SmartCamera: A Low-cost and Intelligent Camera Management System

Amin Roudaki Jun Kong* Shane Reetz
{amin.roudaki@ndsu.edu, jun.kong@ndus.edu, shane.reetz@ndsu.edu}
Department of Computer Science
North Dakota State University

ABSTRACT

Intelligent camera management systems were developed to automatically record meetings for videoconferencing. These systems provided many benefits, such as reducing the production cost and conveniently documenting events. However, automatically recorded videos in general were not visually engaging. This paper presents a novel approach that intelligently controls camera shots and angles to improve the visual interest. We use 3D infrared images captured by a Kinect sensor to recognize active speakers and their positions in a meeting. A movable camera, constructed by placing a wireless PTZ (pan-tilt-zoom) camera on top of a motorized rail, can automatically move its position to frame an active speaker in the center of the screen. Without interrupting the meeting, a speaker can seamlessly switch video sources through gesture-based commands. We have summarized and implemented a set of heuristic rules to simulate a human director. These rules can be visually edited through a graphical user interface. The customization of a virtual director makes our system applicable in various scenarios. We conducted a user study, and the evaluation results justified the quality of an automated video.

Keywords: Automatic Camera Management, Video Conferencing, 3D Camera.

1. INTRODUCTION

The video communication technique improves the interactivity in remote collaboration, such as supporting synchronous meetings [Tan12] or entailing access to artifacts and resources in a remote space [Nor12]. It brings many benefits to automatically record a meeting, such as facilitating the future review or reducing the production cost. However, an automated video is often not engaging and lacks a visual variety. On the other hand, a professional director produces engaging videos by controlling multiple cameras with a variety of interesting views [Ran10], but human-operated video recording is labor intensive and costly. Various automatic camera management systems [Yu10] have been developed to produce an engaging video. In the automatic video production process, it is critical to recognize a dominant speaker and then accordingly adjust the camera's shot and angle without human involvement. Various approaches have been proposed for controlling a camera to produce diverse shots, such as an omni-directional camera [Rui01, Foo00, Cut02] with a 360-degree view or an array of microphones and PTZ cameras [Ran10].

Previous approaches in general placed a camera in a fixed position. Because each camera could only cover a portion of the whole scene, it was necessary to calibrate multiple cameras. We presented a novel system (i.e., *SmartCamera*), which was featured with a movable camera for covering the whole scene. SmartCamera intended to produce engaging videos with a proper rhythm by considering the following aesthetic elements, i.e., selecting a suitable view (i.e., an overview or close-up shot), providing smooth view transition and determining an appropriate shot length. More specifically, SmartCamera was built on a Kinect sensor that provided an overview shot and captured 3D images for tracking speakers and their gestures. Based on the Kinect sensor, we developed efficient algorithms to detect a dominant speaker (See Section 5.1) and accordingly adjust the angle of a PTZ camera (See Section 6.1). After detecting an active speaker, a movable camera, constructed by mounting a PTZ camera on a motorized rail, moved to an appropriate position with a suitable angle based on the head orientation. In order to provide aesthetic and engaging videos, a virtual director was implemented to manage cameras and determine shot lengths based on a set of heuristic rules. Our work focused on recording a meeting where all speakers' activities were covered within the range of an overview camera, i.e., a Kinect sensor. The meetings where a speaker has his/her back towards the overview camera were out of the scope of this paper. In summary, the major contributions of this paper are summarized as follows.

- **A movable camera.** We designed a movable camera, which was flexible to capture every speaker without any calibration. Furthermore, a movable camera assured that the camera can always directly face the active speaker.
- **Seamlessly switching video sources.** Because it is useful for speakers to access multimedia documents (such as PowerPoint) without interrupting the meeting, our approach supports gesture- and voice-based commands to naturally interact with multimedia contents and to seamlessly switch the video source between a multimedia document and a speaker.
- **A customizable virtual director.** A virtual director is essential to produce an engaging video. It includes a set of director rules that intelligently control a movable camera according to different events, such as the change of speakers. Each director rule is defined through a workflow that specifies a series of camera activities to respond to a specific event. A workflow can be modified visually through a graphical user interface. This customization makes our approach applicable in various scenarios.
- **Efficiently detecting speakers.** To the best of our knowledge, we are the first to utilize the Kinect sensor to automatically record videos. 3D infrared images captured through Kinect are applicable in different lighting conditions to detect the skeleton of each speaker. In addition, Kinect is equipped with multiple microphones which allow us to detect the direction/angle of various active sound sources. The integration of the sound and vision tracking in the Kinect sensor further simplifies the hardware setup.

The remainder of the paper is organized as follows. Section 2 reviews Related Work. Section 3 overviews our approach. Section 4 presents the system architecture. Section 5 discusses the speaker detection and gesture/speech recognition. Section 6 illustrates the camera management. Section 7 explains the virtual director. Sections 8, 9, and 10 introduce an empirical study and analyze the data collected from the study, followed by the conclusion and future work in Section 11.

2. RELATED WORK

An automatic meeting capturing system should address three challenging issues:

1. *Track speakers, their movements, voices, and gestures.* It is challenging to track and capture speakers in a meeting because speakers have different behaviors over time (such as changing his/her body orientation, and starting or stopping speech) [Pol97, Ran08]. Furthermore, speakers may have various requests.
2. *Produce an engaging and attractive video.* An automatically produced video is often unattractive. Consequently, people get bored [Rub02, Ran08, Ran10]. This is mainly caused by the lack of various shots from different angles [Ino95], which forces people to watch the video from a fixed view. Furthermore, an engaging video depends on providing right shots based on the events in a meeting. In order to address this issue, it is important to apply TV production rules to the video production process [Ino95, Bia98, Liu01, Ran08].
3. *No human involvement.* In a meeting, speakers need to focus on the discussion. They should not be disrupted to guide an automatic system to record the meeting [Ran06, Rui03]. Therefore, an automatic system should be intelligent with only minimal manual inputs from speakers.

This section reviews existing approaches that address the above issues.

2.1 Tracking Techniques

Speakers' activities in a meeting provide the necessary information for a director to decide what shots should be taken. Therefore, tracking speakers is an essential component in an automatic video production system.

Various sensing techniques have been proposed in order to track speakers. A microphone array [Bra01, Lee02, Liu01] was developed to recognize the location of a sound source in a physical environment. Rui *et al.* [Rui01] combined a microphone array with an omni-directional camera. This approach was suitable for recording meetings in which people were seated in a circle. Lee *et al.* [Lee02] used an omni-directional camera together with four microphones. This approach used the motion analysis and skin detection to recognize speakers. In order to reduce the hardware cost, Cutler *et al.* [Cut02] used multiple inexpensive cameras to replace an omni-directional camera. These cameras were deployed as a ring to provide a 360-degree view. Similarly, the FLYCAM system [Foo00] used an array of inexpensive cameras, organized in a circular manner, to cap-

ture every direction. This approach also supported the motion detection technique (such as hand movements) to identify active speakers. Nickel *et al.* [Nic05] used two microphones to recognize the sound source and applied the image processing technique to track speakers in a meeting. Ranjan *et al.* [Ran08] used infrared cameras, combined with passive reflective markers, to identify all speakers in a meeting. Although this approach efficiently recognized speakers, it required sticking markers on speakers in advance. Later, Ranjan *et al.* [Ran10] used a high-resolution camera to recognize speakers' faces and positions. In addition, they placed a microphone in the front of each speaker to identify the sound source. Andrey *et al.* [And10] used multiple omni-directional and PTZ cameras on the wall and ceiling to detect all speakers using the face detection technique. Based on Radial Basis Function Networks, Howell and Buxton [How02] recognized speakers' gestures and accordingly adjusted the focus of a camera. Based on existing state-of-the-art audio and video preprocessing blocks, Motlicek *et al.* [Mot13] designed a system that provided real-time analysis of complex audio-visual signals/events. Recently, Gadanac *et al.* [Gal14] proposed to use the Kinect sensor for tracking an active speaker and accordingly controlling a fixed PTZ camera. Different from the above method, our approach controls a moving camera and we implemented a prototype to evaluate the quality of an automatically recorded video.

