Improving Meeting Capture by Applying Television Production Principles with Audio and Motion Detection

Abhishek Ranjan[1], Jeremy Birnholtz[1,2], Ravin Balakrishnan[1]

[1]Department of Computer Science
University of Toronto
www.dgp.toronto.edu
aranjan | ravin @dgp.toronto.edu

[2]Department of Communication
Cornell University
www.comm.cornell.edu
jpb277 @cornell.edu

ABSTRACT

Video recordings of meetings are often monotonous and tedious to watch. In this paper, we report on the design, implementation and evaluation of an automated meeting capture system that applies television production principles to capture and present videos of small group meetings in a compelling manner. The system uses inputs from a motion capture system and microphones to drive multiple pan-tilt-zoom cameras and uses heuristics to frame shots and cut between them. An evaluation of the system indicates that its performance approaches that of a professional crew while requiring significantly fewer human resources.

Author Keywords

Meeting capture, automated camera control, video

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Meetings are a frequent and necessary aspect of life in most organizations, with a 1999 white paper reporting that 37% of employee time in the United States is spent attending meetings, and that there are over 11 million business meetings held daily [18]. Moreover, in a recent survey, 50% of the respondents indicated that attending face-to-face meetings was a waste of their time [17]. These statistics highlight two problems. First, the amount of time that some employees spend in meetings makes it difficult for them to either attend all of them or get anything else done [17]. Second, it can be difficult to recall what was said or accomplished in each one.

These problems have sparked significant recent interest in the ability to capture meetings for later review by those who missed the meeting or for archival reference purposes (e.g. Classroom2000 [1], Quindi [26], MeetingSense [19], Indico [9], WLAP [37]). Some of these capture systems have been reported to be successful [25], indicating potential value in capturing and archiving meetings.

CHI 2008, April 5–10, 2008, Florence, Italy.

Such meeting capture systems typically use some combination of video and audio recordings, combined with presentation media (e.g., PowerPoint), and agendas. One problem with the recording component of these systems, however, is that the video often does not capture the most relevant aspects of the meeting. This makes the reviewing and retrieval task cumbersome and is often the reason video archives are not used, rather than people not wanting meetings to be captured in the first place [12]. In particular, many systems use only one camera to capture video, and, unless dedicated staff are on hand to manage the system, the shot from this camera does not change often [1, 23]. Without multiple views, users may lack the visual information required to understand the context [7]. Further, a relatively static camera typically results in a video that is boring to watch [11, 15]. Having dedicated video production staff does improve the video, but this comes at a significant cost that would likely be prohibitive for most meetings. This is illustrated by Figure 1 which shows the setup required for a professional television crew to record a modest meeting of three people.

Figure 1. An example television production crew setup

Accordingly, there has been significant recent interest in automating camera control to make videos that look more like those shot by professionals. This is accomplished by understanding what the camera should be aimed at [21, 24, 31], how it should be moved [3, 14] and when to cut between cameras [11, 15, 28]. Professional crews notice

and respond to cues such as who is speaking, who is likely to speak next, gestures and other body language, and a set of heuristics about when to cut between shots [2, 38].

This paper describes our iterative development and evaluation of an automated camera control system that leverages television production principles and uses input from a motion tracking system and several microphones.

BACKGROUND

There are essentially three criteria that an effective meeting video must meet.

1. *It must capture enough visual information to allow viewers to understand what took place.* Capturing the desired visual information can be challenging in that meetings may involve rapid dialogs, physical artifacts, presentation media, whiteboards, etc. [22, 30]. This requires either a single camera shot that can include everything [23], or the capacity for multiple shots via a movable camera or multiple cameras [6, 11, 14].
2. *It must be compelling to watch.* People's expectations for, and ability to engage with, video recordings they view are shaped by their prior experience in viewing video recordings [27]. The problem with the fixed wide-shots used by many existing capture systems is that the video itself (apart from content) is monotonous when compared with professionally produced video [11]. In this regard, it could be useful to understand the techniques [2, 38] that make television more compelling for viewers.
3. *It must not require substantial human effort.* Meeting participants are primarily there to attend a meeting, and typically do not operate cameras reliably when user controls are provided [6, 24]. Similarly, professional crews are only affordable for some events [3, 29].

If we assume that the third criterion requires an automated solution for everyday use, our problem then becomes one of automatically creating a video that captures necessary visual information and is compelling to watch. Capturing and recording a meeting is fundamentally comprised of three tasks, executed repeatedly: 1) determining what is or is likely to soon be the most important piece of visual information in the setting (e.g., the face of the person talking), 2) getting an appropriately framed shot of that bit of information, and 3) cutting to that shot. We now turn to the problems and prior work in achieving these goals.

Finding the most important thing

The first task in a complex environment is to determine what the viewer will want to see. In a meeting setting, this is typically the person who is talking, and prior efforts reflect this. Several systems [11, 15, 28], for example, use speaker detection algorithms to determine who is talking and select a camera known to have a shot of that person.

While effective in determining the speaker, this approach can lack the variety of shots that provide viewers with contextual information about other attendees. To address this issue, Inoue et al. [11] augmented a speaker detection system and cut between multiple camera views using an algorithm based on shot content and transition probabilities gleaned from professionally produced television shows. This approach adds shot variety, but their implementation, like the system cited above, does not account for human movement in transitioning between shots.

