Videography for Telepresentations

Yong Rui, Anoop Gupta and Jonathan Grudin

October 2, 2001

Technical Report
MSR-TR-2001-92

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

Videography for Telepresentations

Yong Rui, Anoop Gupta, Jonathan Grudin

ABSTRACT

Our goal is to help automate the capture and broadcast of lectures to remote audiences. There are two components to the design of such systems. The first is the technology component that includes the hardware (e.g. video cameras) and associated software (e.g., computer vision software to track the lecturer). The second is the aesthetic component that embodies the rules and idioms that human videographers follow to make the video visually engaging. In this paper we present detailed video-production rules obtained from professional videographers as they critiqued a state-of-the-art lecture room automation system. To facilitate the construction of similar systems by other practitioners in the field, we present generalized rules for different lecture room environments, assuming different numbers and types of cameras. We further discuss areas of gap between what the professional videographers require and what is technologically feasible today.

Keywords

Automated camera management, video, videography, lecture capture and broadcast.

1 INTRODUCTION

Online broadcasting of lectures, both live and on-demand, is increasingly popular in universities and corporations. For instance, MIT's OpenCourseWare (OCW) initiative intends to make all of MIT's courses available on the Web to anyone anywhere [10]. They expect to have about 500 courses available online within the next two years. As an example of corporate education, in 1998 alone, Microsoft supported 367 on-line training lectures with more than 9000 online viewers [8].

As people experience online viewing, expectations of lecture availability increase. However, the cost of capturing content can be prohibitive, primarily due to the cost of hiring professional videographers. In our organization, that cost is over $500 per lecture, while the disk storage cost is less than $1. One way to breach this cost barrier is to build automated camera management systems, where little or no human intervention is needed. Even if the product of such systems does not match the quality of professional videographers, who can still be used for the most important broadcasts, the systems may allow the broadcast (or capture for subsequent viewing) of presentations that otherwise would be available only to physically present audiences. Two major components are needed in such a system:

1. A technology component: The hardware (cameras, microphones, PCs) and software to track and frame lecturers when they move around and point; to promptly detect and frame audience members who ask questions.
2. An aesthetic component: The rules and idioms that human videographers follow to make the video visually engaging. Online audiences have developed expectations based on viewing lectures produced by professional videographers. The automated system should meet such expectations.

In prior work [9], we collected "base-level" video production rules from human experts, built an automated lecture capture system using these rules, and evaluated the performance of the system by regular audiences. The system has since been used on a daily basis in our organization and freed our human videographer for more important tasks. During the past year, the system has continued to improve. In this paper, we extend our research in several new directions:

- Based on previous feedback, we have made important improvements. We present results of re-evaluation.
- While the regular audience was quite positive in evaluation of our previous system [9], we recognized the system had a long way to go. We present further evaluation of our system by four professional videographers. In addition, based on our detailed discussions with them, we now provide a significantly refined set of video-production rules, covering camera positioning, lecturer tracking and framing, audience framing, and shot transitions.
- The system we have currently deployed uses multiple cameras placed in a medium size lecture room. Clearly, human videographers may use different rules as the environment changes, e.g. if there is only one camera vs. multiple cameras or if the room is small or large. Based on discussion with the professional videographers, we provide a set of best practices for 9 common situations. We interestingly find that adapting to new situations requires only a few well-defined changes.
- Finally, some of the rules suggested by the videographers cannot be achieved using existing technology. We present technology feasibility analysis given the state-of-the-art of today's computer vision and signal processing techniques.

Our goal of this paper is to facilitate other practitioners to easily construct similar lecture room automation systems to meet their own needs. The paper is organized as follows. Section 2 is a review of related research on lecture room automation. In section 3, we provide an overview of the system and its recent improvements. We describe the methodology and design of our study in Section 4. In Section 5, we identify high-level results from our surveys of professionals and regular audiences. In Section 6 we present detailed rules suggested by the professionals, and consider the technology feasibility for system automation. In Section 7 we describe how the approach might be applied in different room and camera configurations, then conclude in Section 8.

2 RELATED WORK

Before covering existing lecture room automation systems, we first mention some enabling techniques. Tracking is required to keep the camera focused on the lecturer and to display audience members when they speak. The tracking techniques can be obtrusive or unobtrusive. Obtrusive techniques require people to wear infrared, magnetic or ultra-sound based sensors [11,12]. Unobtrusive tracking is transparent to people and employs computer vision and microphone array techniques [13,14,16]; their quality is approaching that of the obtrusive measures, especially in the context of lecture room automation. We use unobtrusive tracking techniques in our system.

Several projects report on lecture room automation [1,11,16]. In [16], Wang and Brandstein report a real-time head tracker that targets automated video conferencing. Such a system is only one component in our system: the lecturer-tracking module. No attempt is made in their work to construct a complete lecture room automation system.