Identifying speakers based on the image processing and face detection techniques can be erroneous [Ran10]. Andrey *et al.* [And10] also reported that the level of illumination can affect the accuracy of face detection. In addition, Yu *et al.* [Yu10] reviewed various smart meeting recording systems and identified an inefficient face recognition algorithm as one major limitation. In contrast to the previous work, we used a Kinect sensor to track speakers. The Kinect sensor provided 3D infrared images. Compared with 2D color images, the depth information in 3D images made it easy and robust to track speakers and their gestures. Furthermore, our approach used the skeleton of a speaker to derive the head orientation, which was useful to directly film the speaker. In addition, the Kinect sensor included a microphone array which was able to track multiple sound sources simultaneously. The combination of the audio and visual tracking aided the accurate identification of a dominant speaker in a meeting. Recently, Takahashi *et al.* [Tak13] developed a hand-free gesture recognition technique based on a time-of-flight camera that measures the depth to objects. The above approach is capable of detecting small finger-tip gestures while our approach focuses on full-body gestures.

The head tracking has been applied to 3D audio systems. Although multiple views in the videoconferencing attracted a lot of attention, only little research focused on the audio spatialization. In such systems, a virtual sound source was rendered for the audience so that the perceived sound source seemed to be originated from the same location as the remote speaker. Song and Zhang [Son11] utilized the head tracking technique for moving the sweet spot to the position of the listener.

2.2 Camera Management Techniques

An engaging video requires multiple camera shots from various angles. An omni-directional camera [Lee02, Rui01] or its variation [Cut02, Foo00] (i.e., organizing an array of inexpensive cameras as a ring) provides a 360-degree view to capture a close-up shot for any speaker. Other approaches [Ran10, Liu02] used PTZ cameras to provide multiple shots. Because a PTZ camera has a fixed position, it is necessary to include multiple PTZ cameras to cover the whole scene. Multiple PTZ cameras increase the cost, and the calibration process complicates the hardware setup.

A camera with a fixed position may limit the variety of camera shots since the camera cannot be adjusted based on the head orientation of a speaker. Jones *et al.* [Jon09] discussed the importance of eye contact when capturing videos. However, it is challenging to recognize the head orientation based on the image processing techniques [Ran10]. Our approach distinctively combines a 3D infrared camera with a movable camera. The depth information from the 3D camera guides the movable camera to an appropriate position which makes the camera directly face a speaker. Without being limited to a fixed position, a movable camera has the capability to explore a variety of positions and angles to provide the best shot for each speaker.

2.3 Video Directing Techniques

Some approaches [Kun90, Nag09] used post-processing techniques to improve the quality of an automated video. However, some scenarios need the real-time recording. A professional director can intelligently control cameras to change shots based on contextual clues. This ability is crucial for producing an engaging and attractive video. Researchers have summarized a list of heuristic rules [Liu01, Ran10] to simulate an experienced director. However, those rules are suitable for managing cameras with fixed positions. Based on previous experiences and by consulting with professional directors, we developed a set of heuristic rules that specifically fit our hardware design with a moving camera. Those rules synchronized various hardware components (e.g., the movable camera and the Kinect sensor) and chose appropriate shots based on the events in a meeting.

One of the important information sources in a meeting is a media file (i.e., PPT, movie or picture). Gestures have been used to support the camera management in previous approaches (such as grabbing a camera's attention [How02] or changing a camera's shot [Ran10]). Our approach enables the speakers to fully control media files through voice- and gesture-based commands.

Automatic video recording systems [Hec07, Muk99] have been successfully used to record class lectures. Those systems used cameras and microphones to capture the video footage, which was automatically edited in a post-processing step. In-

stead, our approach focuses on the real-time recording. Furthermore, different from a lecture, frequent speaker changes in a meeting make it complicated to cut shots between speakers. With the growth of digitally recorded lectures, it is useful to automatically analyze the contents in a lecture video for highlighting points of interest. Wang *et al.* [Wan07] proposed a novel framework that edited lecture videos based on the analysis of poses, gestures and texts. Especially, a finite state machine that represented editing rules was implemented to produce an engaging lecture video that suited the pace of a presenter's gestures and postures. Later, Wang *et al.* [Wan08] further proposed an efficient gesture detection algorithm for real-time applications. Recently, Zhang [Zha12] analyzed gestures to derive the significant pedagogic information in a single-instructor lecture video. Lecture video systems in general only consider a single presenter/speaker and focus on recognizing the presenter's gesture. On the other hand, SmartCamera emphasizes on providing smooth view transitions between different speakers.

2.4 Techniques Comparison

In all of the above mentioned systems, the ultimate goal is to deliver the best shot at a specific time according to the meeting context. This requirement needs to accurately track meeting participants and active speaker. An engaging video also needs to provide various shots with appropriate angles by framing a person's eye on two thirds of the screen height. Therefore, we compared our approach with other approaches from two large categories, i.e., person detection and camera shot. The two categories are elaborated into ten features, as summarized in Table 1. Based on the comparison, we observed either omni directional camera or multiple cameras are used to simultaneously track multiple meeting participants, while only our approach implements a moving camera to cover all participants. Omni-directional cameras are limited to framing speakers and do not provide diverse shots and multiple cameras increase the complexity of calibration. Our approach is featured by tracking the head angle, which is only supported by one other system [Mot13]. The detection of head angle is important to accurately frame a person, which is an important principle in the video production. The 3D image used in our approach is more reliable than 2D images in other approaches. Our system is one of the few systems, which values the hand-free gestures to control video sources. From the videography point of view, it is important to flexibly capture the video from various angles. Our approach is distinct from other approaches by designing a moving camera, which provides virtually unlimited different shots and thus increases the video engagement. On the other hand, PTZ cameras have the limitation of a fixed position, which provides the same view point. This is different from how a professional records a video. Furthermore, as mentioned by [Ran10], due to a fixed position, a PTZ camera cannot appropriately frame a person if he/she moves or tilts or moves away from a camera. Multiple cameras are used to address this issue [Ran10], but increase the effort of calibrating multiple cameras. Instead, our approach used only a single PTZ camera, but had the full ability to provide any view point.