Human television crews are able to overcome these issues because they are able to see and anticipate people's movements [5, 38]. The ability to make these predictions comes partly from experience, but also from the ability to recognize subtle cues (e.g., gaze, gestures) that people are getting ready to talk or move. Reflecting this approach, Takemae et al. [32] used gaze direction as a cue in editing video recordings of conversation. They proposed that in a meeting, the focus of attention can be predicted by finding the participant who is being gazed at by the maximum number of participants.

Getting the shot

After determining what the viewer is likely to want to see, the next step is ensuring that a shot is available. This involves locating the object in space, determining which camera is best suited to get a shot of it, and framing that shot properly.

While locating the object is typically easy for human directors for reasons discussed above, it is difficult or impossible for systems without some sort of motion tracking component. Previous systems [3, 15] coarsely tracked a single individual, such as a speaker at the front of an auditorium, using vision techniques, but most systems to date have not leveraged the potential of "seeing" objects or people in the 3D space.

Once objects can be located precisely, determining the camera to get the shot can be simplified by employing camera placement heuristics used in television studios. In a typical 3-camera studio (Figure 3), on which our prototype system is modeled, one camera is placed in the center and the other two are placed to the sides. Each of the side cameras is then responsible for shots of the participants opposite them, and the center camera typically provides wide shots as well [2, 38]. Depending on which camera is "live" at any given moment, there may be some variation in how cameras are actually used to get required shots.

Finally, framing the shot also requires the ability to locate objects in space. Assuming this capability is present, television production heuristics can again assist with this process. In particular, the notion of "headroom" suggests that some space be left above people's heads in framing close-up shots. And the notion of "noseroom" and "leadroom" suggests that, when people are not looking or moving directly toward the camera, some extra space be left

on the side of the screen toward which they are looking or walking. This serves to both make the shot look more pleasing, and to anticipate future movement by allowing room for it to occur [38]. While Liu et al. [15] noted the importance of these principles, they could not implement them due to inadequate technology.

Cutting to the shot

The final step in the process is cutting to the shot. While this may seem obvious, this step is actually subtle and nuanced. Television directors are trained to avoid certain types of cuts (e.g., "jump cuts" where a person appears to "jump" on the screen) and to pay attention to visual signals, such as gaze or physical movements (e.g., cutting from a close-up to a wide shot while somebody stands up rather than after the head has already left the shot [5]).

Liu et al. [15] draw on these heuristics to automate camera control in an auditorium setting where only a single speaker is typically of interest. However, our setting is that of small meetings which are inherently dynamic and complex, with several participants of interest. Inoue et al. [11] tackle a similar setting using probabilistic shot transitions. However, their system was limited to organized meetings where people strictly took turns to talk one by one [10].

OUR ITERATIVE SYSTEM DESIGN PROCESS

In this section, we describe the design process we used in developing our prototype system. Throughout this process, we worked from the principles described above and sought guidance from two people with professional television directing training and experience. One of them currently is a professor in the television arts program at a local university and has over 30 years of experience in the television industry as a director and camera operator. The other is a member of our research team, who spent 8 years training and working in the television industry.

Initial Prototype Design

Meeting Room Layout and Camera Positioning

In our prototype system, we considered a small informal meeting scenario with three collocated participants. Such a meeting is common in many settings and provides us with a basis for design that is realistic, but not so complicated as to render prototyping and testing intractable.

The seating arrangement and the room layout are shown in Figure 3. The camera placement was based on typical studio designs [2, 38] and suggestions of our two expert directors. The meeting room had a rectangular table in the center, and the three participants were seated around it. Since whiteboards are often used as a medium to present ideas in small group meetings [22], we placed one near a corner of the table, visible to participants and the cameras.

Equipment

We used three Sony SNC-RZ30 PTZ cameras (640x480 pixels resolution, IP enabled) to capture video and three Shure SLX wireless clip-on lavaliere microphones to capture audio. The wireless microphone system allowed participants to move in the meeting space without losing the audio input.

To allow the system to locate people in the meeting space, we tracked participants' location and motion using a Vicon motion tracking system [36]. Each participant wore a head-band with passive markers. These markers were visible to an array of infrared cameras in our lab space and allowed us to track participant head position and orientation in real-time. While these headbands with markers were required for our prototype system, it is expected that, as computer vision technologies improve, there will eventually be no need for physical markers [20].

Tracked Events

In order to use as many cues as possible to determine the most important visual information, the system tracked the following events using the microphones and the motion tracking system:

1. Speaker change: Each microphone was constantly polled to read audio signals from each participant, and change in sound energy level was used to differentiate speech from silence.
2. Posture change (sitting, standing, or moving): The height of the participant's head was calibrated to differentiate between sitting and standing positions, and head movement range was calibrated to detect if the participant was moving.
3. Head orientation: Head orientation has been shown to be a good approximation for gaze [33]. We tracked head orientation in 3D space by applying methods used by Birnholtz et al. [4].

Figure 2. Close-up shot (left) and overview shot (right) used in the initial prototype

Shot Transition: When to cut

The system used the aforementioned cues to determine what the viewer might want to see and frame a shot of it. In particular, whenever there was a speaker change detected, the system showed a close-up shot of the new speaker. When multiple speakers started to speak at the same time or took turns quickly, the system cut to an overview or wide shot that showed all three participants (see Figure 2). Furthermore, whenever a participant's posture changed from sitting to standing or walking, the system showed the overview shot to convey the posture change to viewers.