In [11], Mukhopadhyay and Smith present a lecture-capturing system. They use an obtrusive sensor to track the lecturer and a static camera to capture the podium area. Though there are overlaps between this system and ours, the focus is quite different. Because their system records multiple multimedia streams independently on separate computers, synchronization of those streams is a key focus for them. In our system, various software modules cooperatively film the lecture in a seamless way, and synchronization is not a concern. Furthermore, our main focus is on sophisticated camera management strategies.

Bellcore's AutoAuditorium [1] is a pioneer in lecture room automation. It uses multiple cameras to capture the lecturer, the stage, the screen, and the podium area from the side. A director module selects which video to show to the remote audience based on heuristics. There is overlap between the AutoAuditorium system and ours, but there are substantial differences in the richness of video production rules, the types of tracking modules used, and the overall system architecture.

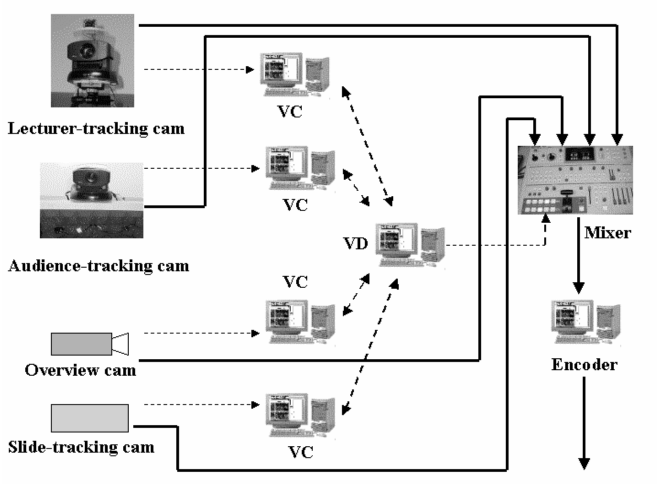


Figure 1. System block diagram. Dashed lines indicate status and command signals. Solid lines indicate video data. VC stands for virtual cinematographers and VD stands virtual director. One thing worth pointing out is that even though we represent various VCs and VD with different computers, they can actually reside in a single computer running multiple threads.

Other aspects of lecture automation are explored in Classroom2000's effort on notes-capturing [3], STREAM's effort on cross-media indexing [5], and Gleicher and Masanz's work on off-line lecture video editing [6]. Work in video mediated communication (e.g., Hydra, LiveWire, Montage, Poleholes, and Brandy Bunch) [4] and Stanford's iRoom [15] is only indirectly relevant to this work.

3 SYSTEM OVERVIEW AND ENHANCEMENT

We briefly describe the component modules in our system, how they work together, and the room in which our system is deployed. We then describe enhancements we have made to the system since our previous study [9].

3.1. System overview

To produce high-quality lecture videos, human operators need to perform many tasks, including tracking a moving lecturer, locating a talking audience member, showing presentation slides, and selecting the most suitable video from multiple cameras. Consequently, high-quality videos are usually produced by a video production team that includes a director and multiple cinematographers. We therefore organize our system according to such a structure. We develop software modules to simulate the cinematographers and the director. They are called the virtual cinematographers (VCs) and the virtual director (VD) in our system. A block diagram of the system is shown in Figure 1.

Considering different roles taken by the VCs and the VD, we develop a two-level structure in our system. At the lower level, VCs are responsible for basic video shooting tasks, such as tracking the lecturer or locating a talking audience. Each VC periodically reports its status to the VD. At the upper level, the VD collects all the necessary information from the VCs, and makes an informed decision on which VC's camera is chosen as the final video output and switches the video mixer to that camera [16]. The edited lecture video is then encoded for both live broadcasting and on-demand viewing. As our first attempt, we have chosen to use one lecturer-tracking VC, one audience-tracking VC, one slide-tracking VC, one overview VC, and one VD in our current system (see Figure 1).

Figure 2 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

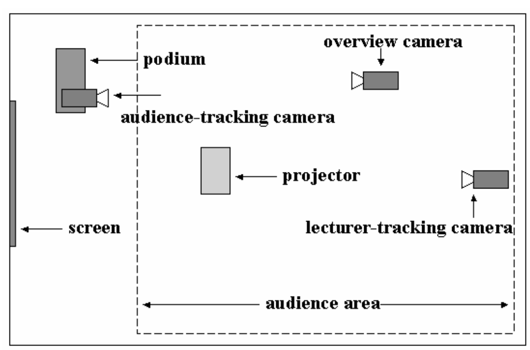


Figure 2. Top view of the lecture room layout.

audience area is in the right-hand side in the figure and includes about 50 seats. There are four cameras in the room: a lecturer-tracking camera (Figure 3a), an audience-tracking camera (Figure 3b), a static overview camera, and a slide-tracking camera (e.g., a scan-converter) that captures whatever is being displayed on the screen. We developed sophisticated computer vision tracking techniques for the lecturer-tracking VC, and sound source localization (SSL) techniques for the audience-tracking VC. The lecturer-tracking technique allows the camera to track and frame the lecturers when they move around. The audience-tracking technique allows the camera to promptly detect and frame the talking audience member. For a detailed description of these techniques, see [16].