	Active speaker detection	Multiple speaker position tracking	Head position tracking	Head angle tracking for best shot	Participant Tracking Technology	Camera Type	Control Camera Angle	Change Camera View Point	Gesture detection	Framing based on body/head Pose	Audio Tracking
Automating camera management for lecture room environments [Liu01]	Yes	No	No	No	2D Image	Multiple Fixed	No	No	No	No	Microphone Array
Videography for telepresentations [Rui01]	Yes	No	Yes	No	2D Image	Omni Camera	No	No	No	No	Microphone Array
Viewing Meeting Captured by an Omni-Directional Camera [Rui03]	Yes	Yes	Yes	No	2D Image	Omni Camera	No	No	No	No	Microphone Array
Distributed meetings: a meeting capture and broadcasting system [Cut02]	Yes	Yes	No	No	2D Image	Omni Camera	No	No	No	No	Microphone Array
FLYCAM [Foo00]	Yes	Yes	No	No	2D Image	Omni Camera	No	No	No	No	Microphone Array
A joint particle filter for audio-visual speaker tracking [Nic05]	Yes	No	No	No	2D Image	Multiple Fixed	No	No	No	No	Two Microphones
Improving Meeting Capture by Applying Television Production Principles with Audio and Motion Detection [Ran08]	Yes	Yes	Yes	No	2D Image Infrared Tagging	Multiple Fixed	No	No	No	No	Multiple Microphones
Automatic Camera Control Using Unobtrusive Vision and Audio Tracking [Ran10]	Yes	Yes	Yes	No	2D Image	Multiple PTZ	Yes	No	Yes	Partial	Multiple Microphones
A video monitoring model with a distributed camera system for the smart space [And10]	Yes	Yes	Yes	No	2D Image	1 PTZ 1 Fixed	Yes	No	No	No	Multiple Microphones
Real-time audio-visual analysis for multiperson videoconferencing [Mot13]	Yes	Yes	Yes	Yes	2D Image	Fixed	No	No	Yes	No	Microphone Array
Kinect-based Presenter Tracking Prototype for Videoconferencing [Gal14]	Yes	No	No	No	No	1 PTZ	Yes	No	Yes	No	Microphone Array
SmartCamera	**Yes**	**Yes**	**Yes**	**Yes**	**3D infrared image**	**1 PTZ on rail**	**Yes**	**Yes**	**Yes**	**Yes**	**Microphone array**

					processing	**1 Fixed**					

Table 1. Summary of comparison

3. APPROACH OVERVIEW

There are two major challenges in SmartCamera. First, it is challenging to sense and track the behavior of speakers under different situations with the occurrence of various noises. Second, it is critical to intelligently guide a video production process that involves various actions, such as selecting an appropriate video source and controlling a movable camera. In order to address the above issues, SmartCamera takes the advanced 3D sensing technology for the speaker recognition. A virtual director autonomously controls the video output and reacts to various sensing events according to a set of customizable heuristic rules. In addition, SmartCamera is equipped with a movable camera which provides the flexibility to shoot speakers at various angles and positions. In summary, SmartCamera has the following five design goals.

3.1 A Natural User Interface

The SmartCamera system not only automatically controls a movable camera to take the best shot in a meeting, but also encourages meeting attendees to interact with multimedia contents in a natural way. With SmartCamera, a speaker controls multimedia contents through predefined gesture-based commands without interrupting the meeting. Such a natural user interface allows speakers to focus on the meeting.

3.2 Robustness

Tracking a speaker can be affected by many environmental factors, e.g., noises, echoes, light conditions, or a speaker's body being partially blocked by an object. SmartCamera can mitigate different types of noises.

- **Ambient noises or sound echoes.** Our approach uses the Kinect microphone array to detect the sound sources. When there are multiple sound sources (e.g., one is coming from the dominant speaker and the other is some ambient sound, such as the air conditioning), SmartCamera first uses the Kinect SDK to eliminate the noises and sound echoes, and then applies some heuristics (refer to Section 5.1) to identify the dominant speaker.
- **Improper lighting or a partial view.** The speaker detection and tracking should not be affected by improper lighting (e.g., too dark or too light). The 3D infrared camera in Kinect is applicable to various conditions, and the depth information reduces the computational complexity. Based on the depth information, the Kinect sensor can derive the positions of different portions of a body, which may be blocked by various objects.

3.3 Low Cost

Previous smart camera systems [Foo00, Liu02, Ran10, And10] in general included an expensive panoramic camera or multiple cameras to cover various angles. The essential hardware components in SmartCamera only include one Kinect Sensor (sensing various events and providing an overview shot), one PTZ camera (providing a close-up shot of the dominant speaker), a motorized rail (moving a camera to an appropriate position), and a laptop (running a virtual director). Especially, the Kinect sensor has several functions, including a 3D tracking device, a regular RGB camera for an overview shot, and a microphone array for sound detection. The usage of the Kinect sensor simplifies the hardware setup and reduces the total cost.

3.4 Ease of Setup and Use

We can easily set up and operate SmartCamera in various environments. All hardware components (e.g., a Kinect sensor, a camera, a laptop, and a rail) are portable and can be flexibly connected through Wi-Fi. The movable camera replaces an expensive automated movable arm which is commonly used in a TV production studio, and assures the best angle for framing a speaker based on his/her head orientation. SmartCamera does not need any calibration during the setup process. We simply place the Kinect sensor in the front to cover all speakers. After the setup, the virtual director automatically controls the movable camera without any human involvement.

3.5. Customizable Virtual Director

Automatically directing a video requires intelligent decision making. In practice, a human director provides various shots based on different events that happen in the scene. In SmartCamera, a heuristic rule specifies camera activities in response to a specific event. With a graphical user interface, heuristic rules can be added, removed or edited to fit different scenarios.

4. SYSTEM ARCHITECTURE

In practice, a professional director first places cameras in appropriate locations to cover the whole scene. Then, the director chooses one camera as the output and switches between different cameras upon certain events, such as the change of dominant speakers, gestures, or verbal requests. In a video production process, standard video production rules must be observed to produce an engaging video, such as the maximum time to show a person or the maximum time of an overview shot. In summary, a professional director observes various events and uses his/her video production knowledge to control cameras based on observed events.

SmartCamera simulates a professional director based on four essential hardware components: a Kinect sensor, a PTZ camera, a motorized rail, and a laptop. The Kinect sensor has three important functions: tracking speakers through an infrared camera, identifying the dominant speaker through a microphone array, and providing an overview shot. A PTZ camera mounted on top of a motorized rail constructs a movable camera which takes a close-up shot for an active speaker. The motorized rail is implemented through an EasyDriver stepper motor driver and a stepper motor. The motor driver controls the stepper motor to move a camera on the rail. The stepper motor used in the prototype is powerful to take 4.6 seconds on moving a camera for 1 meter while it is small enough to attach to the rail. In addition, SmartCamera requires a laptop that connects the above hardware components together. The laptop collects from Kinect the environment-sensing information which is used to adjust the position of the movable camera.

The SmartCamera software consists of four main components as shown in Figure 1, i.e., Environment Sensing, Camera Manager, Media Manager, and Virtual Director. The environment sensing component is responsible to analyze and convert environmental sensory data (such as 3D images or sound) to video production events (such as gestures or a switch between active speakers). The camera manager component controls the movable camera to a designated location. The media manager component accesses to multimedia contents. The virtual director component directly interacts with the above three software components. In response to different events received from the environment sensing component, the virtual director either commands the camera manager to take a shot at an appropriate position and angle or switches the video source to a multimedia file through the media manager.

