Automating lecture capture and broadcast: technology and videography

Yong Rui, Anoop Gupta, Jonathan Grudin, Liwei He

Microsoft Research, One Microsoft Way, Redmond, WA 98052-6399, USA
e-mail: {yongrui, anoop, jgrudin, lhe}@microsoft.com

Abstract. Our goal is to help automate the capture and broadcast of lectures to online audiences. Such systems have two interrelated design components. The technology component includes hardware and associated software. The aesthetic component comprises the rules and idioms that human videographers follow to make a video visually engaging; these rules guide hardware placement and software algorithms. We report the design of a complete system that captures and broadcasts lectures automatically and report on a user study and a detailed set of video-production rules obtained from professional videographers who critiqued the system, which has been deployed in our organization for 2 years. We describe how the system can be generalized to a variety of lecture room environments differing in room size and number of cameras. We also discuss gaps between what professional videographers do and what is technologically feasible today.

Key words: Lecture capture – Automated camera management – Video – Videography – Virtual director

1 Introduction

Online broadcasting of lectures and presentations, live and on-demand, is increasingly popular in universities and corporations as a way of overcoming temporal and spatial constraints on live attendance. For instance, at Stanford University, lectures from over 50 courses are made available online every quarter [25]. Similarly, the University of Washington's Professional Master's Program offers its courses online to help people further their educational and professional goals [20]. As an example of corporate education, Microsoft supported 367 online training lectures with more than 9000 online viewers in 1999 alone [12].

Although online viewing provides a convenient way for people to view lectures at a more convenient time and location, the cost of capturing content can be prohibitive, primarily due to the cost of hiring professional videographers. This could be addressed with automated camera management systems requiring little or no human intervention. Even if the resulting quality does not match that of professional videographers,

the professionals can handle the most important broadcasts, with a system capturing presentations that otherwise would be available only to physically present audiences.

Two major components are needed in such a system:

1. A technology component: Hardware (cameras, microphones, and computers that control them) and software to track and frame lecturers when they move around and point and to detect and frame audience members who ask questions.
2. An aesthetic component: Rules and idioms that professionals follow to make the video visually engaging. The automated system should make every effort to meet expectations that online audiences have based on viewing lectures produced by professional videographers.

These components are interrelated: aesthetic choices will depend on the available hardware and software, and the resulting rules must in turn be represented in software and hardware. In this paper, we address both components. Specifically, we present a complete system that automatically captures and broadcasts lectures and a set of video-production rules obtained from professional videographers who critiqued it. The system [15,21,22] has been used on a daily basis in our organization for about 2 years, allowing more lectures to be captured than our human videographer could have handled.

The goal of this paper is to share our experience building such a system with the practitioners in the field to facilitate their construction of similar systems, and to identify unsolved problems requiring further research. The rest of the paper is organized as follows. Section 2 reviews research on lecture room automation. In Sect. 3, we present the system and its components, including the hardware and the lecturer-tracking, audience-tracking, and virtual director software modules. In Sect. 4, we describe the design and results of a user study. In Sect. 5, we present rules obtained from professional videographers and analyze the feasibility of automating them with today's technologies. In Sect. 6, we describe how the system can be generalized to a variety of lecture room environments that differ in room size and number of cameras. Concluding remarks follow in Sect. 7.

2 Related work

In this section, we provide a brief review of related work on individual tracking techniques, videography rules, and existing automated lecture capture systems.

2.1 Tracking techniques

Tracking technology is required both to keep the camera focused on the lecturer and to display audience members when they talk. There are obtrusive tracking techniques, in which people wear infrared, magnetic, or ultrasound sensors, and unobtrusive tracking techniques, which rely on computer vision and microphone arrays.

Obtrusive tracking devices emit electric or magnetic signals that are used by a nearby receiver unit to locate the lecturer. This technique has been used in commercial products [17] and research prototypes [16]. Although obtrusive tracking is usually reliable, wearing an extra device during a lecture can be inconvenient.

A rich literature in computer-vision techniques supports unobtrusive tracking. These include skin-color-based tracking [26], motion-based tracking [8], and shape-based tracking [2]. Another unobtrusive technique, based on microphone array sound source localization (SSL), is most suited for locating talking audience members in a lecture room. Various SSL techniques exist as research prototypes [3,5,14,27] and commercial products (e.g., PictureTel [18] and PolyCom [19]).

To summarize, obtrusive solutions are more reliable but less convenient. The quality of unobtrusive vision and microphone-array-based techniques is quickly approaching that of obtrusive solutions, especially in the context of lecture room camera management.

2.2 Videography rules

Various directing rules developed in the film industry [1] and for graphics avatar systems [11] are loosely related to our work. However, there is a major difference. In film and avatar systems, a director has multiple physically or virtually *movable* cameras that can shoot a scene from almost any angle. In contrast, our camera shots are constrained: we have pan/tilt/zoom cameras, but they are physically anchored in the room. Therefore, many film industry rules are not applicable to a lecture capture system and serve only as high-level considerations.

2.3 Related systems

In [16], Mukhopadhyay and Smith present a lecture-capturing system that uses a magnetic device to track the lecturer and a static camera to capture the podium area. Because their system records multiple multimedia streams independently on separate computers, synchronization of those streams was their key focus. In our system, various software modules cooperatively film the lecture seamlessly, so synchronization is not a concern. Our main focus is on sophisticated camera management strategies.

Bellcore's AutoAuditorium [4] 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. The AutoAuditorium system concerns overlap with ours but differ substantially in the richness of video production rules, the types of tracking modules used, and the overall system architecture. Furthermore, no user study of AutoAuditorium is available. Our system, in contrast, has been in continuous evolution and use for the past 2 years, as described below.