One consequence of these shot transition rules was that, since people in meetings frequently speak at the same time or in rapid succession, the system cut to the overview shot more often than we would have liked. This issue is further addressed below.

In television production there is a notion of screen duration which refers to the duration for which a shot stays "on the air". In order to avoid extremely short or extremely long shots, screen duration often has a lower and upper limit. Rui et al. also used this notion in their system [29]. In our system, every shot had a minimum length of 3 seconds and a maximum length of 15 seconds. These bounds were decided after consulting our two expert directors and performing iterative adjustments.

Getting the Shot

If, based on the shot transition rules described above, the needed shot was not immediately available, the system then had to allocate a camera for this task.

Even though there were three cameras available, this was sometimes nontrivial as one of the cameras was always "on the air" and could not be moved quickly (as that would be jarring to the viewer). Thus, at any given moment we actually only have two available cameras for getting a new shot. Given the amount of interpersonal interaction taking place, this sometimes meant that the system had to cut to an intermediate "transition" shot to free up a camera to get the shot that was actually needed.

Professional directors often approach this problem by cutting to a reaction shot from another participant or a "back-up" shot such as a wide shot for a short duration and then using the previously live camera to frame the new shot. In our initial prototype, an overview shot was used as the intermediate shot whenever the live camera needed to switch shots. We reserved one camera for an overview shot at all times and used the other two cameras to frame close-up shots of the three participants. However, as we will discuss in a later section, this choice resulted in several issues related to predictability and lack of variety.

Shot framing

Once a camera was allocated to get a particular shot, the next step was to frame that shot appropriately. Our system draws on the heuristics described earlier, which are implemented as follows.

First, we make use of participant head position and orientation data from the motion capture system. Headroom was created by locating the topmost point of the person's forehead and leaving 250 millimeters space above this point when framing the shot along the vertical dimension. Similarly, using the motion tracker system we located a point approximately 100 millimeters in-front of the foremost headband marker. This point was used as an approximation for the nose position and the center of the frame along the horizontal axis. This resulted in appropriate noseroom and leadroom under different view angles.

Expert Feedback on the Initial Prototype

We captured a 23-minute long meeting using our initial prototype. The meeting involved three participants discussing the Arctic Survival Task [8]. This task was selected to ensure a substantive discussion and active participation. We gathered feedback on the video from our two expert directors by having them watch the video, comment via email, and then meet with the system developers. Their comments fit into four major categories.

Monotonous and Predictable

As noted above, our initial prototype used an overview shot as a back-up shot when cameras were not immediately ready with the next needed shot. Since the discussion in the meeting we recorded was rich with multiple people talking at the same time and people taking turns quickly, the cameras often were not ready to show the new speaker. This resulted in the system defaulting to the overview shot, which led the experts to comment that the system was monotonous and highly predictable. One of the experts commented as follows:

"*There is too much of the wide shot, in my directorial view, so the overall feeling of the video is somewhat repetitive…. Television (and conversation on television) is about people and their faces – we want to see them talk as they converse.*"

Unexpected Cuts

Since participants were often talking over each other, the system could not always determine the focal person based on the available information. This resulted in some awkward cuts. For example, in one case a participant was talking and the system was showing a close-up shot of that person, but suddenly another attendee started talking over the speaker. The system switched the focus to the new speaker and the old speaker could not be seen in the shot at all, even though they were taking rapid turns back and forth. One of the experts suggested that the speaker should not be moved out of the shot halfway through a sentence, and emphasized the following mantra: "*There is a rhythm as to when to cut, and when not to*".

This issue indicates that a system based only on speaker detection may not be effective for capturing meetings with rich discussion since there could be multiple speakers at the same time, and finding the appropriate focus of attention is a difficult problem in these cases.

Slow Reaction Time

In television production, prediction plays an important role in shot framing and cuts. Camerapersons often predict and anticipate how people will move and frame their shots accordingly [5, 13]. Similarly, directors often try to predict the most likely next speaker and try to have a shot of this person ready to show as soon as they begin to talk. In our initial prototype, we did not have any notion of prediction. The system waited until someone spoke; it framed a shot (if not already framed) as soon as the person spoke, and cut to

the shot if the previous shot had been shown for longer than the pre-assigned minimum screen duration. These steps made the system's response time noticeably long. One of the experts commented:

"*The reaction time has to be quicker on the cuts – somebody starts speaking, camera repositions (if necessary) and then cut right away. That's the way a high-speed director works and keeps the audience much more engaged.*"

Lack of Variety

The initial prototype showed two types of shots: close-up shots of attendees and a fixed overview shot. The experts suggested including various overview shots using different cameras and shots with props and artifacts. One of them commented:

"*Consider other shots – for example, when they are talking about 'the list' make it possible to show the list, even if a human being is not standing next to it.*"

Deciding when to frame a shot of artifacts is a difficult problem since recognizing an artifact as the focus of attention (such as a list in the above comment) requires understanding the role of the artifact in the context of the discussion in real-time. However, some non-verbal cues (e.g. gaze, posture) could also be used to estimate the role of such artifacts. In the revised prototype, as we will discuss in a later section, we used this information in combination with the notion of noseroom to make some of these types of shots possible.