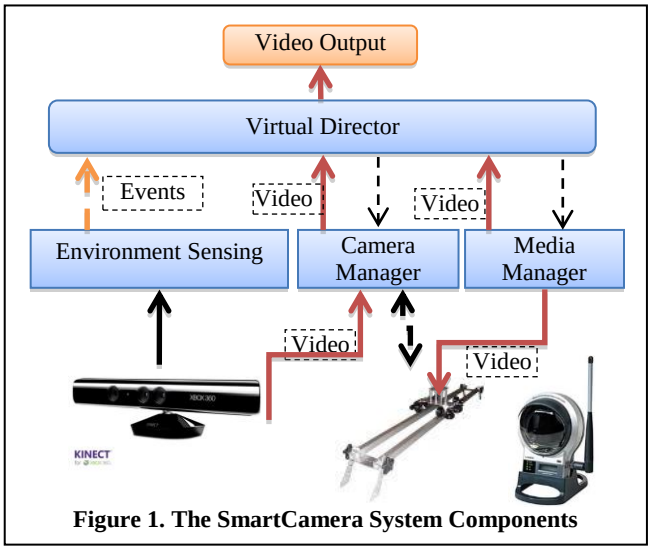

Figure 1. The SmartCamera System Components

5. ENVIRONMENT SENSING

The Microsoft Kinect sensor captures a wide range of sensory information which includes RGB videos, 3D infrared images, and sounds. The sensory data are converted to important video production events which are classified into three categories: 1) user recognition events that are implemented by the user tracking module; 2) speech-based commands that are recognized by the speech analyzer; and 3) gesture-based commands that are identified by the gesture analyzer.

5.1 User tracking

The user tracking module recognizes all meeting attendees, identifies a dominant sound source, and derives the dominant speaker by mapping the dominant sound to one attendee. Different from previous approaches [Ran10, Nic05] which were developed based on 2D images, this paper uses 3D infrared images for a robust body and face recognition. Our approach works even when some parts of a body are not visible. In addition, the depth information is useful to efficiently detect the head orientation of a speaker, which guides a camera to directly face the speaker.

In the user tracking module, the sound analyzer is essential to identify the dominant sound and its position. More specifically, the sound analyzer continuously tracks sound beams and their angles. Each detected sound beam has a dynamic power value which implies the strength of the detected sound. A higher power value means a more reliable and continuous sound. The power value of a newly detected sound is set as zero in the beginning. As the sound is continuously detected, its power value increments every 100 milliseconds, until it reaches a predefined maximum value. Similarly, when an existing sound is not detected any more, its power value decrements every 100 milliseconds and finally reaches zero. Among all detected sound beams, their values are compared every 100 milliseconds and the sound with the largest power value is considered as the dominant sound while other sound beams are treated as noises. The use of a power value avoids falsely recognizing a

sound burst as a new speaker. Since the duration of a sound burst is very short, its power value remains at a low level. Furthermore, the power value is useful to continuously track a speaker. During a meeting, a speaker may pause for a while. During the pause, although the power value of the dominant speaker gradually decrements, it still remains at a relatively high level that avoids losing the dominant speaker. Because the power value of each sound beam is updated and compared every 100 milliseconds, the odds that two sources have the same value are very small. In the case that two sources have the same power value, we randomly choose one as the dominant speaker. In the case of multiple speakers, the speaker, who spoke for the longest time, has the largest power value and thus is recognized as the dominant speaker. After detecting the dominant sound, the analyzer matches its angle to a specific speaker, as shown in Figure 2. More specifically, the center of the head of each person detected by the Kinect depth camera is denoted as (Ux. Uy, Uz). The Kinect SDK allows us to construct a 3D representation of a person head, and accurately calculates the angle that the person head is facing. In order to recognize the dominant speaker, we first calculate the angle of each person with respect to the camera based on the equation User_Angle = ArcTan(Uz/Ux), and then find the minimal absolute angle difference between each person's angle and that of the detected dominant sound. The person with the minimal difference is considered as the dominant speaker. In summary, based on the microphone array, we introduced the power value for each sound beam to eliminate short sound burst and mapped the location of the detected sound to a person to eliminate continuous ambient sound.

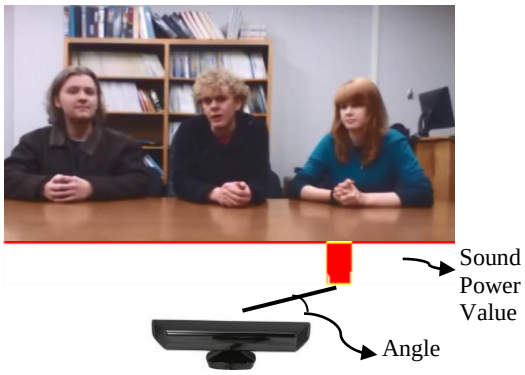

Figure 2. Matching a sound source with a participant

We have evaluated the time taken to recognize a new speaker. In the test, we switched speakers for 20 times and measured the time spent on identifying the new speaker. The result showed that SmartCamera took on average 1.06 seconds (STD=0.19) to report a new speaker.

In summary, the integration of 3D infrared images and a microphone array provides several benefits. First, it avoids the placement of multiple microphones around the scene, and thus simplifies the hardware setup. Second, the 3D infrared images provide a robust user tracking. Finally, the depth information in 3D images can efficiently detect the head orientation of a speaker for taking an engaging shot.

5.2 Speech Analyzer

Speech is a natural communication means in our daily life and has an important role in the video production. It is especially useful when a speaker asks the director to switch the video output to a media file. Furthermore, a speaker may require a specific shot in some scenarios. The speech analyzer component allows a speaker to control a camera or a media file through voice-based commands.

SmartCamera supports seven voice-based commands. In order to issue a voice command, a speaker must first speak the keyword "CAMERA" to activate the command recognition process and then speak one of the seven commands. The two-stage recognition process reduces the probability of recognizing a false command from a conversation in a meeting. The seven commands are listed as follows: 1. Overview shot; 2. Close-up shot; 3. Show this; 4. Show movie; 5. Show picture; 6. Show next; and 7. Show previous. The first three commands are used to control a camera shot. The "Show this" command makes the camera focus on the object in the dominant speaker's hand. These four commands allow a speaker to select a specific shot when needed. The fourth and fifth commands switch the video output to a media file. A speaker uses the last two commands to browse different pages in a media file.

5.3 Gesture Analyzer

Gestures enable people to control multimedia information without interrupting the meeting. In a video production process, gestures can eliminate the sound distraction in a meeting. In order to differentiate gesture-based commands from normal hand movements in a meeting, the gesture analyzer is triggered only after a multimedia document is opened through a voice command. We have defined four gestures to control a media file:

- *Slide a hand to the left.* Flip to the next page.
- *Slide a hand to the right.* Flip to the previous page.
- *Slide a hand top-down.* Stop playing the media file.
- *Slide a hand bottom-up*. Start playing the media file.

The gesture recognition algorithm is implemented by tracking multiple points of a joint. Briefly speaking, we define a starting point for a specific joint and its subsequent movements. A gesture is recognized if the predefined movement of a joint is detected. For example, considering the gesture of sliding a hand to the left, the starting point for the hand joint is defined on right side of the body, and the end point for the same joint is at the left side of the body. All movements between the starting and end points are completed within 1 second.

In summary, the environment sensing component supports controlling the video source through hand-free gestures, which avoids passing a remote controller among meeting participants.

6. VIDEO SOURCES

The output of a video production can be either video streams captured by a camera or a multimedia file. This section discusses the switch between different video sources through the camera manager and the media manager.

6.1 Camera Manager

In SmartCamera, it is critical to intelligently manage cameras to produce the best shot at any time. SmartCamera includes one fixed camera and one movable camera. The fixed camera, which is embedded in the Kinect sensor, is placed in the front to cover all attendees and provides an overview shot, while the movable camera takes close-up shots.