Liu et al. recently developed a lecture broadcasting system that allows multiple operators to manually control one or more lecture room cameras [14]. They propose an optimization algorithm to mediate between different operator requests. This is an interesting way to utilize human knowledge. Song et al. at UC-Berkeley independently developed a similar technique to allow multiple human controls of a remote camera [23,24]. These systems focused on improving manual control by nonprofessionals; our system aims to automate the lecture capture and broadcast process.

Several other lecture room automation projects focus on different aspects of classroom experience. For example, Classroom2000 [6] focuses on recording notes in a class. It also captures audio and video, but by using a single fixed camera it limits the coverage and avoids the issues addressed in our research. Cruz and Hill [7] discuss efforts on cross-media indexing. Gleicher and Masanz [10] deal with offline lecture video editing. There is also a rich but tangential literature on video-mediated communication systems (e.g., Hydra, LiveWire, Montage, Portholes, and Brady Bunch) surveyed in [9].

To summarize, few prior efforts to build automated camera management systems involve a complete system. There exist diverse computer-vision and microphone-array tracking techniques, but their integration in a lecture room environment has not been deeply studied. Furthermore, there is almost no attempt to construct software modules based on observations of professional video production teams. Finally, there are few systematic studies of professional video production rules. This paper focuses on the integration of individual tracking techniques in lecture room environments, detailing the design of a camera management framework that studies indicate can achieve results approaching those of a professional video production team. In the next few sections, we present the system/technology and the aesthetic/videography components.

3 System and technology component

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 cameramen. We organize our system according to such a two-level structure (Fig. 1). At the lower level, multiple virtual cameramen (VC) are responsible for basic video shooting tasks such as tracking the lecturer or locating a talking audience. At the upper level, a virtual director (VD) col-
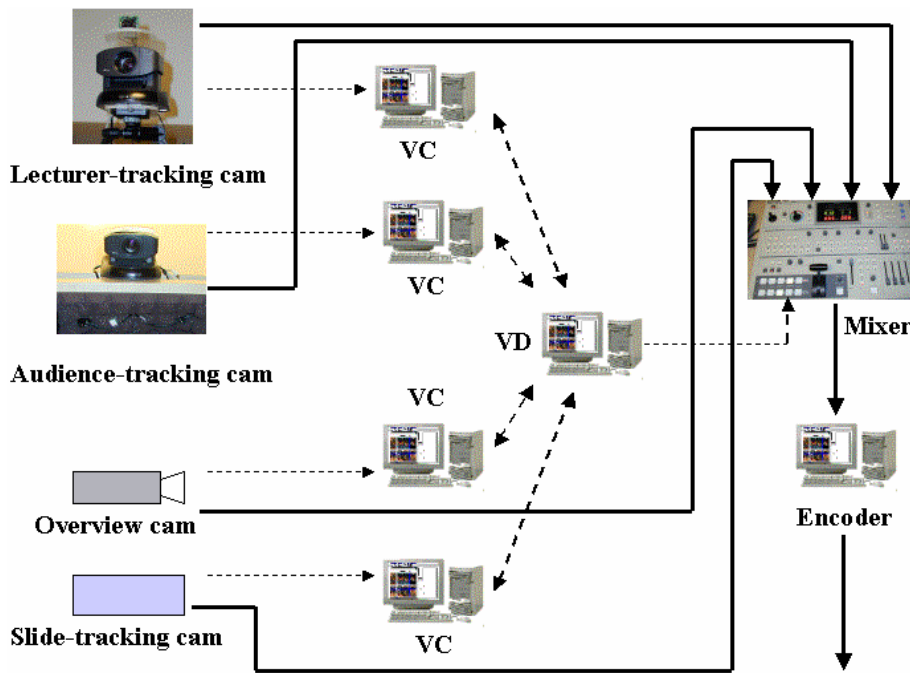
Fig. 1. *Dashed lines* indicate status and command signals. *Solid lines* indicate video data. VC stands for virtual cameramen and VD stands virtual director. It is worth noting that, even though we represent various VCs and VD with different computers, they can actually reside in a single computer running multiple threads

lects all the necessary information from the VCs, makes an informed decision as to which should be the final video output, and switches the video mixer to that camera. The edited lecture video is then encoded for both live broadcasting and on-demand viewing. For our first trial, we chose to use one lecturer-tracking VC, one audience-tracking VC, one slide-tracking VC, one overview VC, and one VD (Fig. 1). Note that, although the various VC/VDs are represented as different computers, they can actually reside in a single computer running different threads.

Figure 2 shows a top view of the lecture room where our system is physically installed. The lecturer normally moves behind the podium and in front of the screen. The audience area, about 50 seats, is to the right of the figure. Four cameras are devoted to lecturer tracking, audience tracking, a static overview, and slide tracking (e.g., a scan converter) to capture the screen display. The following is a list of the system AV hardware:

Fig. 2. Top view of the lecture room layout

- Two Sony EVI-D30 pan/tilt/zoom cameras for capturing lecturer and audience. The EVI camera pans between $[-100°, +100°]$, tilts between $[-25°, +25°]$, and a highest zoom level of $12\times$.
- A Super Circuit PC60XSA camera to monitor a lecturer's movement. It has a horizontal field of view (FOV) of $74°$.
- A Pelco Spectra II camera for the overview shot. We use this particular camera because it had been installed in the lecture room before our system was deployed. Nothing prevents the use of a low-end video camera, such as a PC60XSA.
- Two cheap Super Circuit PA3 omnidirectional microphones used in detecting which audience member is talking.
- A Panasonic WJ MX50 audio-video mixer. This low-end analog mixer takes four inputs and can be controlled by a computer via RS 232 link. We are currently working on a purely digital solution to render this MX50 mixer unnecessary.
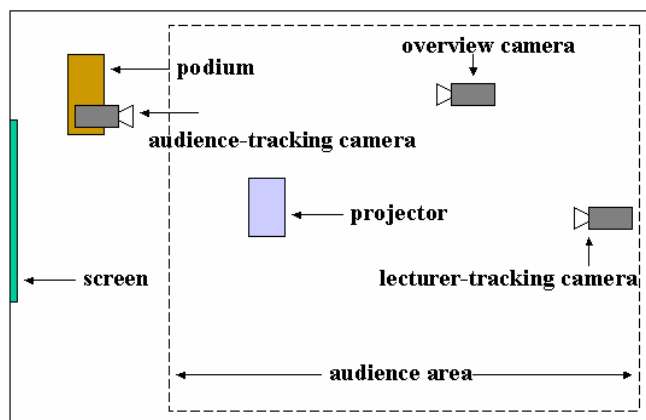
The user interface for the remote audience is shown in Fig. 3. To the left is a standard Windows MediaPlayer window. The output of lecturer-tracking, audience-tracking, and overview cameras are edited by the VD, and one is displayed in this window. The output of the slide-tracking camera is displayed to the right. An alternative would be to eliminate the latter window and integrate the output of the slide-tracking camera with the others. However, the Fig. 3 interface was already in use by our organization's lecture-capture team for lectures captured by a professional videographer. To obtain a controlled comparison, we use the same interface for our system. Note that a similar user interface was used in [16].

Because the overview VC constantly and statically views the whole lecture room, no tracking is needed. For the slide VC it is also relatively simple – it uses color histogram difference to determine if a new slide is shown. The lecturer-tracking and
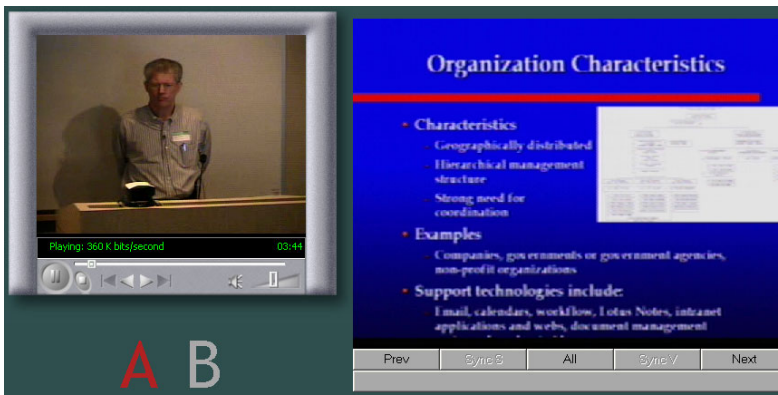
Fig. 3. User interface for rewmote audience

audience-tracking VC modules require much more complex sensor arrangements and framing rules. We next discuss these two modules as well as the critical VD module.

3.1 Lecturer-tracking VC

The lecturer-tracking VC must follow the lecturer's movement and gestures for a variety of shots: close-up to focus on expression, median shots for gestures, and long shots for context. As detailed in Sect. 2, various tracking techniques are available. We excluded obtrusive tracking techniques because of their unnecessary inconvenience. Of computer-vision and microphone-array techniques, the former is better suited for tracking the lecturer. In an unconstrained environment, reliable tracking of a target remains an open computer vision research problem. For example, some techniques can only track for a limited duration before the target begins to drift away; others require manual initialization of color, snakes, or blob [2]. While perfectly valid in their targeted applications, these approaches could not provide a fully automated system.

A lecture room environment imposes both challenges and opportunities. On one hand, a lecture room is usually dark and the lighting conditions change drastically when a lecturer switches from one slide to another. Most color-based and edge-based tracking cannot handle poor and variable lighting. On the other hand, we can exploit domain knowledge to make the tracking task manageable:

1. A lecturer is usually moving or gesturing during the lecture, so motion information can be an important tracking cue.
2. A lecturer's moving space is usually confined to the podium area, which allows a tracking algorithm to predefine a tracking region to help distinguish the lecturer's movement from that of the audience.

The first of these allows the use of simple frame-to-frame difference to conduct tracking for a real-time system. The second allows us to specify a podium area in the video frame so that a motion-based tracking algorithm is not distracted by audience movement. We mounted a static wide-angle camera on top of the lecturer-tracking camera and used the video frame difference from the wide-angle camera to guide the active camera to pan, tilt, and zoom (Fig. 4a). This tracking scheme does not require a lecturer to wear extra equipment, nor does it require human assistance.

A noticeable problem with our first prototype system [15] was that the lecturer-tracking camera moved too often – it continuously chased a moving lecturer. This could distract viewers. The current system uses the history of a lecturer's 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.

Let (x_t, y_t) be the location of the lecturer estimated from the wide-angle camera. Before the VD cuts to the lecturer-

Fig. 4. a Lecturer-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

a   **b**

tracking camera at time t, the lecturer-tracking VC will pan/tilt the camera such that it locks and focuses on location (x_t, y_t). To determine the zoom level of the camera, the lecturer-tracking VC maintains the trajectory of lecturer location in the past T seconds, $(X,Y) = \{(x_1, y_1), \ldots, (x_t, y_t), \ldots, (x_T, y_T)\}$. Currently, T is set to 10 s. The bounding box of the activity area in the past T seconds is then given by a rectangle (X_L, Y_T, X_R, Y_B), where they are the leftmost, topmost, rightmost, and bottom-most points in the set (X,Y). If we assume the lecturer's movement is piecewise stationary, we can use (X_L, Y_T, X_R, Y_B) as a good estimate of where the lecturer will be in the next T′ seconds. The zoom level Z_L is calculated as follows:

$$Z_L = \min\left(\frac{HFOV}{\angle(X_R, X_L)}, \frac{VFOV}{\angle(Y_B, Y_T)}\right), \qquad (1)$$

where *HFOV* and *VFOV* are the horizontal and vertical field of views of the Sony camera, and $\angle(,)$ represents the angle spanned by the two arguments in the Sony camera's coordinate system.