Based on the feedback from the experts, we revised our prototype design and ran an evaluation on the revised version.

Revised Prototype Design

The revised prototype was designed for a similar scenario: three participants informally meeting around a table and using a whiteboard. The number of cameras and other hardware were also the same; however, the camera placement and algorithm to select and drive the camera movement were significantly modified.

Modifications in Camera Placement

Following the principle of "camera blocking" from television production, two of the cameras were moved further apart (see Figure 3). This improved the composition of close-up shots (compare Figure 2 (left) with Figure 4 (left)). A person's close-up shot was framed only by the camera directly opposite to him or her. This also provided more depth in the overview shots (see Figure 5).

Use of Gaze and Speaker History for Prediction

In television production, professionals often anticipate the next speaker by determining focus of attention of the participants. In meetings, gaze direction has been shown to indicate people's attention [32, 34, 35].

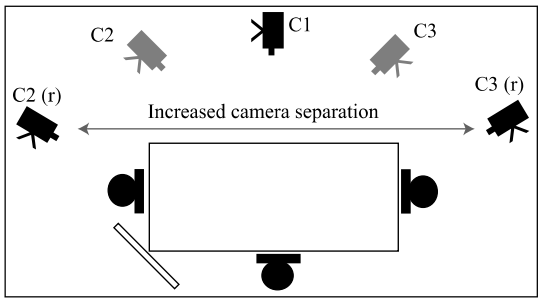

Figure 3. Room layout: C1, C2, C3 represent camera positions in the initial prototype; C1, C2(r), C3(r) represent camera positions in the revised prototype.

In the revised prototype, we used head orientation as an approximation for gaze direction and used it to resolve the focus of people's attention when multiple participants were speaking at the same time. The system tracked the head orientation and estimated the person who was the most popular gaze target. The system then framed a close-up shot of the target and cut to it.

A purely gaze-based prediction and transition, however, could result in a sequence of quickly changing close-up shots if the participants engage in a heated discussion. Therefore, we decided to use this approach only when the current shot "on the air" was an overview shot and multiple participants started talking.

For cases in which the current shot "on the air" was a close-up shot, and multiple participants started talking, we use another prediction strategy that leverages speaker history. This strategy was motivated by the observation that when two people quickly take turns it is possible to predict the next speaker. In our revised prototype, whenever two speakers took turns quickly, the system switched to a two person shot of last two speakers (see Table 1). This increased the probability of keeping the speaker in the shot when a new person starts speaking. This approach also addressed the issue of unexpected cuts in that when the camera shows the two person shot, the previous speaker still remains on screen along with the new speaker.

Variety in Shots

Based on feedback and suggestions from the experts, we included a wider variety of shots in the revised prototype. These shots are commonly used in television production studios to shoot talk shows [2, 38]. Various shots used in the final prototype are shown below.

1. *Close-up shot* (Figure 4 (left)): Often the speaker was shown using this shot. This shot was used in the initial prototype, but the modifications in camera positions now made it possible to frame it more accurately. This shot was also used as a reaction shot we describe later.
2. *Two-person shot* (Figure 4 (right)): Two participants talking at the same time or taking turns quickly.

3. *Overview shot* (Figure 5). Depending on the camera that framed the shot, one of the participants was typically in full facial view while the others were viewed from the side.
4. *Shot of artifacts* (Figure 6): We did not make provisions for explicit shots of artifacts. However, the use of noseroom and view direction allowed a close up of the whiteboard in the vicinity of participants.

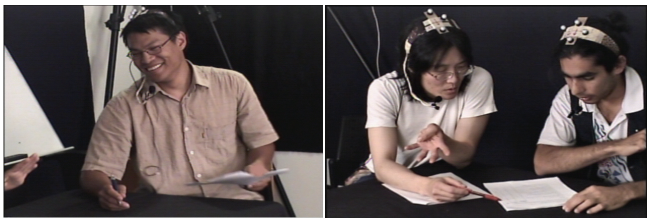

Figure 4. (left) Close-up shot, (right) Two-person shot

Figure 5. Samples of overview shots

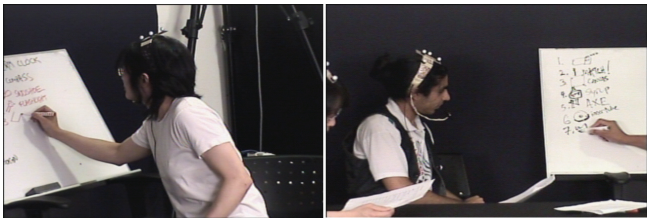

Figure 6. Samples of artifacts shots

Modifications in Camera Control and Shot Transition

The experts commented that our initial prototype defaulted to the wide shot too often. To address this issue, we modified the camera control algorithm. Whenever a camera was not on the air, it framed a close-up shot of a participant directly opposite to it. A constraint was placed so that two cameras did not frame the same person. This configuration had two advantages: (1) if one of the two already framed persons spoke, a close-up shot would be immediately available to cut to, and (2) a close-up reaction shot, instead of a monotonous overview shot, could be used as a transition shot.