A movable camera, as presented in Figure 3, provides the flexibility to exercise a wide range of possible views. Especially, the capability of changing its position assures that a speaker directly looks at the camera, which makes the video more engaging. More specifically, the advantage of using a track-based camera versus a series of PTZ cameras is the ability to place the camera on any position along the track. This gives the system greater potential to create a well-composed shot, especially if the speaker moves at all during the video shoot. There is also the potential to capture a video while the camera moving on the track, which could add significant aesthetic interest.

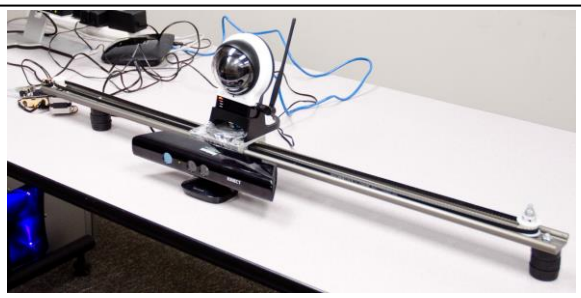

Figure 3. A PTZ camera on a rail

In an initial design, we constructed a movable camera by mounting a camera on top of an iRobot. However, we identified several issues. First, the iRobot makes loud noises when it is moving. Second, the iRobot is often moving off a straight line, which requires an additional camera to monitor and adjust the movement of the iRobot. Third, the movement of the iRobot is slow. These problems lead us to create a new motorized rail system which provides rapid and accurate movements. Furthermore, the rail system does not require any calibrations. We tested the moving speed between two ends (distance = 1 meter) along the track for 20 times. It took on average 4.61 seconds to move a camera from one end to the other (STD=0.34).

The movable camera system is powered by a stepper motor through a belt and two pulleys attached to the rail. The camera itself is fixed to a small cart that fits to the rail with a low friction material. The camera manager communicates with an Arduino microcontroller through a USB port to control the stepper motor and receives video streams from the PTZ camera through Wi-Fi. The movable camera has two states: *Ready* and *On-Move*. By default, the movable camera is in the state of *Ready*. When the movable camera receives a moving command, its state is changed to *On-Move*. When the camera arrives at the designated position, the state is changed back to *Ready*.

Based on the position and the head orientation of a speaker, it is critical to calculate the position of the PTZ camera on the rail and the panning/tilting angle, so that the camera straightly points at the speaker. The position of the PTZ camera is specified as the distance relative to the center of the rail, i.e., d_m in Figure 4. In order to calculate d_m, we need to first determine the panning angle a_1 (refer to Figure 4). Ideally, the Kinect sensor can calculate the panning angle based on the orientation of a speaker's face. We create a 3D model of the person head and combine that with face detection in order to accurately calculate the person head orientation. In the case that a speaker's face is partially visible to the Kinect sensor, we use the shoulder angle to imply the orientation. Eq.1 shows the calculation of the panning angle a_1 based on the positions of the right shoulder (SR_x, SR_y, SR_z) and the left shoulder (SL_x, SL_y, SL_z). Using the panning angle a_1, d_m is calculated according to Eq.2, in which (d_{px}, d_{py}, d_{pz}) is the center of the head of the dominant speaker detected by the Kinect sensor. In the case that a speaker turns his head/body extremely to one side, the calculated d_m may be larger than the half size of the rail, i.e., d_{max}. Then, we move the camera to the end of the rail (i.e., d_{max}) and calculate the camera panning angle a_1' based on Eq.3. The tilting angle a_2 is calculated according to Eq. 4, where d_{py} is the height of the speaker's head center relative to the Kinect sensor. The calculation must consider the height of the PTZ camera relative to the Kinect sensor, i.e., *CameraHeight*, which is 15 cm in our implementation. Tilting the camera according to the head position of a speaker can properly frame the head of the speaker in the center.

$$a_1 = \frac{\pi}{2} - \tan^{-1}\left(\frac{SL_z - SR_z}{SL_x - SR_x}\right) \qquad Eq.1$$

$$d_m = \frac{d_{pz}}{\tan(a_1)} - d_{px} \qquad Eq.2$$

$$a_1{}' = \tan^{-1}\left(\frac{d_{pz}}{d_{px} + d_{max}}\right) \qquad Eq.3$$

$$a_2 = \tan^{-1}\left(\frac{d_{py} - CameraHieght}{d_{pz}}\right) \qquad Eq.4$$

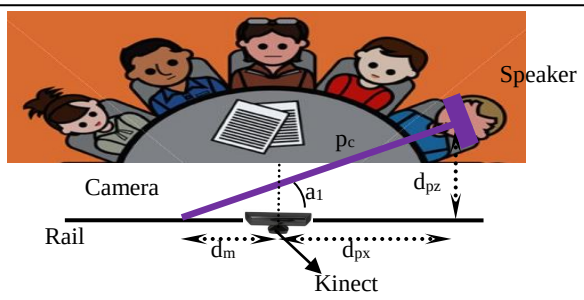

Figure 4. Adjusting a camera based on the speaker

6.2 Media Manager

During a meeting, it is useful to show some pictures or videos related to the meeting, since the presentation media can increase the engagement of audiences and make the topic more understandable and interesting. The media manager component produces video outputs from a media file in a predefined media folder. Meeting attendees can activate the media output by using voice-based commands and then control the media file by speeches or gestures.

7. VIRTUAL DIRECTOR

It is complicated to direct a scene and produce a video. A director should carefully follow the meeting topic and identify the dominant speaker at any time. In addition, the director should continuously listen to various requests from speakers. Based on all signals received from various sources in a meeting, the director decides an appropriate action (e.g., moving cameras to an appropriate location, selecting the output feed, or changing a shot) to make the video engaging.

As presented in Figure 5, the virtual director receives environmental events (such as speaker change) and decides corresponding actions based on a set of heuristic rules. Action commands are sent to various components (e.g., the movable camera or the media manager) to actually perform those actions. Some actions may be performed immediately (such as voice commands), while some may have a delayed effect (e.g., it takes a while to move a camera to a designated position).

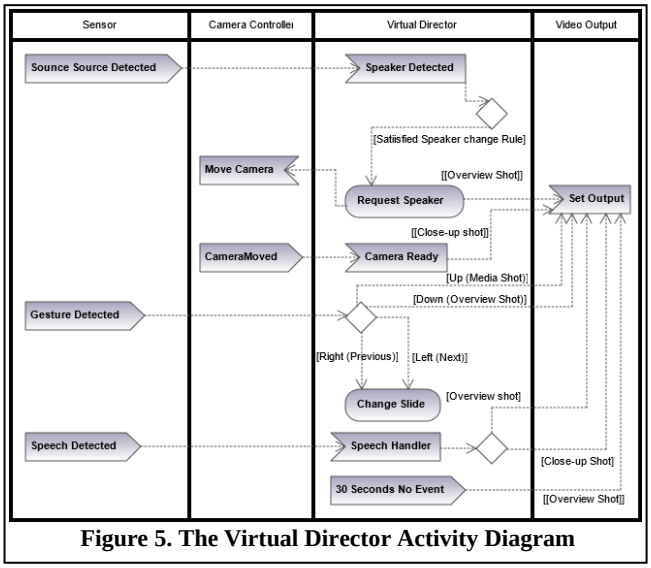

Figure 5. The Virtual Director Activity Diagram

Heuristic rules, which represent video production knowledge, are implemented through a workflow engine. The workflow is embedded in SmartCamera through an open interface. Therefore, we can update a heuristic rule through a graphical user interface without changing the source code. This versatile design can adapt the virtual director to various scenarios. For example, a user may need to switch cameras between speakers rapidly in a fast-paced debate, or to stay on a single camera longer in a slow-paced documentary style. The mapping between an action and a gesture/voice command can also be customized. In addition, the decoupling between the virtual director and other components supports reusing those heuristic rules. For example, if our system is extended with multiple Kinect sensors or cameras, a portion of the virtual director can be reused in the new system. The heuristic rules are grouped into three categories, which are discussed in details in the following subsections.