3.2 Audience-tracking VC

Showing audience members when they ask questions is important in making useful and interesting lecture videos. Because the audience area is usually quite dark and audience members may sit close to each other, computer-vision-based audience tracking will not work. A better sensing modality is based on microphone arrays, where the audience-tracking VC estimates the sound source using the microphones and uses this to control a camera.

As mentioned in Sect. 2, there are commercial products that implement SSL steered tracking cameras (e.g., Picture-Tel [18] and PolyCom [19]). However, they do not expose their APIs and do not satisfy our framing strategies. For example, their response time is not quick enough and they do not accept commands such as "pan slowly from left to right". To have full control of the audience-tracking VC module, we developed our own SSL techniques. Within various SSL approaches, the generalized cross-correlation (GCC) approach receives the most research attention and is the most successful [5,28]. Let $s(n)$ be the source signal and $x_1(n)$ and $x_2(n)$ be the signals received by the two microphones:

$$\begin{aligned} x_1(n) &= as(n-D) + h_1(n) * s(n) + n_1(n) \\ x_2(n) &= bs(n) + h_2(n) * s(n) + n_2(n) \end{aligned}, \qquad (2)$$

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

$$\begin{aligned} D &= \arg\max_{\tau} \hat{R}_{x_1x_2}(\tau) \\ \hat{R}_{x_1x_2}(\tau) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} W(\omega) G_{x_1x_2}(\omega) e^{j\omega\tau} d\omega, \end{aligned} \qquad (3)$$

where $\hat{R}_{x_1x_2}(\tau)$ is the cross correlation of $x_1(n)$ and $x_2(n)$, $G_{x_1x_2}(\omega)$ is the Fourier transform of $\hat{R}_{x_1x_2}(\tau)$, i.e.,

the cross power spectrum, and $W(w)$ is the weighting function.

In practice, choosing the right weighting function is of great significance in achieving accurate and robust time delay estimation. As seen in Eq. 2, there are two types of noise in the system: background noise $n_1(n)$ and $n_2(n)$ and reverberations $h_1(n)$ and $h_2(n)$. Previous research suggests that the maximum likelihood (ML) weighting function is robust to background noise and the phase transformation (PHAT) weighting function is better at dealing with reverberations [28]:

$$W_{\mathrm{ML}}(\omega) = \frac{1}{|N(\omega)|^2}$$

$$W_{\mathrm{PHAT}}(\omega) = \frac{1}{|G_{x_1x_2}(\omega)|},$$

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

These weighting functions are at two extremes: $W_{\mathrm{ML}}(w)$ puts too much emphasis on "noiseless" frequencies, whereas $W_{\mathrm{PHAT}}(w)$ treats all frequencies equally. We developed a new weighting function that simultaneously deals with background noise and reverberations [28]:

$$W_{MLR}(\omega) = \frac{|X_1(\omega)||X_2(\omega)|}{2q|X_1(\omega)|^2|X_2(\omega)|^2+(1-q)|N_2(\omega)|^2|X_1(\omega)|^2+|N_1(\omega)|^2|X_2(\omega)|^2},$$

where $q \in [0,1]$ is the proportion factor and $X_i(\omega)$, $i = 1, 2$, is the Fourier transfer of $x_i(n)$. Experimentally, we found $q = 0.3$ is a good value for typical lecture rooms.

Once the time delay D is estimated by the above procedure, the sound source direction can be estimated given the microphone array geometry. As shown in Fig. 5, let the two microphones be at locations A and B, where AB is called the baseline of the microphone array. Let the active camera be at location O, whose optical axis is perpendicular to AB. The goal of SSL is to estimate the angle $\angle COX$ such that the active camera can point in the right direction. When the distance of the target, i.e., |OC|, is much larger than the length of the baseline |AB|, the angle $\angle COX$ can be estimated as follows [5]:

$$\angle COX \approx \angle BAD = \arcsin\frac{|BD|}{|AB|} = \arcsin\frac{D \times v}{|AB|}, \qquad (4)$$

where D is the time delay and $v = 342$ m/s is the speed of sound traveling in air.

To estimate the panning angles of the active camera, we need at least two microphones in a configuration similar to that in Fig. 5. If we want to estimate the tilting angle as well,
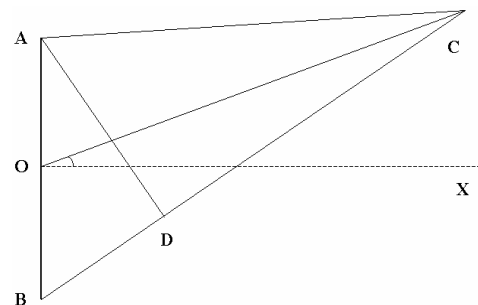

Fig. 5. Sound source localization

we need a third microphone. By having four microphones in a planar grid, we can estimate the distance of the sound source in addition to the pan/tilt angles [5]. Of course, adding more microphones increases the system complexity. In our particular application, however, simpler solutions are available. Because audience members are typically sitting on their seats, if the active camera is mounted slighted above eye level, tilting is not necessary. Furthermore, because estimating sound source distance is still less robust than estimating sound source direction, in our current system we focus our attention only on how to accurately control the panning angles of the active camera. In our hardware configuration (Fig. 4b), two microphones (PA3) are put below the audience-tracking camera, while the horizontal centers of the microphone array and the camera are aligned.

3.3 VC-VD communication protocol

Our system's two-level virtual camera-virtual director structure simulates a professional video production crew. This arrangement also allows for clean separation between policy and mechanism. The VCs handle the low-level chores of controlling the camera (e.g., tracking the lecturer or the audience member raising a question) and periodically report their status to the VD. The VD module, which encodes the high-level policies, then makes an informed decision on which VC's camera is chosen to broadcast. The VC-VD communication protocol is therefore of crucial importance to the success of the system.