3.2. System enhancement

Based on the results of our previous study [9], we made several important enhancements to our system, as described below.

- *Lecturer-framing strategy*. A noticeable problem with the original system was that the lecturer-tracking camera moved too often – it continuously chased a moving lecturer. This can distract viewers, even to the point of motion sickness. The current system uses the history of lecturers' activity to anticipate future locations and frames them accordingly. For example, for a lecturer with an "active" style, the lecturer-tracking VC will zoom out to cover the lecturer's entire activity area instead of continually chasing with a tight shot. This greatly reduces unnecessary camera movement [16].
- *Audience-tracking techniques*. Viewers using the original system commented that the audience-tracking camera responded slowly and inaccurately. Improvement requires overcoming two obstacles: background noise and room reverberation. We subsequently have developed a sophisticated hybrid weighting function for sound source localization. It combines both the maximum likelihood (ML) method, robust to background noise, and the phase transformation (PHAT) method, robust to room reverberation [16].
- *Status and command information*. The original system supported limited status and commands. For example, the VD only informed a VC if its camera was being selected as the output camera, and VCs only reported to the VD if they were ready or not ready. Sophisticated rules, such as audience panning and slide changing, were not sufficiently supported. Our current system employs a more comprehensive set of status and commands. The VCs report the following status information to the VD:
  - **Mode**: Is the camera panning, focusing, static or dead?
  - **Action**: Is the camera aborting, waiting, trying, doing or done with an action that the VD requested?
  - **Scene**: Is there activity in the scene: is the lecturer moving, audience talking, or slide changing?
  - **Score**: How good is this shot, e.g., what is the zoom level of the camera?
  - **Confidence**: How confident is a VC in a decision; e.g., that a question comes from a particular audience area.

  The VD sends the following commands to the VCs:
  - **Mode**: Let the camera do a pan, focus, or static shot;
  - **Status**: If the VC's camera will be selected as preview, on air or off air.

(a) (b)

Figure 3. Cameras used. (a) Lecturer-tracking camera: the top camera is a static wide-angle sensing camera and the bottom camera is a pan/tilt/zoom camera (SONY EVI-D30); (b) Audience-tracking camera: the lower portion is a two-microphone array used to perform sound source location (SSL).

The above status and commands allow the VD and VCs to exchange information effectively and support more sophisticated video production rules. For example, we now provide a slow pan shot of the audience, and the duration for which we focus on an audience member is a function of the confidence we have in the quality of the sound-source localization estimate.

4 METHODOLOGY AND DESIGN OF STUDY

Our system is deployed in one of our organization's lecture rooms (Figure 2). It is used on a daily basis for broadcast of lectures for live and on-demand viewing. To compare our system and human videographers, we restructured the lecture room so that both the videographer and our system had four cameras available: they shared the same static overview and slide projector cameras, while each controlled separate lecturer-tracking and audience-tracking cameras placed at similar locations. They also used independent video mixers. A series of four one-hour lectures on collaboration technologies given by two HCI researchers was used in the study.

There were two groups of participants: professional videographers and the remote audience watching from their offices. The four videographers were recruited from a professional video production company. They are all experienced videographers who have worked in the filed for 3-12 years. Each of them recorded one of the four lectures. After a recording, we interviewed the videographer for two hours. First, we asked them what they had done and what rules they usually followed, pressing for details and reviewing some of their video. They then watched and commented on part of the same presentation as captured by our system. They then filled out and discussed answers to a survey covering system quality (Table 1). Finally, we asked them how they would position and operate cameras in different kinds of rooms and with different levels of equipment, described in Section 7.

Table 1. Survey results. We used a 1-5 scale, where 1 is strongly disagree, 2 disagree, 3 neutral, 4 agree and 5 strongly agree. ***P*** values refer to comparisons of the third and fourth (regular audience rating) columns using a Wilcoxon Test. Results shown as: Median (Mean).

Survey questions	Profess. evaluate system	Audience evaluate system	Audience evaluate profess.	***P***
1. Shot change frequency	2.5 (2.8)	3.0 (2.6)	4.0 (3.4)	0.01
2. Framed shots well	1.5 (1.8)	3.0 (2.7)	4.0 (3.6)	0.02
3. Followed lecturer smoothly	2.0 (2.0)	2.0 (2.3)	4.0 (3.5)	0.01
4. Showed audience questioner	3.5 (3.5)	3.0 (2.8)	2.0 (2.7)	0.73
5. Showed audience reaction	4.0 (3.5)	2.0 (2.3)	2.0 (2.3)	1.00
6. Showed facial expression	3.0 (2.8)	2.5 (2.8)	3.0 (3.2)	0.23
7. Showed gestures	3.5 (3.2)	4.0 (3.2)	4.0 (3.5)	0.06
8. Showed what I wanted to watch	3.0 (3.2)	4.0 (3.4)	4.0 (3.9)	>.05
9. Overall quality	2.0 (2.0)	3.0 (2.8)	4.0 (3.8)	<.01
10. As compared with previous experience	1.5 (1.5)	3.0 (3.1)	3.0 (3.6)	0.11

In addition, 18 employees in the organization watched one or more of the lectures from their offices at their own initiative and filled out the survey described below. The interface they saw is shown in Figure 4. The left portion is a standard Microsoft MediaPlayer window. The outputs of lecture-tracking camera, audience-tracking camera, and overview camera were first edited by the VD and then displayed in this window. The output of the slide-tracking camera was displayed to the right. Each lecture was captured simultaneously by a videographer and by our system. Remote viewers were told that two videographers, designated A and B (see bottom-left portion of Figure 4), would alternate every 10 minutes, and were asked to pay attention and rate the two following the lecture.