7.1 Speaker shot

When the user tracking component (Refer to Section 5.1) recognizes a new dominant sound source and identifies the head or body orientation of the new dominant speaker, it triggers a *Speaker Detected* event. At this point, if all of the following conditions are satisfied, the workflow first changes the video output to an overview shot and then repositions the movable camera according to the position of the detected speaker for a close-up shot.

1. The output is not set to a media file (e.g., a slide show).
2. At least 7 seconds have passed since the last shot switch. This condition prevents fast camera switches between different shots, which can be disorienting to the audiences and offset the visual pace of the video. This delay can be shortened or lengthened to match a specific scenario.

3. If the current speaker is the same as the newly detected speaker, the head/body orientation of this speaker must have been changed by more than 15 cm or 20 degrees.
4. The movable camera is not moving. The dominant speaker may be changed when the movable camera is moving to a designated position p_1. In this case, we do not move the camera to a new position until it reaches p_1 first.
5. The total number of speaker transitions within the last minute is less than 5 times. This condition prevents a camera from changing speakers with a close-up shot frequently during a heated discussion. Once the limitation is reached, SmartCamera will provide an overview shot.

7.2 Command

Voice- and gesture-based commands are triggered through *Speech Detected* and *Gesture Detected* events. For example, when a speaker says *Show Picture*, the virtual director will set the video output as a slide show. Then, the speaker can go through the slides by gestures or speeches. All commands, except the *close-up shot* command, are performed immediately once they are recognized. For the close-up shot command, the virtual director first provides an overview shot during the movement of the movable camera. Once the camera reaches the designated location, the output is switched to a close-up shot.

7.3 Timing

In order to make a video more engaging, the director needs to change the shots over time. The following timing rules specify the cutting between different shots.

Close-up Shot Timer. When the virtual director starts a close-up shot, the close-up timer is activated (i.e., *30 Seconds No Event* in Figure 5). This timer will be expired after 30 seconds. The expiration switches a close-up shot to an overview shot, which avoids a continuous fixed view.

No Change Timer. A shot change is disabled, if the last change happened within 7 seconds. This timer prevents a fast switch between different shots, which may cause unpleasant distraction.

Frequent Speaker Transition Timer. By allowing at most five switches of close-up shots during one minute, this timer prevents frequent switches among speakers.

8. AN EMPIRICAL STUDY

The major goal of this study is to evaluate the video production quality and engagement of the SmartCamera system. In this study, we invited three people from an *Improv Comedy Club* at a Midwest university to discuss their club. The SmartCamera system automatically filmed the discussion. At the same time, a professional camera-man was recruited to rec-

ord the discussion. The SmartCamera system produced an output video in real-time, while the professional needed a post-processing step to mix and select shots to produce the final output. In the evaluation, all important features in our system, i.e., different video sources (between a multimedia file and a camera), different speakers, and different camera views (overview and close-up). Similar scenarios, which included three speakers, were tested in other studies, such as [Ran08] and [Ran10].

The objective of our system is to provide engaging videos. Therefore, we chose to compare our system with a professional camera man. If comparing with an automatic system, we cannot fully justify the quality of our system even when our rating is better than that of a related system since both systems may not be engaging. Furthermore, the evaluation result may be biased if we did not set up the related system appropriately.

We apply a between-group evaluation to compare the quality of the video produced by SmartCamera with that produced by the professional. Participants are randomly divided into two groups. The first group (referred to as **G1** in the following) watches the video produced by SmartCamera, and the second group (referred to as **G2** in the following) watches the other video. Each group has 27 subjects. In order to avoid biased opinions, participants were not informed on how a video was produced. After watching the video, each participant was asked to complete a questionnaire that used a 5-point Likert-type scale (i.e., ranging from "1- Strongly disagree" to "5- Strongly agree"). The questionnaire was designed to evaluate the quality a produced video from the perspectives of view selection (i.e., an overview or close-up shot), shot length and view transition (See Tables 1 and 2).

Research Question and Hypotheses

Based on the Goal Question Metric (GQM) approach [Bas94], we define the following goals and hypotheses.

Goal 1: Compare the engagement and the overall quality of the video produced by SmartCamera with the video produced by a human professional.

Hypothesis 1: Participants rated the SmartCamera video similar to the human directed video in terms of the engagement and overall production quality.

Goal 2: Compare SmartCamera and human directed videos in terms of the overview and close-up shots.

Hypothesis 2: Participants rated the SmartCamera video similar to the human directed video in terms of the overview and close-up shots.

Participating Subjects

Fifty four undergraduate students enrolled in the Computer Science and Information Science programs at a Midwest University participated in this study. The subjects were randomly selected and were not specifically targeted to benefit the study results.

Artifact/Videos

Both videos have the same resolution, frame rate, length and contents. Each video mainly consists of two sessions. In the first session, speakers introduce their club through PowerPoint slides which are controlled by voice and gesture commands. SmartCamera automatically provides the PowerPoint slides as the designated output. In the human-directed video, the professional needs to manually choose the slides as the output in the post-processing step. In the second session, speakers give detailed information about the work they had performed. The second session involves speaker recognition, speaker transition and camera switches between the overview and close-up shots.

9. DATA ANALYSIS

This section provides an analysis on the data collected from the study. An alpha value of 0.05 was selected for judging the significance of the results.

	Median		**STD**		**Mean Ranks**		**P**
	G1	**G2**	**G1**	**G2**	**G1**	**G2**	
1. You find the video interesting or engaging.	3	3	0.92	1.06	28.02	26.98	0.82
2. Rate the quality of the video production as a whole.	3	3	0.68	0.89	23.55	31.44	0.067
3. The camera views were switched with the right amount of times.	3	3	1.13	0.94	25.3	29.7	0.31
4. Overall, the selected view and the timing of shot switch were implemented appropriately.	3	3	0.85	0.79	20.15	34.85	0.0006
5. The director framed the speaker well.	2	4	1.05	0.9	19.26	35.74	0.0001

Table 2. Overall production quality result

H1: COMPARISON OF SUBJECTIVE FEEDBACK ON THE ENGAGEMENT AND VIDEO PRODUCTION QUALITY

Participants were asked to evaluate the engagement and the overall video quality based on five questions listed in Table 2, in which G1 represents the first group that watched the SmartCamera video, while G2 indicates the second group that watched the human-directed video. Since there were two independent unpaired groups, we performed a two-tail Mann-Whitney's U test. Shaded cells in Table 2 indicate a significant difference (i.e., $p < 0.05$). We found that the rates about the engagement were close without a significant difference between two videos. This result indicated we achieved the major design goal of producing an interesting and engaging video, which was one of the largest challenging issues in the automatic video production [Rub02, Ran08, Ran10, Yu10]. The human-directed video had a higher mean rank on the overall quality (i.e., the second question in Table 2) than the SmartCamera video. However, no significant difference was observed ($p=0.067$). Similarly, both videos had similar rates on the third question.

