6. Brandstein, M., A pitch-based approach to time delay estimation of reverberant speech, *Proc. IEEE ASSP Workshop Appls. Signal Processing Audio Acoustics*, 1997.
7. Brotherton, J. & Abowd, G., Rooms take note: room takes notes!, Proc. AAAI Symposim on Intelligent Environments, 1998, 23-30.
8. Buxton, W., Sellen, A., & Sheasby, M., Interfaces for multiparty videoconferences, *Video-mediated communication* (edited by Finn, K., Sellen, A., & Wilbur, S.), Lawrence Erlbaum Publishers.
9. Cruz, G. & Hill, R., Capturing and playing multimedia events with STREAMS, *Proc. ACM Multimedia'94*, 193-200.
10. Cutler, R. Cutler and Turk, M., View-based Interpretation of Real-time Optical Flow for Gesture Recognition, *IEEE Automatic Face and Gesture Recognition*, April 1998
11. He, L., Cohen, M., & Salesin, D., The virtual cinemato-grapher: a paradigm for automatic real-time camera control and directing, *Proc. of ACM SIGGRAPH'96*, New Orleans, LA. August 1996.
12. He, L., Grudin, J., & Gupta, A., Designing presentations for on-demand viewing, *Proc. of CSCW'00*, Dec. 2000
13. Jancke, G., Grudin, J., Gupta, A., Presenting to local and remote audiences: design and use of the Telep system, *Proc. CHI'00*
14. Jiang, W. & Malvar, H., Adaptive speech noise reduction, *Microsoft Research Technical Report*, Aug. 1999.
15. Mukhopadhyay, S., & Smith, B., Passive Capture and Structuring of Lectures, *Proc. of ACM Multimedia'99*, Orlando.
16. Omuigui, N., He, L., Gupta, A., Grudin, J., & Sanock, E., Time-compression: system concerns, usage, and benefits, *Proc. CHI'99*
17. ParkerVision, http://www.parkervision.com/
18. PictureTel, http://www.picturetel.com/
19. PolyCom, http://www.polycom.com/
20. Stanford Online, http://stanford-onlines.stanford.edu/
21. Stiefelhagen, R., Yang, J., & Waibel, A., Modeling focus of attention for meeting indexing, Proc. of ACM Multiemdia'99.
22. Wang, C. & Brandstein, M., A hybrid real-time face tracking system, *Proc. of ICASSP98*, May 1998, Seattle, 3737-3740.
23. Zhai, S., Morimoto C. & Ihde, S., Manual and gaze input cascaded (MAGIC) pointing, *Proc. of CHI'99*, 246-253.