In the revised prototype, since there were more shot types and multiple cues, the shot transition rules were more complex (see Table 1). One of the most important differences was the introduction of two-person shot. Although Inoue et al. [11] also used two person shots in their system, but the transition to this shot was purely probabilistic. In our system, most of the transitions were based on verbal or non-verbal cues, since that is how professionals usually decide on shot transitions [5].

Table 1. Shot transition table: the system switches from 'Current shot' to 'Next shot' when the corresponding 'Action/event' happens. A close-up shot or a two-person shot always shows the most recent speaker or the two most recent speakers, respectively.

Current Shot	Action/event	Next shot
Close-up	One person speaks	Close-up
	Two people speak	Two-person
	More people speak	Overview
	Silence	Close-up/Overview (50% probability)
	Maximum screen duration exceeded	Reaction shot of the current speaker's gaze target
Two-person	One person speaks	Close-up
	More people speak	Two-person
Overview	One person speaks	Close-up
	Two people speak	Two-person
	More people speak	Reaction shot of the most popular gaze target

Whenever the system detected that two persons were talking over each other, it framed a two-person shot using the camera which was offline and was opposite to one of the two speakers, and cut to that camera.

A cut to an overview shot was made when: there was silence or everyone was talking at the same time, or someone was standing or moving. The camera to frame the overview shot was selected based on the most recent speaker. This selection added variety and depth to overview shots and made the recent speaker the focus of the shot.

The experts emphasized the role of reaction shots in keeping the video interesting. In order to incorporate this in the revised prototype, whenever a speaker was on-screen for more than the maximum screen duration, the system showed a reaction shot of the speaker's most recent gaze target.

SYSTEM EVALUATION

The goal of the evaluation was to see how well the recordings made by the automated system meet the three conditions stated earlier: 1) informative enough for viewers to understand what took place; 2) compelling to watch; and 3) cost-effective in terms of human production effort.

Since there is no absolute measure to assess how well our system meets these conditions, we opted for a relative evaluation where we compared people's response to a video shot using the automated system to one recorded by a professional television production crew. Our intent was not for the automated system to surpass the performance of the

professional crew, but rather to see how it measured up and if we could gain insights from the comparison.

Both videos were about 40 minutes long and involved 3 people under the Arctic Survival scenario used in our initial prototyping phase. Different sets of three people were used in the two recordings to ensure that the participants in the subsequent comparison phase of the study did not get bored watching two videos with roughly the same content. To ensure a valid comparison, however, both videos were recorded in the same space in our laboratory and using the same cameras. Though the details of the discussion differed across the two videos, they were largely similar in terms of their overall patterns of interaction and artifact usage.

The professional crew were instructed to replicate a professional television studio as closely as possible. A control room was set up in an adjacent space using nine video monitors (3 for camerapersons, 3 for showing camera feeds to the director, 1 for preview, 1 for program, and 1 for transitions), an audio mixer and a video switcher. A director selected and requested shots from the camera operators who controlled the PTZ cameras with a mouse-based interface. They practiced using this interface for about 20 minutes before the recording began. We decided to use the same PTZ cameras for both videos to ensure that the two videos were as similar as possible, and to see how a professional crew made use of them.

Comparative User Evaluation

We selected an approximately 15-minute clip from each video. These clips were selected such that they included frequent interaction with the whiteboard. Since whiteboards are common artifacts in most meetings, this allowed us to compare how well the system handled it as compared to the crew.

Participants and Procedure

11 participants (4 females, M_{age}=26) were recruited at a large North American University and asked to carefully watch these two videos (without rewind or forward) in our laboratory. They were instructed to pay attention to both the content and the quality of the recording, but they were not told that they were evaluating a camera control system. They provided feedback in the following two ways.

First, they were provided with a physical slider at the beginning of the experiment. By moving the slider head, they were able to continuously express their satisfaction (at the integral scale of -3 to +3) with what they were seeing. The center of the slider represented the neutral rating (or 0). Similar techniques have previously been used in focus groups and for measuring emotional responses [16, 30]. There was a small window on the screen showing the value corresponding to the slider head position. These values were recorded by the system once per second. The participants were instructed to use the slider as often as necessary so that it always reflected their satisfaction level with the video coverage (and not the content).

Second, questionnaires were administered at the halfway and end point of each video. They consisted of Likert scale and free response items that asked participants about the video contents and the quality of the coverage. The content questions were asked to ensure and validate that participants were paying attention to the video.

The order in which the two videos were presented was balanced across participants.

Results: How did the videos compare?

Our first question concerned participants' overall satisfaction with the two recordings. To make this comparison, we calculated the mean slider value for each participant under the two conditions by taking the sum of all the slider values and dividing it by the duration in seconds. A dependent sample Wilcoxon Signed Ranks test on the mean values (M_{crew} = 1.1, SD = 0.8; $M_{automatic}$ = 0.6, SD = 0.6) indicated that participants were, on the whole, more satisfied with the crew video than with the automatically shot video by a statistically significant margin ($Z = -2.8$, $p < 0.05$).). The video presentation order did not result in any quantifiable transfer effect.

While we were slightly disappointed that the system did not perform as well as the crew, we were pleased that the average satisfaction level for the automated system was positive, and that the difference between the recordings was not that great (< 1 SD).