Our first prototype [15] supported limited communication. For example, the VD only informed a VC if its camera was being selected as the output camera, and the VCs only reported to the VD if they were ready or not ready. Sophisticated rules, such as audience panning and slide changing, were not 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; for example, what is the *zoom level* of the camera?
- **Confidence**: How *confident* is a VC in a decision, for example, 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 is selected as *preview*, *on air*, or *off air*.

This communication protocol allows the VD and VCs to exchange information effectively in support of more sophisticated video production rules. For example, we can provide a slow pan of the audience, and the duration of focus on a questioner is a function of our confidence in the SSL.

3.4 Virtual director

The responsibility of the VD module is to gather and analyze reports from different VCs, to make intelligent decisions on which camera to select, and to control the video mixer to generate the final video output. Just like video directors in real life, a good VD module observes the rules of cinematography and video editing to make the recording more informative and entertaining. Here we focus on how a flexible VD module can easily encode various editing rules. We equipped the VD with two tools: an event generator to trigger switching from one camera to another and a finite state machine (FSM) to decide which camera to switch to.

3.4.1 Event generator – when to switch

The event generator generates two types of events that cause the VD to switch cameras: *STATUS_CHANGE* and *TIME_EXPIRE*.

STATUS_CHANGE events

When there is a scene change, such as an audience member speaking, or an action change, such as a camera changing from doing to done, the event generator generates a *STATUS_CHANGE* event. The VD then takes actions to handle this event (e.g., switches to a different camera).

TIME_EXPIRE events

In video production, switching from one camera to another is called a *cut*. The period between two cuts is called a video *shot*. An important video editing rule is that a shot should not be too long or too short. To enact this rule, each camera has a minimum shot duration D_{MIN} and a maximum allowable duration D_{MAX}. If a shot length is less than D_{MIN}, no camera switch is made. On the other hand, if a camera has been on longer than its D_{MAX}, a *TIME_EXPIRE* event is generated and sent to the VD. Currently, D_{MIN} is set to 5 s for all cameras, based on the professionals' suggestions.

Two factors affect a shot's length D_{MAX}: the nature of the shot and the quality of the shot. The nature of a shot determines a base duration D_{BASE} for each camera. For example, lecturer-tracking shots are longer than overview shots because they are in general more interesting. The quality of a shot is defined as a weighted combination of the camera zoom level Z_L and tracking confidence level C_L. The quality of the shot affects the value of D_{MAX} in that high-quality shots should be allowed to last longer than low-quality shots. The final D_{MAX} is therefore a product of the base length D_{BASE} and the shot quality:

$$D_{\mathrm{MAX}} = D_{\mathrm{BASE}} \times (\alpha Z_L + (1-\alpha)C_L),$$

where α is chosen experimentally. We use $\alpha = 0.4$ in our current system.

3.4.2 FSM – where to switch

In Sect. 3.4.1, we discussed how event generation triggers the VD to switch cameras. In this section we discuss which camera the VD switches to upon receiving a triggering event. Because the slide-tracking camera's video appears in a separate window (Fig. 3), only the lecturer-tracking, audience-tracking, and overview cameras are dispatched by the VD.

In [11], He et al. propose a hierarchical FSM structure to simulate a virtual cinematographer in a virtual graphics environment. This influenced our design of the VC and VD modules. Unlike their system, our system works in the real world, which imposes physical constraints on how we can manipulate cameras and people. For example, we cannot obtain a shot from an arbitrary angle. Furthermore, although their system can assume all the cameras are available at all times in the virtual environment, our system cannot, because targets may not be in the field of view of some cameras. This leads to greater complexity in our VD module.

To model different camera functionalities, each camera can have one or multiple states. In our case, the lecturer-tracking camera has one state: *Lecturer_Focus*; the audience-tracking camera has three: *Audience_Focus*, *Audience_Pan*, and *Audience_Static*; and the overview camera has one: *Overview_Static*. Figure 6 shows this five-state FSM. When the system enters a state, the camera associated with that state becomes the active camera. At design stage, the designer specifies the states associated with a camera and sets of events that cause the state to transition to other states.

Professional video editing rules can easily be encoded in this framework. For example, a cut is more often made from the lecturer-tracking camera to the overview camera than to the audience-tracking camera. To encode this rule, we make the transition probability of the former higher than that of the latter. The following pseudocode illustrates how the system transits from *Lecturer_Focus* to other states:

```
if (CurrentState == Lecturer_Focus) {
  if (the shot is not very good any more ||
      the shot has been on for too long) {
          GotoOtherStatesWithProbabilities(
          Audience_Static, 0.2,
          Audience_Pan, 0.1,
          Overview_Static, 0.7);
  }
}
```
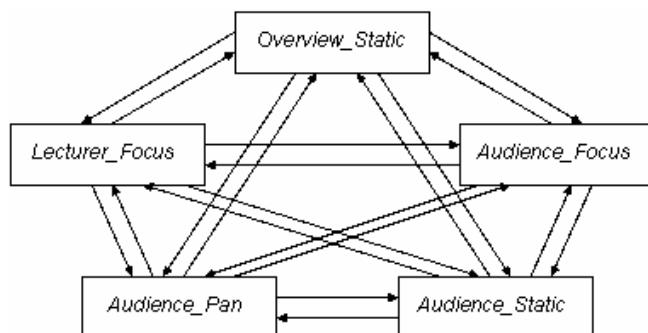

Fig. 6. A five-state FSM

Note that when the system changes state, the transition probabilities guide the transitions. In the case just described, the system goes to *Audience_Static* with probability 0.2, *Audience_Pan* with probability 0.1, and *Overview_Static* with probability 0.7. This provides VD designers the flexibility to tailor the FSM to their needs. At the microscopic level, each camera transition is random, resulting in less predictability, which can make viewing more interesting. At the macroscopic level, some transitions are more likely to happen than others, following the video editing rules. Experimental results in the next section reveal that such an FSM strategy performs well in simulating a human director.