5 EVALUATION RESULTS

This section covers highlights of professionals evaluating our system, and remote audience evaluating both our system and the professionals. The results are presented in Table 1. The first seven questions relate to individual aspects of lecture-recording practice, and the last three questions focus on overall lecture-watching experience.

Individual aspects

The professionals rated our system quite well for questions 4, 5 and 7 (median ratings of 3.5 to 4.0; see Table 1 for all means). They gave us the highest ratings for Q4 and Q5 relating to capturing audience reactions/questions. In fact, their scores

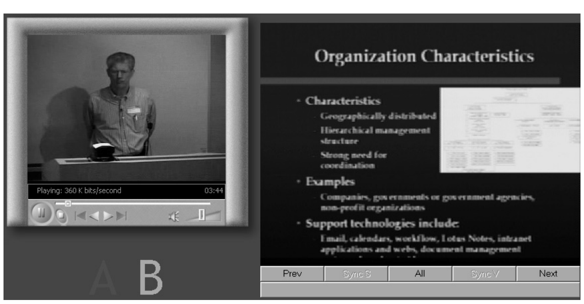


Figure 4. The user interface for remote audience.

were even higher than those given by the remote audience, among the few exceptions in the whole survey (see Table 1) -- they said many times our system found the questioner faster than they did. Q7 related to showing lecturer gestures. Both the professionals and the remote audience gave our system high scores of 3.5 and 4.0, respectively. They thought our system's medium-to-close lecturer shots caught the gestures well.

The professionals gave our system moderate scores on Q1 (shot change frequency: 2.5) and Q6 (showing facial expressions: 3.0). On shot change frequency, the professionals felt that there was a reasonably wide range based on personal preference, and we were within that range. The audience, however, significantly preferred videographers shot change frequency (p=0.01). Some videographers did point out to us that our shot change frequency was somewhat mechanical (predictable). For Q6, because our lecturer shots were not very tight, they covered the lecturer's gestures well (Q7), but were less effective in capturing lecturer's facial expressions (Q6).

The videographers gave our system very low scores on Q2 and Q3. They were most sensitive to Q2 on framing. This is where they have spent years perfecting their skills [1], and they made comments like why was the corner of screen showing in lecturer shot (see Figure 5b). This was recognized by remote audience as well, and they thought the videographers framing was significantly better than our system's (p=0.02).

On Q3 (following lecturer smoothly) the videographers were critical when our system let the lecturer get out of the frame a few times and then tried to catch up the lecturer again. The remote audience also recognized this, and they thought the videographers' lecturer tracking was significantly better than our system's (p=0.01).

Overall experience

Individual aspects of lecture recording practice are important, but the overall experience is even more important to the end users. We asked three overall quality questions. Q8 put less emphasis on aesthetics and asked "The operator did a good job of showing me what I wanted to watch". The professionals gave our system a score of 3.0 and the remote audience gave us their highest score of 4.0. One of the professionals said *"Nobody running the camera … this is awesome … just the concept is awesome"*. Another said *"It did exactly what it was supposed to do … it documented the lecturer, it went to the questioner when there was a question"*.

Our second overall question (Q9) had greater emphasis on aesthetics and asked, "Overall, I liked the way the operator controlled the camera". The videographers clearly disagreed with our proposition giving a score of 2.0. In detailed discussion, lack of aesthetic framing, smooth tracking of lecturer, and semantically motivated shot cuts were the primary reasons. The remote audience also clearly preferred the overall quality of video from the professionals (p < .01), while giving our system a neutral score of 3.0.

Our third overall question (Q10) focused on how the quality compared to their previous online experiences. The audience thought the quality of both our system and professionals was equivalent to their previous experiences, giving scores of 3.0.

It is interesting to note that while the ratings on the individual aspects of our system were low, the ratings of our system's overall quality were about neutral or higher as judged by the end-users – they never gave a >4.0 score even for professionals. These ratings provide evidence that our system was doing a good job satisfying remote audience's basic lecture-watching need. Given that many organizations do not have the luxury of deploying professionals for recording lectures – e.g. most Stanford online lectures are filmed by undergraduate students – the current system can already be of significant value.

6 FINE-GRAINED RULES & TECHNOLOGY FEASIBILITY

Most of the existing systems have not been based on systematic study of video production rules or the corresponding technical feasibility. The eight high-level rules employed in our own previous effort proved insufficiently comprehensive [9]. In this section we consider fine-grained rules for video production based on our interviews with the professional videographers (represented as ***A***, ***B***, ***C*** and ***D***).