To understand the details of these scores, we analyzed the frequency of each satisfaction level in the two videos. We aggregated the time spent by all users under different satisfaction levels and calculated the frequencies. Since the slider values were logged every second, the percentage frequency of a particular satisfaction level (or the corresponding slider value) indicates the percentage of total time the participants felt that particular level of satisfaction while watching the corresponding video (see Figure 7).

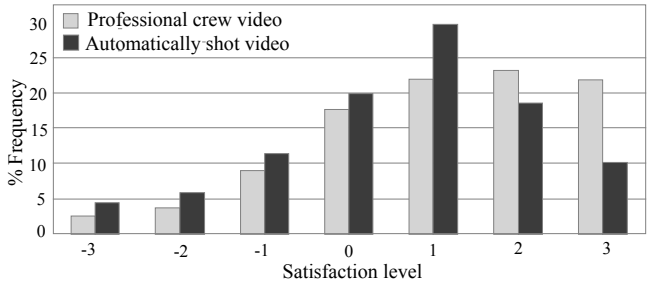

Figure 7. Percentage of time (for all participants) spent under different satisfaction levels for different videos

The frequency data suggest that participants while watching the crew video spent 85% of the playback time in neutral or positive satisfaction level, with approximately equal amount of time in each positive satisfaction level. Whereas, for the automatically shot video, they spent 78% of the playback time in neutral or positive satisfaction level, with 10% of the playback time in the high satisfaction level. This

analysis indicates that the crew video had more instances where participants were highly satisfied, whereas, in the automatically shot video participants were more likely to be neutral or moderately satisfied.

To better understand why participants seemed to be more satisfied with the crew video, we turned to our questionnaire data. One question asked participants how often they wished they could have seen something that was taking place, but could not. Figure 8 (left) shows that most of the participants indicated that this happened "sometimes" when watching the automatically shot video. However, three participants indicated that this happened "often." When we analyzed the free-response comments and midpoint questionnaire results for these three participants, we observed that their responses were at the "sometimes" level after watching the first half of the video, but towards the end they changed it to "often". The reason for this change was our system's inability to properly capture the whiteboard.

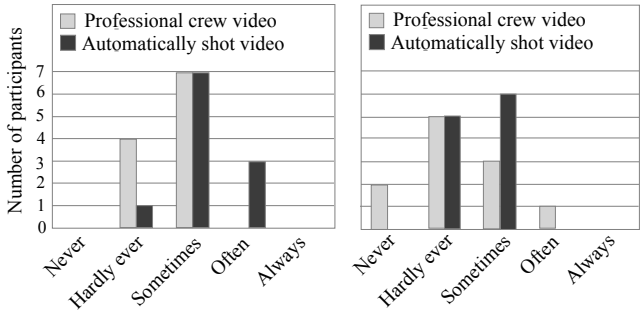

Figure 8. (left) Response distribution for how often participants wished to see desired information but could not. (right) Response distribution for difficulty in figuring out why a particular shot was chosen.

We were also interested in how often participants saw something, but wondered why they were seeing it or wished they could see something else. As it can be seen in Figure 8 (right), there were few times when participants could not figure out why they saw a particular shot, though this did occur somewhat more frequently in the automatically shot video. This indicates that, for both systems, most of the participants were able to figure out most of the time why they saw a particular shot.

Since we significantly modified the shot transition algorithm in the revised prototype, we were interested in estimating the effectiveness of shot transitions. While we relied on the professionals for detailed feedback about this, we asked study participants whether they felt the system was making too many, too few, or about the right number of cuts. As Figure 9 shows, most of the participants were split between "About right" and "A bit too frequently" options.

We further analyzed the data to estimate how much influence the whiteboard coverage had on the participants' responses. The analysis showed that 3 out of 11 participants changed their response from "About right" to "A bit too frequently" after watching the second half of the video. One of them explicitly commented that frequent camera switches away from the whiteboard towards the end were tiring. Furthermore, when we performed Wilcoxon Signed Ranks test on slider values for two halves of the videos separately, we observed that the difference in the rating was significant only in the second half ($Z=-2.2$, $p=0.03$), but not in the first half ($Z=-1.9$, $p=0.06$).

In both the halfway and end point questionnaires, we asked participants if they enjoyed watching the videos. While 9 out of 11 participants had generally positive responses, one was neutral and the other had a negative response. The participant who did not enjoy the videos also rated the shot change frequency as "A bit too frequently" for both videos, and generally preferred wide shots.

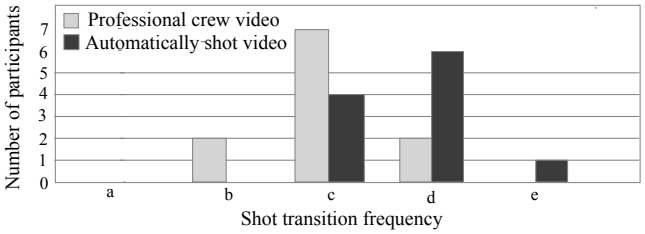

Figure 9. Response distribution for perceived shot transition frequency for the two videos (a: Way too infrequently, b: A bit too infrequently, c: About right, d: A bit too frequently, e: Way too frequently)

Expert Feedback

We showed the videos to an independent expert (initially unaware of our research) who has professional experience in television studio production and is currently an editor in a television studio (this is a different person from the two experts who advised in our design phase). In order to get an unbiased opinion, we showed him both the videos without mentioning their sources. He was also unaware that the two videos were shot live without any post-production step. When asked to compare the two videos, he commented about the automatically shot video:

"*Overall the video was pretty good, because the editing engaged me a little more than the first* [camera crew] *video. Even though it was somewhat lacking in close ups the multiple angles made it somewhat more interesting."*

He also commented that the shot transition frequency was about right in his editorial view. However, he mentioned that the correct shot transition frequency is highly subjective. As discussed previously, this subjective nature is also evident from the distribution obtained in our comparative user study. When we later told him that one of the videos was shot by an automatic system and recorded live, he was surprised. When asked about the effects of the aforementioned issues in the video on the audience, he said that people often look over problems in live settings that would be simply unacceptable if they occurred in movies.