4 Study and professional critique of automated system

Our system has been in use for 2 years. It employs a basic set of camera and transition management rules based on our reading of the literature and discussions with a videographer in our organization, and was enhanced following a study reported in [15]. To identify weaknesses and possible enhancements, we devised a study that involved four professional videographers from outside our organization as well as the natural audience for a series of talks being given in the lecture room in which the system is deployed (Fig. 2).

The room is used on a daily basis for lectures that are viewed by a local audience while our system automatically broadcasts them throughout the company and digitally records them for on-demand viewing. To compare our system against human videographers, we restructured the lecture room so that both a videographer and the 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 1-h lectures on collaboration technologies given by two HCI researchers was used in the study.

Thus, there were two groups of participants in this study. Four professional videographers, each with 3 to 12 years' experience, were recruited from a professional video production company. Each recorded one of the four lectures. After a recording, we interviewed the videographer for 2 h. We asked them what they had done during the lecture 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 were not told about our system in advance; all were interested in and did not appear threatened by such a system.) They then filled out and discussed answers to a survey covering system quality. Finally, we asked them how they would position and operate cameras in different kinds of rooms and with different levels of equipment.

Employees who decided on their own initiative to watch a lecture from their offices were asked if they were willing to participate in an experimental evaluation. Eighteen agreed. The interface they saw is shown in Fig. 3. The left portion is a standard Microsoft MediaPlayer window. The outputs of the lecturer-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 simulta-

neously by a videographer and by our system. Remote viewers were told that two videographers, designated A and B (see bottom-left portion of Fig. 3), would alternate every 10 min and were asked to pay attention and rate the two following the lecture. A and B are randomly assigned to the videographer and our system for each lecture. Following the lecture, they filled out the survey discussed below.

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. We use a scale of 1–5, where 1 is strongly disagree, 2 disagree, 3 neutral, 4 agree, and 5 strongly agree. Because the answers are in ranking order, i.e., 1–5, the WilCoxon test is used to compare different testing conditions. The p-value in the table indicates the probability that the comparison results are due to random variation. The standard in psychology is that if p is less than 0.05, then the difference is considered significant. The first seven questions in the table relate to individual aspects of lecture recording, and the last three questions focus on the overall lecture-watching experience.

Individual aspects

The professionals rated our system quite well for Q4, Q5, and Q7 (median ratings of 3.5 to 4.0; all ratings are medians unless indicated otherwise; 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 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 the 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 the 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, and they made comments like why was the corner of the screen showing in the lecturer shot (Fig. 4b). 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 with 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 evaluator to rank the statement "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 the evaluator to rank the statement "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,

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. The 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-value
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)	>0.05
9. Overall quality	2.0 (2.0)	3.0 (2.8)	4.0 (3.8)	<0.01
10. As compared with previous experience	1.5 (1.5)	3.0 (3.1)	3.0 (3.6)	0.11

smooth tracking of lecturer, and semantically motivated shot cuts were the primary reasons given. The remote audience also clearly preferred the overall quality of video from the professionals ($p < 0.01$), 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 the professionals was equivalent to their previous experiences, giving scores of 3.0.

It is interesting to note that, although 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 a 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 Detailed rules and technology feasibility

Most existing systems are not based on systematic study of video production rules or the corresponding technical feasibility. The high-level rules employed in our previous effort proved insufficiently comprehensive [15]. In this section we consider detailed rules for video production based on interviews with professional videographers (represented as *A*, *B*, *C*, and *D*). We also analyze automation feasibility employing current state-of-the-art technologies.

6.1 Camera positioning rules

The professionals generally favored positioning cameras about 2 m 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. 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 Fig. 5a). 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 did not. *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 [1]. 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 the lecturer properly, failing to provide the proper 10 to 15 cm of space above the head and sometimes losing the lecturer entirely (Fig. 7). 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 butterlike 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 over slowly 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. Our lecturer framing relies partly on this strategy.

Automation feasibility

Although base-level lecturer tracking and framing rules are achievable, as with 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,
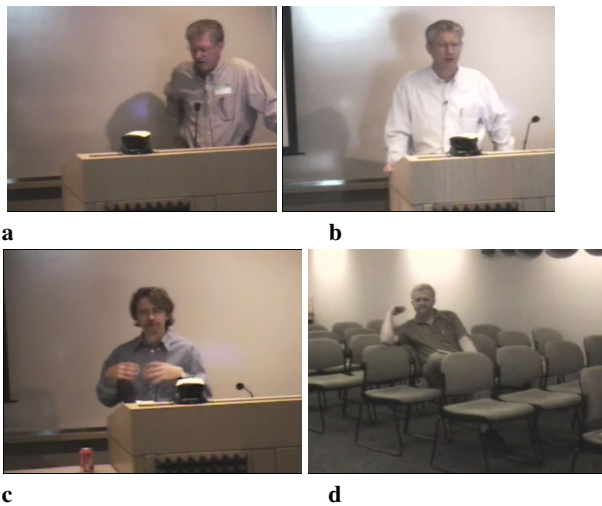
Fig. 7a–d Examples of bad framing. **a** Not centered. **b** Inclusion of screen edge. **c** Too much headroom. **d** Showing an almost empty audience shot

for eye gaze detection, an effective technique is the two IR light sources used in the IBM BlueEye project [29]. 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 the 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, 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, use a wide audience shot or remain with lecturer.*

Rule 3.2. *Do not show relatively empty audience shots.* (See Fig. 4d 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 had been 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 the previous section, 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.

Automation feasibility

Our sophisticated SSL technique 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 there are 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 commented, "… keep the shots mixed up so (viewers) can't totally predict …." All the 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., $3 \sim 5$ s, 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 s for overview shots. For audience shots the durations mentioned are in the range of 4–10 s 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 person's face).