The human-directed video had a higher rate on both the camera switch time (i.e., question 4) and framing (i.e., question 5). This result meets our expectation since a human director can understand the content of a discussion and accordingly control cameras faster and more precisely. In the evaluation, we observed that SmartCamera had some delays to move a camera to a designated position when speakers switched frequently and fast in a discussion. The delay is caused by the following three reasons. First, it takes a while to move a camera to the designated position (refer to Section 6.1). Second, although the power value in the sound analyzer avoids falsely recognizing a sound burst, it increases the time to detect a new sound source. However, the delay caused by the sound analyzer is limited within 1.06 seconds (refer to Section 5.1). Third, the 7-second rule in the virtual director (refer to Section 7), which avoids a disrupting and frequent switch between different shots, causes delays in pointing the camera at a new speaker. More specifically, if the first speaker speaks for less than 7 seconds, the movable camera will not point at the second speaker until the 7-second timer is expired. In order to minimize the delay of shot switches, we can consider updating the hardware with a faster and more powerful motor, which requires a tight touch between belt and pulley and thus increases challenge in the engineering. More important, the camera movement can be optimized. During an overview shot, we can move the camera from its current position to the center of the rail, which may potentially reduce the average moving distance. Finally, the virtual director (especially the 7-second rule) should be elaborated to balance the frequent switch between various shots and the delay of pointing the movable camera at a speaker. The quality of framing in SmartCamera is reduced due to the following two reasons. First, if a speaker turns his head/body to one side, the Kinect sensor cannot capture the speaker's face/body clearly, which causes inaccurate body recognition. Second, when a speaker is speaking, his/her head/body movements need to adjust the position of the camera. However, SmartCamera currently only considers large movements (more than 15 cm or 20 degrees). In summary, the framing issue is caused by body recog-

nition, especially when a speaker does not face the Kinect sensor well. One possible improvement is to employ multiple sensors in various positions to capture different angles of a speaker and accordingly construct a full 3D view [Wil12]. In addition, a longer rail for a PTZ camera could provide a better angle and more space to move the camera.

SmartCamera can potentially benefit from the newest development in the vision-base head tracking. Recently, based on the new camera models that applied the structured light technology (e.g., the Kinect sensor), some robust, real-time head tracking algorithms were proposed. For example, Suau *et al.* [Sua12] utilized the depth information to estimate the position of the head, which showed high robustness against partial occlusions and fast movements. The inclusion of those new achievements can increase the accuracy of the head tracking to adjust the movable camera more precisely in SmartCamera.

In summary, the engagement and the overall quality of the SmartCamera video are close to that of a human directed video. However, the quality of camera switches and framing in SmartCamera still needs improvements.

H2: COMPARISON OF SUBJECTIVE FEEDBACKS ON TIMING /LENGTH OF OVERVIEW AND CLOSE-UP SHOTS

In addition to the engagement and the overall quality, we also performed a two-tail Mann-Whitney's U test on the quality of the overview and close-up shots from the perspectives of timing and length, as presented in Table 3. A significant difference is observed on questions 2 and 3 in Table 3. In SmartCamera, the length of an overview shot is co-related with the timing of switching to a close-up shot. More specifically, in order to switch to a close-up shot, SmartCamera first needs to move a camera to the designated position. While the camera is moving, SmartCamera continues providing an overview shot as the output. Consequently, the duration of an overview shot is lengthened, while the duration of a close-up shot is shortened. In order to address the above issue, it is critical to reduce the transition time from an overview shot to a close-up shot by moving the camera faster to its designated position. On the other hand, SmartCamera can immediately switch from a close-up shot to an overview shot. Therefore, participants were satisfied with the timing of switching to an overview shot (question 1 in Table 3).

	Median		**STD**		**Mean Ranks**		**P**
	G1	**G2**	**G1**	**G2**	**G1**	**G2**	
1. The director switched to the overview shot at the right time.	3	4	0.97	0.79	23.59	31.40	0.069
2. The length of the overview shot was appropriate.	3	4	1.12	0.83	21.59	33.40	0.006
3. The director switched to the close-up shot at	2	3	1.02	0.83	19.69	35.31	0.0002

the right time.							
4. The length of the close-up shot was appro-priate.	3	4	0.98	0.89	23.31	31.68	0.052

Table 3. Overview/close-up shots

10. QUANTITATIVE ANLSYSIS

This section provides an in-depth comparison between two videos according to the number of overview/close-up shots and the timing of camera switches. Such a comparison quantitatively evaluates the quality of an automated video.

The total length of both videos is 5 minutes and 45 seconds. Table 3 compares the timing of different camera shots in both videos. The first column displays the timing of each shot in the human directed video, along with the selected type of the shot (i.e., an overview or close-up shot), while the second column shows the corresponding information in the SmartCamera video. Overall, 12 close-up shots are captured by a human director, versus 14 close-up shots in SmartCamera. 11 close-up shots are matched between two videos. Therefore, SmartCamera has 92% recall rate compared with the human directed video. The high recall rate indicates that the sound analyzer in SmartCamera is efficient in detecting a dominant speaker, and the virtual director makes a correct decision to switch from an overview shot to a close-up shot. Out of 15 overview shots in the SmartCamera video, 13 shots match with the human directed video, i.e., 87% recall rate. After watching both videos, we observed that the automated video switched shots more frequently than the human-directed one, especially in the beginning and end of a video. According to the discussion context, the beginning or the end of a meeting is to introduce or summarize the meeting topic. Therefore, a professional chose an overview shot. On the other hand, SmartCamera did not understand the context to decide the beginning or end of a meeting.

Human Director		**SmartCamera**		**Diff**
Time	**Shot**	**Time**	**Shot**	
00:00	Overview	00:00	Overview	0
		00:06	Close-up	
00:14	Slides	00:14	Slides	0
01:50	Overview	01:50	Overview	0
02:00	Close-up	02:00	Close-up	0
		02:11	Overview	
		02:18	Close-up	
02:20	Overview	02:22	Overview	2
02:34	Close-up	02:28	Close-up	6
02:47	Overview	02:40	Overview	7
02:53	Close-up	02:46	Close-up	7
02:56	Overview	02:50	Overview	6
03:01	Close-up	02:57	Close-up	4
03:07	Overview	03:01	Overview	6
03:15	Close-up	03:19	Close-up	4
03:19	Overview	03:25	Overview	6
03:29	Close-up			

03:41	Overview			
03:50	Close-up	03:49	Close-up	1
03:56	Overview	03:58	Overview	2
04:04	Close-up	04:04	Close-up	0
04:07	Overview	04:09	Overview	2
04:17	Close-up	04:15	Close-up	2
04:47	Overview	04:46	Overview	1
04:54	Close-up	04:51	Close-up	3
04:59	Overview	04:59	Overview	0
05:05	Close-up	05:02	Close-up	3
05:17	Overview	05:19	Overview	2
05:23	Close-up	05:25	Close-up	2
05:30	Overview	05:32	Overview	2
		05:37	Close-up	
		05:43	Overview	

Table 3. Comparison of shots timing

In addition to the correct selection of a camera shot, the timing of each shot is also important to the quality of a video. The last column in Table 3 provides the time difference on each shot between two videos in seconds. The average difference on close-up shots is 2.91 seconds (STD=2.26), while the average difference on overview shots is 2.77 seconds (STD=2.55). In summary, the timing of shot switches in the SmartCamera video has less than 3-second difference from the human directed video.