DISCUSSION

We started our iterative design process with the goal of meeting the three conditions outlined earlier. In this section, we assess if we met these three conditions.

Does It Capture Enough Visual Information?

Assessing a system's ability to capture visual information is non-trivial since there is no standard metric for it. In our comparative evaluation, we assessed it based on the user's response to if they could see what they wanted to see. The results indicate that the system succeeded most of the time in providing enough visual information.

Sometimes when users could not see the desired visual information, it was due to the system's inability to capture various artifacts (whiteboard, papers etc.) in the meeting. Previous systems [15] approached this problem in specialized auditorium settings by showing the electronic whiteboards or slides in a separate window. In our more general setting, we attempted to address this general problem by including the artifacts (e.g non-electronic artifacts such as papers, books, coffee mugs etc) in the shots with people using noseroom. While this approach successfully conveyed to the viewer that some activities were being performed on the whiteboard or on the list, it could not effectively capture the details of that activity.

Is It Compelling to Watch?

The analysis of slider data indicates that participants were highly satisfied (+3 level) for approximately 10%, mostly satisfied (+1 to +2 level) for approximately 50%, and neutral (0 level) for approximately 20% of the total playback length of the automatically captured video. The questionnaire data further support this in that participants mostly enjoyed watching the video. The expert's comments were also encouraging in that he found the editing sufficiently engaging. However, a common issue raised by some participants in their comments was that the shot transitions were a bit too frequent.

As far as shot framing was concerned, one participant specifically liked two person shots: "*It did help when two people were shown in frame, to see who was talking. The 3rd voice that was heard often said short phrases, and it could be easily extrapolated that he/she was talking even when not visible.*" Two participants pointed out a few framing issues in the video where a person's forehead went out of the frame, or a part of their body was not included in the frame which should have been included.

Is It Cost Effective?

Although the slider and questionnaire data analysis show that the video produced by our automatic system was not at par with the crew video, the differences were not large. The mean slider value mean difference indicates an overall difference of 0.5 on a 7-point scale, which is less than one standard deviation. Furthermore, the percentage of time for which participants were dissatisfied with the automatically shot video was 22% and that for the crew video was 15%, which is also a relatively small difference.

While the crew video did surpass the automatically shot video in inducing a very high level of satisfaction (22% vs. 10% of the time), this quality came at the cost of three professional camerapersons and a director working for approximately 2 hours (including the set up and planning time) to shoot an approximately 40 minute long meeting. Our automatic system required approximately 10 minutes preparation time and no human intervention during the shooting. To be sure, it did require a substantial investment in motion capture and sound detection equipment. It is our expectation, however, that as the cost of computer vision technologies drops, these costs will fall substantially.

Implications for Practice

Our design process demonstrated that lessons can be learnt from the experts in television production to make meeting capture videos compelling. The use of audio signal level and some non-verbal cues (gaze, posture) in the design have realized a performance approaching that of a professional television production crew. Furthermore, as noted by one of the experts who advised on the design, the prototype has an interesting property that most human television production crews do not have: it does not require the content of the conversation to operate. This makes this prototype essentially language independent. Our evaluation of the prototype also suggests the importance of capturing the usage of non-electronic artifacts in meetings.

The real-time nature of the system means that our results and techniques apply not only to those interested in meeting capture, but also to those developing real-time applications such as video conferencing or webcasting.

LIMITATIONS, CONCLUSIONS AND FUTURE WORK

Visual information captured from meetings is well-known to be monotonous to watch, whereas, the information when captured by professionals is often compelling. Motivated by this observation, we designed and implemented an automatic meeting capture system that uses audio detection and motion tracking to apply various television production principles for capturing meetings.

While prior systems have applied some of these principles to capture lectures in auditorium settings, we extensively explored them to capture dynamic environment of small meeting rooms. A user evaluation of the system indicated that despite its limitations the videos were compelling to watch, and comparable to those shot by professionals.

We plan to further explore issues in handling artifacts in meetings, compare various shot transition strategies, and conduct a long term study to analyze the effects of such meeting capture systems on participants' behavior.

ACKNOWLEDGEMENTS

We thank Dana Lee, Mark Toller, David Solomon and crew, John Hancock, study participants, and DGP members.