Rule 4.4. *Shot transitions should be motivated.*

Rule 4.5. *A good time for a transition 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 noted above, opinions differed 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 can be used not only as a safety backup but also 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 that they 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. 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 an overview shot. It would be nice if we could detect if a lecturer is pointing to the screen, which is a good time to make motivated transitions. As for detecting if a lecturer is finishing his/her thoughts, that is an extremely difficult problem. It requires high-accuracy speech recognition in a noisy environment and real-time natural language understanding, both requiring years of research. However, we could provide a partial solution – we can detect if the lecturer is talking. This way, at least we will not make a transition when the lecturer is still talking.
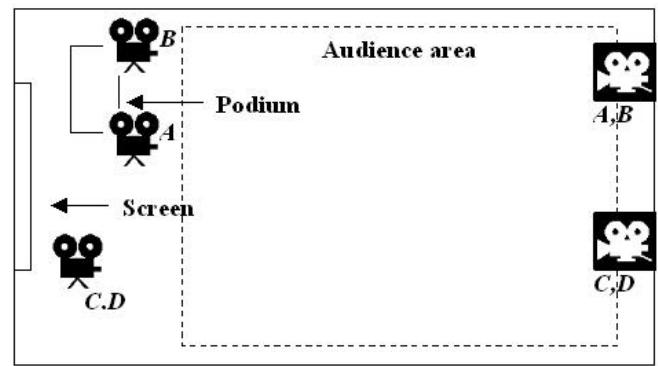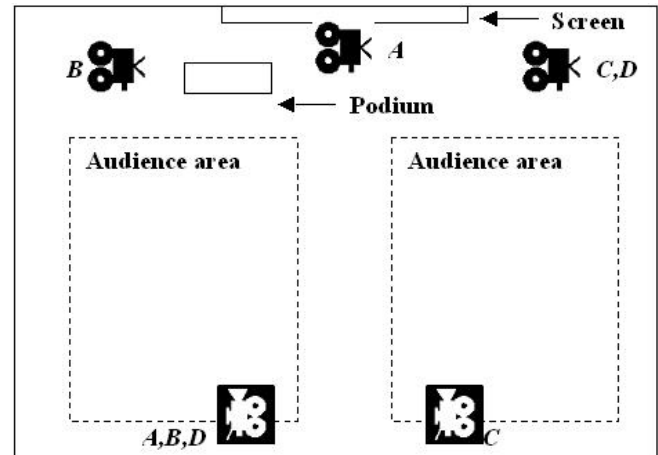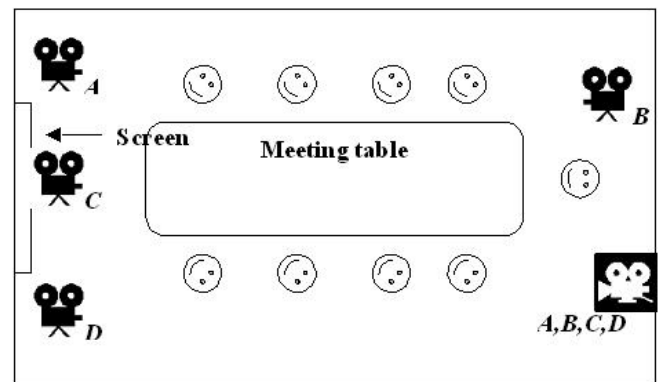
Fig. 8. a Medium-sized lecture room camera position (R1C3). **b** Large auditorium camera position (R2C3). **c** Meeting room camera position (R3C3). The three room configurations. White cameras are lecturer-tracking/overview units, black cameras are audience-tracking units. *Letters* indicate the different videographers' choices. Slide cameras are implicit – they just capture the screens

7 Generalization to different settings

Our discussion so far has focused on a medium-sized lecture room with multiple cameras. We would like to accommodate different lecture venues and different levels of technology investment. We asked the videographers how the rules and camera setup would change in different environments. We asked them to consider three common venue types: (R1)

medium-sized lecture room ($\sim$ 50 people), (R2) large auditorium ($\sim$ 100+ people), and (R3) small meeting room ($\sim$ 10–20 people). For the small meeting room, we are interested in the presentation scenario than the discussion scenario. The arrangements are shown in Fig. 5. We asked them to also consider three levels of technology investment: (C1) A single dual-function lecturer-tracking plus overview camera (such as our lecturer-tracking camera with its wide-angle camera on top); (C2) two cameras – C1 plus a slide/screen-capturing camera; and (C3) three cameras – C2 plus an audience-tracking camera. This leads to nine combinations (R1–R3 $\times$ C1–C3). For simplicity, we use RnCm to represent the case where camera configuration Cn is used in the lecture room Rm.

7.1 Camera positioning

Figure 8 shows camera positions proposed by videographers **A**, **B**, **C**, and **D**. When the audience camera or slide camera was not present, the videographers did not suggest changing the position of the lecturer-tracking/overview camera, so cases R1C3, R2C3, and R3C3 cover all nine combinations.

The layout in Fig. 5a (R1C3) represents the lecture room where our system is installed. The videographers' assessment of it was described in the previous section – for instance, the differing preferences for face-on and angled views of a lecturer.

For the auditorium (5b: 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 (5c: R3C3), the audience faces different directions and the cameras are closer to the people, 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. This variability may reflect that this is a less common professional setting; it also suggests greater flexibility for automated systems.

7.2 Shots and transitions

We discussed the shots and transitions for configuration R1C3. Based on our interviews with the professionals, most rules for R1C3 generalize to R2C3 and R3C3. A major exception is the meeting room (R3C3), where the audience-tracking camera often can only see half of the audience. The videographers suggested two possible actions when a person in such a blind zone asks a question: Avoid transitioning to an audience shot, or use the lecturer-tracking camera to cover the shot if it can, using the overview camera as the transition. In the latter case, the sequence would subsequently be reversed: audience to overview to lecturer. Recall that the lecturer-tracking/overview camera is a dual-function unit – the top static camera providing overview shots while the bottom camera is pan/tilt/zoom controllable.