6.1. Camera positioning rules

The professionals generally favored positioning cameras about two meters from the floor, close to eye level but high enough to avoid being blocked by people standing or walking. However, ***A*** and ***C*** felt that ceiling-mounted cameras, as used in our room, were acceptable as well. ***A*** also liked our podium-mounted audience-tracking camera (Figure 3b). All videographers wanted audience-tracking cameras in the front of the room and lecturer-tracking cameras in the back. However, with the podium toward one side of the room, two videographers (***A*** and ***B***) preferred direct face-on camera positioning and two (***C*** and ***D***) preferred positioning from an angle (shown in Figure 6a). Summarized as rules for camera positioning:

Rule 1.1. *Place cameras at the best angle to view the target. This view may be straight on or at a slight angle.*

Rule 1.2 *Lecturer-tracking and overview cameras should be close to eye level but may be raised to avoid obstructions from audience.*

Rule 1.3. *Audience-tracking cameras should be high enough to allow framing of all audience area seating.*

Two rules important in filming were also discussed:

Rule 1.4. *A camera should avoid a view of another camera.* This rule is essential in film, and it is distracting if a videographer is visible behind a camera. But a small camera attached to the podium or wall may not be distracting, and one in the ceiling can be completely out of view. Two of the videographers noted that they followed this rule, but the other two didn't. ***A*** in particular noted that our podium-mounted audience-tracking camera, although in range of the lecturer-tracking camera, was unobtrusive.

Rule 1.5. *Camera shots should avoid crossing "the line of interest"-- This line can be the line linking two people, the line a person is moving along, or the line a person is facing* [1]. For example, if a shot of a subject is taken from one side of the line, subsequent shots should be taken from the same side [7]. It was noted by the videographers that rule 1.5 did not apply in our setting because the cameras did not focus on the same subject.

6.2. Lecturer tracking and framing rules

Rule 2.1. *Keep a tight or medium head shot with proper space (half a head) above the head.* The videographers all noted failures of our system to center lecturers properly, failing to provide the proper 10 to 15 centimeters space above the head and sometimes losing the lecturer entirely (see Figure 5). They differed in the tightness of shots on the lecturer though; two got very close despite the greater effort to track movement and risk of losing a lecturer who moves suddenly.

Rule 2.2. *Center the lecturer most of the time but give lead room for a lecturer's gaze direction or head orientation.* For example, when a lecturer points or gestures, move the camera to balance the frame. ***A*** explicitly mentioned the "rule of thirds" and ***B*** emphasized "picture composition."

Rule 2.3. *Track the lecturer as smoothly as possible, so that for small lecturer movements camera motion is almost unnoticed by remote audiences.* As compared to our system the videographers had tremendous ability to predict the extent to which the lecturer was going to move and they panned the camera with butter-like smoothness.

Rule 2.4. *Whether to track a lecturer or to switch to a different shot depends on the context.* For example, ***B*** said that if a lecturer walked over quickly to point to a slide and then returned to the podium, he would transition to an overview shot and then back to a lecturer shot. But if the lecturer walked slowly over and seemed likely to remain near the slide, he would track the lecturer.

Rule 2.5. *If smooth tracking cannot be achieved, restrict the movement of the lecturer-tracking camera to when a lecturer moves outside a specified zone.* Alternatively, they suggested zooming out a little, so that smaller or no pans would be used. The lecturer-framing strategy we describe in Section 3.2 partly relies on this strategy.

Automation Feasibility

While base-level lecturer tracking and framing rules are achievable, as we did in our system, many of the advanced rules will not be easy to address in the near term future. For rule 2.2, real-time eye gaze detection and head orientation estimation are still open research problems in computer vision. For instance, for eye gaze detection, an effective technique is the two IR light sources used in the IBM BlueEye project [18]. Unfortunately, such a technique is not suitable in this application.

For rules 2.1-2.4, the system must have a good predictive model of lecturer's position and movements, and the pan/tilt/zoom camera must be smoothly controllable. Unfortunately, neither is easily satisfied. Because the wide-angle sensing camera has a large field of view (see Figure 3a), it has very limited resolution of the lecturer. Given the low resolution, existing techniques can only locate the lecturer roughly. In addition, current tracking cameras on the market, e.g., Sony's EVI D30 or Canon's VC-C3, do not provide smooth tracking in the absolute position mode. Given the above analysis, instead of completely satisfying all the rules, we focus on rule 2.5 and implement others as much as possible.

6.3. Audience tracking and framing rules

All videographers agreed on the desirability of quickly showing an audience member who commented or asked a question if that person could be located in time. Beyond that they differed. At one extreme, ***B*** cut to an audience for comedic reactions or to show note-taking or attentive viewing. In contrast, ***D*** avoided audience reaction shots and favored returning to the lecturer quickly after a question was posed. Thus, agreement was limited to the first two of these rules:

Rule 3.1. *Promptly show audience questioners. If unable to locate the person, show a wide audience shot or remain with the lecturer.*

Rule 3.2. *Do not show relatively empty audience shots.* (See Figure 5d for a violation by our system.)