REFERENCES

1. Abowd, G. Classroom 2000: an experiment with the instrumentation of a living educational environment. *IBM Systems Journal*, *38,* 4 (1999). 508-530.
2. Arijon, D. *Grammar of the film language*. Silman-James Press, 1991.
3. Bianchi, M.H., AutoAuditorium: a fully automatic, multi-camera system to televise auditorium presentations. In *Joint DARPA/NIST Smart Spaces Technology Workshop* (1998).
4. Birnholtz, J.P., Ranjan, A. and Balakrishnan, R. Using motion tracking data to augment video recordings in experimental social science research. In *E-Social Science* (2007) (in online proceedings).
5. Donald, R. and Spann, T. *Fundamentals of TV production*. Blackwell Publication, 2000.
6. Gaver, W., Sellen, A., Heath, C. and Luff, P. One is not enough: multiple views in a media space. In *InterCHI Conference* (1993), 335-341.
7. Gaver, W.W., The affordances of media spaces for collaboration. In *ACM CSCW* (1999), 17-24.
8. Human-Synergistics, The subarctic survival simulation (http://www.human-synergistics.com.au/content/products/simulations/survival.asp).
9. Indico, http://indico.cern.ch.
10. Inoue, T., Okada, K. and Matsushita, Y. Evaluation of a videoconferencing system based on TV programs. In *IEEE 19th International Convention of Electrical and Electronics Engineers in Israel* (1996), 436-439.
11. Inoue, T., Okada, K. and Matsushita, Y. Learning from TV programs: application of TV presentation to a videoconferencing system. In *ACM UIST* (1995), 147-154.
12. Jaimes, A., Omura, K., Nagamine, T. and Hirata, K. Memory cues for meeting video retrieval. In *ACM CARPE* (2004), 74-85.
13. Kuney, J. *Take one: television directors on directing*. Praeger Publishers, 1990.
14. Liu, Q., Kimber, D., Foote, J., Wilcox, L. and Boreczky, J. FLYSPEC: a multi-user video camera system with hybrid human and automatic control. In *ACM Multimedia* (2002), 484-492.
15. Liu, Q., Rui, Y., Gupta, A. and Cadiz, J.J. Automating camera management for lecture room environments. In *ACM CHI* (2001), 442-449.
16. Lottridge, D. Hedonic affective response as a measure of human performance. http://www.imedia.mie.utoronto.ca/IML/model/technical_reports.php, University of Toronto, Interactive Media Lab, Toronto, 2007.
17. Meetings in america V: meeting of the minds, http://e-meetings.verizonbusiness.com/meetingsinamerica/pdf/MIA5.pdf (2003).
18. Meetings in america: a study of trends, costs, and attitudes toward business travel and teleconferencing, and their impact on productivity, http://e-meetings.verizonbusiness.com/meetingsinamerica/uswhitepaper.php (1999).
19. MeetingSense, www.meetingsense.com/.
20. Nickel, K. and Stiefelhagen, R. Pointing gesture recognition based on 3-D tracking of face, hands and head orientation. In *ACM ICMI* (2003), 140-146.
21. Ou, J., Oh, L.M., Fussell, S.R., Blum, T. and Yang, J. Analyzing and predicting focus of attention in remote collaborative tasks. In *ACM ICMI* (2005), 116-123.
22. Poltrock, S.E. and Engelbeck, G. Requirements for a virtual collocation environment. In *ACM GROUP* (1997), 61-70.
23. Polycom, http://www.polycom.com/.
24. Ranjan, A., Birnholtz, J.P. and Balakrishnan, R. An exploratory analysis of partner action and camera control in a video-mediated collaborative task. In *ACM CSCW* (2006), 403-412.
25. Rosenschein, S.J. Meeting capture: an essential part of the collaboration toolkit, http://www.cxoamerica.com/pastissue/article.asp?art=268314&issue=202#top.
26. Rosenschein, S.J., Quindi meeting companion: a personal meeting-capture tool. In *ACM CARPE* (2004), 112-113.
27. Rubin, A.M. The uses-and-gratifications perspective of media effects. *Media Effects: Advances in theory and persuasion* (2002), 525-548.
28. Rui, Y., Gupta, A. and Cadiz, J.J. Viewing meeting captured by an omni-directional camera. In *ACM CHI* (2001), 450-457.
29. Rui, Y., Gupta, A. and Grudin, J. Videography for telepresentations. In *ACM CHI* (2003), 457-464.
30. Sellen, A.J. Speech patterns in video-mediated conversations. In *ACM CHI* (1992), 49-59.
31. Stiefelhagen, R., Yang, J. and Waibel, A. Modeling focus of attention for meeting indexing. In *ACM Multimedia (Part 1)* (1993), 3-10.
32. Takemae, Y., Otsuka, K. and Mukawa, N. Video cut editing rule based on participants' gaze in multiparty conversation. In *ACM Multimedia* (2003), 303-306.
33. Takemae, Y., Otsuka, K. and Yamato, J. Automatic video editing system using stereo-based head tracking for multiparty conversation. In *ACM CHI extended abstracts* (2005), 1817-1820.
34. Vertegaal, R. The GAZE groupware system: mediating joint attention in multiparty communication and collaboration. In *ACM CHI* (1999), 294-301.
35. Vertegaal, R., Weevers, I., Sohn, C. and Cheung, C. GAZE-2: conveying eye contact in group video conferencing using eye-controlled camera direction. In *ACM CHI* (2003), 521 - 528.
36. Vicon, http://www.vicon.com/.
37. WLAP, http://www.wlap.org/.
38. Zettl, H. *Television production handbook*. Thomson Wadsworth, 2005.