For all three rooms, the rules for case C2 were similar to those in C3. However, with the audience camera unavailable in C2, there were a few rule suggestions for audience shots. One was to simply ignore audience shots. The other was to use the lecturer-tracking camera to cover the audience when possible, with the following shot transitions: lecturer to overview to audience to overview to lecturer.

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

- 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 also to capture slides at slide transitions. Switch to the overview camera when the lecturer-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 will not affect the positioning of existing cameras. Second, for all three rooms, downgrading the equipment investment from C3 to C2 or C1 requires 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 well-studied lecture room (R1). These findings should enable other practitioners to construct systems for their environments.

8 Concluding remarks

We described the technology and features of a lecture room automation system that is in daily use and its assessment by viewers and professional videographers. To enable researchers and practitioners to build on the results, we presented detailed video production rules and analyzed their automation feasibility. Advanced rules will require considerable further research, but basic rules that can be realized today may suffice to cover lectures when professional camera operation and editing are unavailable. Requests to use our system are evidence of a strong perceived need.

We also apply rules for different room and camera configurations, finding that the changes are few and well defined. However, the fact that professional videographers do differ to some extent in applying rules indicates that there is flexibility, which is grounds for optimism. Successful lecture room automation could have a major impact on how people attend and learn from lectures. The hardware cost for such systems is already reasonable and is dropping. By eliminating the need to hire human videographers in some cases, more presentations can be made accessible online in a range of settings.

References

1. Arijon D (1976) Grammar of the film language. Hastings House, New York
2. Baumberg A, Hogg D (1994) An efficient method for contour tracking using active shape models, TR 94.11. University of Leeds
3. Benesty J (2000) Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. J Acoust Am 107:384–391
4. Bianchi M (1998) AutoAuditorium: a fully automatic, multi-camera system to televise auditorium presentations. In: Proceedings of the Joint DARPA/NIST workshop on smart spaces technology
5. Brandstein M, Silverman H (1997) A practical methodology for speech source localization with microphone arrays. Comput Speech Lang 11(2):91–126
6. Brotherton J, Abowd G (1998) Rooms take note: room takes notes. In: Proceedings of the AAAI symposium on intelligent environments, pp 23–30
7. Cruz G, Hill R (1994) Capturing and playing multimedia events with STREAMS. In: Proceedings of ACM Multimedia'94, pp 193–200
8. Cutler R, Turk M (1998) View-based interpretation of real-time optical flow for gesture recognition. In: Proceedings of the IEEE conference on automatic face and gesture recognition (FG'98), pp 416–421
9. Finn K, Sellen A, Wilbur S (eds) (1997) Video-mediated communication. Erlbaum, Mahwah, NJ
10. Gleicher M, Masanz J (2000) Towards virtual videography. In: Proceedings of ACM Multimedia'00, Los Angeles
11. He L, Cohen M, Salesin D (1996) The virtual cinematographer: a paradigm for automatic real-time camera control and directing. In: Proceedings of ACM SIGGRAPH'96, New Orleans
12. He L, Grudin J, Gupta A (2000) Designing presentations for on-demand viewing. In: Proceedings of CSCW'00
13. Kleban J (2000) Combined acoustic and visual processing for video conferencing systems. MS thesis, The State University of New Jersey, Rutgers
14. Liu Q, Kimber D, Foote J, Wilcox L, Boreczky J (2002) FLY-SEPC: a multi-user video camera system with hybrid human and automatic control. In: Proceedings of ACM Multimedia 2002, Juan-les-Pins, France, pp 484–492
15. Liu Q, Rui Y, Gupta A, Cadiz JJ (2001) Automating camera management in lecture room environments. In: Proceedings of ACM CHI 2001, Seattle
16. Mukhopadhyay S, Smith B (1999) Passive capture and structuring of lectures. In: Proceedings of ACM Multimedia'99
17. ParkerVision. http://www.parkervision.com/
18. PictureTel. http://www.picturetel.com/
19. PolyCom. http://www.polycom.com/
20. Professional Master's Program, University of Washington. http://www.cs.washington.edu/education/dl/course_index.html
21. Rui Y, He L, Gupta A, Liu Q (2001) Building an intelligent camera management system. In: Proceedings of ACM Multimedia, Ottawa, Canada, pp 2–11
22. Rui Y, Gupta A, Grudin J (2003) Videography for telepresentation. In: Proceedings of ACM CHI 2003, Ft Lauderdale, FL, pp 457–464
23. Song D, Goldberg K (2003) ShareCam Part I: Interface, system architecture, and implementation of a collaboratively controlled robotic webcam. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems, Las Vegas, October 2003
24. Song D, Goldberg K, Pashkevich A (2003) ShareCam Part II: Approximate and distributed algorithms for a collaboratively controlled robotic webcam. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems, Las Vegas, October 2003
25. Stanford Online. http://scpd.stanford.edu/scpd/students/online-class.htm
26. Stiefelhagen R, Yang J, Waibel A (1999) Modeling focus of attention for meeting indexing. In: Proceedings of ACM Multiemdia'99
27. Wang C, Brandstein M (1998) A hybrid real-time face tracking system. In: Proceedings of ICASSP98, Seattle, May 1998, pp 3737–3740
28. Wang H, Chu P (1997) Voice source localization for automatic camera pointing system in video conferencing. In: Proceedings of ICASSP'97
29. Zhai S, Morimoto C, Ihde S (1999) Manual and gaze input cascaded (MAGIC) pointing. In: Proceedings of CHI'99, pp 246–253