Rule 3.3. *Occasionally show local audience members for several seconds even if no one asks a question.*

B, perhaps the most artistically inclined, endorsed rule 3.3. He favored occasional wide shots and slow panning shots of the audience – the duration of pans varied based on how many people were seated together. The other videographers largely disagreed, arguing that the goal was to document the lecture, not the audience. However, ***A*** and ***C*** were not dogmatic: the former volunteered that he liked our system's audience pan shots a lot, and the latter said he might have panned the audience on occasion if it were larger. The strongest position was that of ***D***, who said of our system's occasional panning of the audience, "*You changed the tire correctly, but it was not flat.*"

As noted in Section 5, our system was relatively highly rated on the audience shots by the remote viewers and even more highly rated by the professionals. For one thing, when the professionals were unfamiliar with the faces, voices, and habits of the audience, our system was faster in locating questioners.

Figure 5. Examples of bad framing. (a). Not centered. (b). Inclusion of the screen edge. (c). Too much headroom. (d). Showing an almost empty audience shot.

Automation feasibility.

Our sophisticated SSL technique (Section 3.2) [16] allows the audience-tracking camera to promptly focus on the talking audience member most of the time. However, detecting "comedic reactions" or "attentive viewing", as ***B*** suggested, is another story. It requires content understanding and emotion recognition which are still open research problems.

On the other hand, detecting roughly how many people are there to avoid "empty audience shots" may not be very difficult. For example, if the lighting is sufficient, face detection algorithms may tell us the number of people. If the lighting is not sufficient, by cumulating the number of SSL results over time, we can also get a rough estimate of the number of audience members.

6.4. Shot transition rules

Some videographers thought our system maintained a good rate of shot change; others thought it changed shots too frequently. This is of course tied to rule 3.3, discussed above. ***D*** further noted that "… keep the shots mixed up so (viewers) can't totally predict …" All videographers felt that there should be minimum and maximum durations for shots to avoid distracting or boring viewers, although in practice they allow quite long (up to a few minutes) medium-close shots of the lecturer.

Rule 4.1. *Maintain reasonably frequent shot changes, though avoid making the shot change sequences mechanical/ predictable.*

Rule 4.2. *Each shot should be longer than a minimum duration, e.g., 3~5 seconds, to avoid distracting viewers.*

Rule 4.3. *The typical to maximum duration of a shot may vary quite a bit based on shot type.* For instance, it can be up to a few minutes for lecturer-tracking shots and up to 7-10 seconds for overview shots. For audience shots the durations mentioned are in the range 4-10 seconds for a static shot where no question is being asked, or the duration of the whole question if a question is being asked, and for panning shots the duration varies based on the number of people that the pan covers (slow enough so that viewers can see each audience's face).

Rule 4.4. *Shot transitions should be motivated*.

Rule 4.5. *A good time for a shot change is when a lecturer finishes a concept or thought or an audience member finishes a question.*

Shot changes can be based on duration, e.g., rule 4.3, but more advanced shot changes are based on events. Unmotivated shot changes, as in a random switch from the lecturer-tracking to the overview camera, can "give the impression that the director is bored." As indicated above, opinions differed some as to what can motivate a transition. Emergencies do motivate shifts to the overview camera, such as when the lecturer-tracking camera loses track of the lecturer, or the audience-tracking camera is being adjusted.

Interestingly, the overview camera not only can be used as a safety backup, it can also be used to capture gestures and slide content. In fact, ***B*** zoomed in the overview camera a little during the talk to cover the lecturer and provide readable slides, although we requested them avoid manipulating the shared overview camera. In summary:

Rule 4.6. *An overview shot is a good safety backup.*
Rule 4.7. *An overview shot can frame a lecturer's gestures and capture useful information (e.g., slide content).*

If the overview camera is a static camera, there is a tradeoff between rules 4.6 and 4.7. If the camera is too zoomed in, it will not serve as a safety backup; but if it is too zoomed out, the shot is less interesting and slides less readable.

Rule 4.8*. Don't make jump cuts—when transitioning from one shot to another, the view and number of people should differ significantly.* Our system occasionally switched from a zoomed-out wide lecturer view to a similar shot from the overview camera. That was an example of "jump cuts" and appeared jarring.

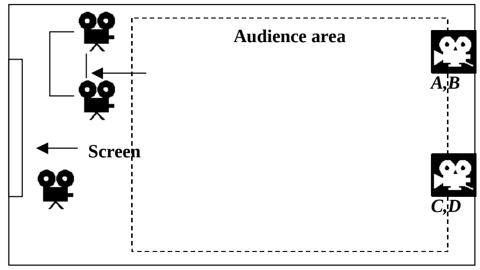
Rule 4.9*. Use the overview camera to provide establishing and closing shots.* The professionals disagreed over the value of overview shots at the beginning and end of a lecture. ***A*** explicitly avoided them and ***D*** explicitly endorsed them.

Automation feasibility.

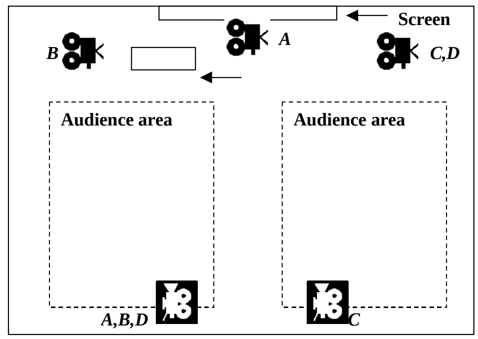
Maintaining minimum/maximum shot duration and good shot transition pace is relatively easy to achieve. Similarly, by carefully incorporating the camera's zoom level, we can avoid "jump cuts". However, for "motivated shot transitions," current techniques can only provide a partial solution. For example, we can easily estimate if a lecturer moves a lot or not to determine if we should cut to the overview shot. It would also be nice if we could detect if the lecturer was pointing to the screen, which would be a good time to make motivated transitions. As for detecting if a lecturer finishes his/her thoughts, that is an extremely difficult problem. It is requires high-accuracy speech recognition in noisy environment and real-time natural language understanding, both needs years of research.

7 GENERALIZATION TO DIFFERENT SETTINGS

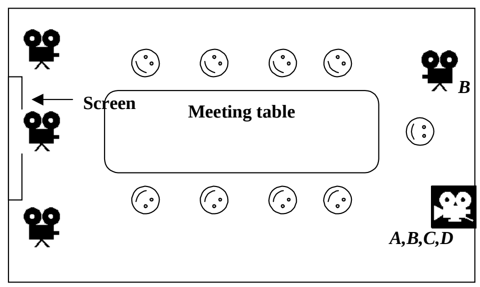
Our discussion so far has focused on a medium sized lecture room with multiple cameras available for filming. For this technology to be widespread, we need to be able to accommodate many different types of lecture venues and different levels of technology investment, e.g., number of cameras. We asked the videographers how the rules and camera setup would change in these different environments. We asked them to consider three common venue types: R1) medium size lecture room (~50 people), R2) large auditorium (~100+ people), and R3) small meeting room (~10-20 people). The arrangements are shown in Figure 6. We asked them to also consider three levels of technology investment: C1) A single dual-function lecturer-tracking plus overview camera – recall from Figure 3a that our lecturer tracking camera already has a wide-angle camera on the top; C2) two cameras – C1 plus a slide/screen capturing camera; and C3) three cameras – C2 configuration plus an audience-tracking camera. This leads to 9 combinations (R1-R3 x C1-C3). For simplicity, we will use R1C1 to represent the case where a single lecturer-tracking/overview camera (C1) is used in the lecture room



(a). Medium sized lecture room camera position (R1C3)



(b). Large Auditorium camera position (R2C3)



(c). Meeting room camera position (R3C3)

Figure 6. The three room configurations. White cameras are lecturer-tracking/overview camera units, black cameras are audience-tracking cameras. Letters indicate the different videographers' choices. Slide cameras are implicit -- they just capture the screens.

(R1). Similarly, R3C3 means camera configuration C3 is used in room type R3.

7.1 Camera Positioning

Figure 6 shows camera positions proposed by videographers ***A***, ***B***, ***C****, and* ***D***. We noted that in cases where the audience camera or slide camera was not present, the videographers did not suggest changing the position of the lecturer-tracking/overview camera. We therefore only need to draw cases R1C3, R2C3 and R3C3 in Figure 6 to cover all the 9 combinations.

The layout in Figure 6a (R1C3) represents the lecture room where our system is installed. The videographers' assessment of it was already described in detail in Section 6 – for instance, the differing preferences for face-on and angled views of a lecturer.

For the auditorium (6b: R2C3), there was little change. It was noted that because the lecturer-tracking cameras were at a greater distance, they could be higher from the floor.

In the meeting room (6c: R3C3), the audience faces in different directions and the cameras are closer to audience/lecturer, leading to more changes. When needed, the lecturer-tracking/overview camera can also be used to show half of the audience. ***A*** and ***B*** placed the audience-tracking camera to view the other half of audience, and ***C***'s ceiling-mounted camera could view them all. ***D*** only captured half the audience face-on. ***D***'s placement avoided cameras viewing one another and eliminated some violations of "the line of interest" rule, as did ***B***'s interesting choice.

7.2 Shots and Transitions

In Section 6, we have discussed in great detail the shots and transitions for configuration R1C3. Based on our interviews with the professionals, interestingly, most of the rules for R1C3 also generalized to R2C3 and R3C3. There was only one main exception corresponding to the meeting room (R3C3). This results because the audience-tracking camera often can only see half of the audience as discussed above. If an audience in such a blind zone was to ask a question, the videographers suggested two options. The first option was simply not to transition to the audience shot. The second option was if the lecturer-tracking camera could have covered the shot then it could be used for that purpose, but using the overview camera as the transition. Then the videographers would follow the reversed sequence back, i.e. audience -> overview -> lecturer. Here, as a reminder, we would like to emphasize again that the lecture-tracking/overview camera is a dual-function unit (Figure 3a) – the top static camera can provide overview shots while the bottom camera is pan/tilt/zoom controllable.

For all the three rooms R1-R3, the rules for case C2 were similar to those in C3. However, because the audience camera was not available at all in C2, there were a few rule changes regarding the audience shots. One was to simply ignore the audience shots. The other was to use the lecture-tracking camera to cover the audience when possible, and go through the following shot transitions: lecturer -> overview -> audience -> overview -> lecturer.

For all the three rooms R1-R3, case C1 is the most challenging, because the videographers had to rely on the lecture-tracking/overview dual-function unit to cover lecturer, slide, and audience. Using case C2 as a reference, the rule changes are the following, mostly on how to capture slides:

- Adjust the position of the overview camera if possible to cover both slides and lecturer more evenly. Use the lecturer-tracking camera to capture the lecturer, and switch to the overview camera at the slide transitions.
- Use the lecturer-tracking camera mostly to capture the lecturer, but to capture the slides at slide transitions. Switch to the overview camera when the lecture-tracking camera is adjusting between the lecturer and the slides.

To summarize this section, three findings make the generalization of our system to other room and camera configurations easy. First, adding/deleting a camera normally won't affect the positioning of existing cameras. Second, for all the three rooms R1-R3, to downgrade the equipment investment from C3 to C2 or C1, there are only a few well-defined rule changes. Third, the camera positioning and rules for the auditorium (R2) and meeting room (R3) are similar to those for the lecture room (R1), which has been well studied. These findings should greatly facilitate other practitioners to construct their own systems.

8 CONCLUDING REMARKS

We have reported evaluations of a daily-used state-of-the-art lecture room automation system. Our goal is to make the system and techniques readily usable by other practitioners. Towards this end, we have presented fine-grained video production rules and analyzed their automation feasibilities. While advanced rules may still require years of research, basic rules have already been realized today and can satisfy people's basic lecture-watching need. Furthermore, to meet requirements of different room and camera configurations, we have further reported new rules obtained from the professionals and have interestingly found that the changes are few and well defined.

Successful lecture room automation systems will make a major impact on how people attend and learn from lectures. The cost of hardware for such systems is already reasonable (under $15K) and is continuously dropping. By further cutting the cost of hiring human videographers, we expect to see a growing number of presentations made accessible online in more and more universities and corporations.

REFERENCES

1. Arijon, D. Grammar of the film language, New York: Communication arts books, Hastings House Publishers, 1976
2. Bianchi, M., AutoAuditorium: a fully automatic, multi-camera system to televise auditorium presentations, *Proc. of Joint DARPA/NIST Smart Spaces Technology Workshop*, July 1998.
3. Brotherton, J. & Abowd, G., Rooms take note: room takes notes!, Proc. AAAI Symposium on Intelligent Environments, 1998, 23-30.
4. Buxton, W., Sellen, A., & Sheasby, M., Interfaces for multiparty videoconferences, *Video-mediated communication* (edited by Finn, K., Sellen, A., & Wilbur, S.), Lawrence Erlbaum Publishers.
5. Cruz, G. & Hill, R., Capturing and playing multimedia events with STREAMS, *Proc. ACM Multimedia'94*, 193-200.
6. Gleicher M., & Masanz, J., Towards virtual videography, *Proc. of ACM Multimedia'00*, LA, Nov. 2000
7. He, L., Cohen, M., & Salesin, D., The virtual cinematographer: a paradigm for automatic real-time camera control and directing, *Proc. of ACM SIGGRAPH'96*, New Orleans, LA. August 1996.
8. He, L., Grudin, J., & Gupta, A., Designing presentations for on-demand viewing, *Proc. of CSCW'00*, Dec. 2000
9. Liu, Q., Rui, Y., Gupta, A. & Cadiz, JJ, Automating camera management for lecture room environments, *Proc. of ACM CHI 2001*.
10. MIT-OCW, http://web.mit.edu/ocw/
11. Mukhopadhyay, S., & Smith, B., Passive Capture and Structuring of Lectures, *Proc. of ACM Multimedia'99*, Orlando.
12. ParkerVision, http://www.parkervision.com/
13. PictureTel, http://www.picturetel.com/
14. PolyCom, http://www.polycom.com/
15. Stanford iRoom, http://graphics.stanford.edu/projects/iwork/
16. Rui, Y., He, L., Gupta, A. & Lu, Q., Building an Intelligent Camera Management System, *Proc. of ACM Multimedia'01*.
17. Wang, C. & Brandstein, M., A hybrid real-time face tracking system, *Proc. of ICASSP98*, May 1998, Seattle, 3737-3740.
18. Zhai, S., Morimoto C. & Ihde, S., Manual and gaze input cascaded (MAGIC) pointing, *Proc. of CHI'99*, 246-253.