11. CONCLUSION AND FUTURE WORK

In this paper, we have presented a low-cost and intelligent camera management system, called *SmartCamera*, which has several advantages, such as being robust and ease of setup. Different from previous approaches based on 2D images, our approach uses 3D images captured by a Kinect sensor to efficiently recognize the skeleton of each meeting attendee. We have designed a movable camera so that a close-up shot is captured based on the head/body orientation of a speaker. Besides, SmartCamera supports voice- and gesture-based commands to control media files (e.g., pictures and movies) during a meeting. This feature allows speakers to naturally interact with multimedia documents without interrupting the discussion. We have summarized a set of video production rules that intelligently direct the video production and control the camera shots based on various sensory data received from Kinect. An empirical user study justifies the quality of an automated video produced by SmartCamera.

The empirical study in this paper focuses on evaluating the quality and engagement of an automatically recorded video. Future evaluation will compare the usability of hand-free gestures with traditional controlling methods, such as a remote controller. Future evaluation will also include comparing our approach with similar systems that were implemented based on multiple PTZ cameras.

The SmartCamera system is not limited to record meetings. For example, by changing the virtual director, we can adapt SmartCamera to record a lecture in a classroom. The flexibility of an open workflow-based virtual director makes it easy to extend or adapt SmartCamera. The future work includes applying and testing SmartCamera in different scenarios. The number of speakers which can be covered in SmartCamera is limited by the Kinect sensor's field of view. If one Kinect sensor cannot cover all speakers in a meeting, we can divide the whole space into several non-overlapping regions, each of which is covered by one Kinect sensor. Therefore, SmartCamera is flexible to cover a small scene, while it is also scalable to support a large number of speakers. In the future, we will extend SmartCamera with several Kinect sensors.

REFERENCES

[And10] Andrey L. Ronzhin, Maria Prischepa, and Alexey Karpov. A video monitoring model with a distributed camera system for the smart space. *Proc. ruSMART/NEW2AN'10*, Springer-Verlag, (2010), 102-110.

[Bas94] Basili, V.R., Caldiera, G., Rombach, H. D. , The Goal Question Metric Approach, *Technical Report, Department of Computer Science*, University of Maryland, (1994), ftp://ftp.cs.umd.edu/pub/sel/papers/gqm.pdf

[Bia98] Bianchi, M. AutoAuditorium: A Fully Automatic, Multi-Camera System to Televise Auditorium Presentation, In *Proc. Joint DARPA/NIST Smart Spaces Technology Workshop*, (1998)

[Bra01] Brandstein, M. and Ward, D. Microphone Arrays: Signal Processing Techniques and Applications. *Springer Verlag*, (2001).

[Cut02] Cutler, R., Rui, Y., Gupta, A., Cadiz, J., Tashev, I., He, I., Colburn, A., Zhang, Z., Liu, Z., and Silverberg, S. Distributed meetings: a meeting capture and broadcasting system. *Proc. Multimedia*, ACM (2002) 503-512.

[Foo00] Foote, J. and Kimber, D. FlyCam: practical panoramic video. *Proc. MULTIMEDIA*. ACM, (2000) 487-488.

[Gad14] Gadanac, D., Ericsson Nikola Tesla d. d., Zagreb, Croatia.,Dujak, M., Tomic, D. and Jercic, D. Kinect-based presenter tracking prototype for videoconferencing *Proc.* MIPRO, (2014) 485-490.

[Hec07] Heck, R., Wallick, M., Gleicher. M. Virtual Videography. *ACM Transactions on Multimedia Computing, Communications, and Applications* (2007), Vol. 3(1).

[How02] Howell, A. J. and Buxton, H. Visually mediated interaction using learnt gestures and camera control. *HCI 2002*. Springer-Verlag. (2002) 272-284.

[Ino95] Inoue, T., Okada, K. and Matsushita, Y. Learning from TV programs: application of TV presentation to a videoconferencing system. *Proc. UIST 1995*, ACM Press, (1995) 147-154.

[Jon09] Jones, A., Lang, A., Fyffe, G., Yu, X., Busch, J., McDowall, I., Bolas, M., and Debevec, P. Achieving eye contact in a one-to-many 3D video teleconferencing system. *ACM Trans. Graph July* (2009). *28, 3,* Article 64.

[Kun90] Kuney, J. Take One: Television Directors on Directing. *Praeger Publishers,* (1990).

[Lee02] Lee, D., Erol, B., Graham, J., Hull, J. and Murata, N. Portable meeting recorder. *Proc. MULTIMEDIA*, ACM (2002), 493-502.

[Liu01] Liu, Q., Rui, Y., Gupta, A., and Cadiz, J. J. Automating camera management for lecture room environments. In *Proc. CHI 2001*. ACM (2001), 442-449.

[Liu02] Liu, Q., Kimber, D., Foote, J., Wilcox, L., and Boreczky, J. FlySPEC: a multi-user video camera system with hybrid human and automatic control. *Proc. Multimedia 2002*. ACM (2002), 484-492.

[Mot13] Motlicek, P., Duffner, S., Korchagin, D., Bourlard, H., Scheffler, C., Odobez, J. M., Galdo, G., Kallinger, M., and Thiergart, O. Real-Time Audio-Visual Analysis for Multiperson Videoconferencing. *Advances in Multimedia* (2013), Volume 2013, Article ID 175745.

[Muk99] Mukhopadhyay, S., Smith, B. Passive Capture and Structuring of Lectures. *Proc. Multimedia'99*, (1999), 477-487.

[Nag09] Nagai, T. Automated lecture recording system with AVCHD camcorder and microserver, *Proc. SIGUCCS* (2009), 47-54.

[Nic05] Nickel, K., Gehrig, T., Stiefelhagen, R., and McDonough, R. A joint particle filter for audio-visual speaker tracking. *Proc. ICMI 2005*. ACM (2005), 61-68.

[Nor12] Norris, J., Schnadelbach, H., Qiu, G. CamBlend: An Object Focused Collaboration Tool. Proc. CHI' 12, (2012) 627-636.

[Pol97] Poltrock, S.E. and Engelbeck, G. Requirements for a virtual collocation environment. In *ACM GROUP* (1997), 61-70.

[Ran06] Ranjan, A., Birnholtz, J.P. and Balakrishnan, R. An exploratory analysis of partner action and camera control in a video-mediated collaborative task. *Proc. ACM CSCW,* (2006) 403-412.

[Ran08] Ranjan, A., Birnholtz, J.P. and Balakrishnan, R., Improving meeting capture by applying television production principles with audio and motion detection. *Proc. CHI 2008,* ACM (2008) 227-236.

[Ran10] Ranjan, A., Henrikson, R., Birnholtz, J., Balakrishnan, R., and Lee, D., Automatic camera control using unobtrusive vision and audio tracking. *Proc. Graphics Interface 2010*. ACM (2010), 47-54.

[Rub02] Rubin, A.M., The uses-and-gratifications perspective of media effects. *Media Effects: Advances in theory and persuasion* (2002), 525-548.

[Rui01] Rui, Y., Gupta, A., and Cadiz, J. J. Viewing Meeting Captured by an Omni-Directional Camera. *Proc. CHI 2001*, ACM (2001), 450-457.

[Rui03] Rui, Y., Gupta, A. and Grudin, J. Videography for telepresentations. *Proc. CHI 2003*, ACM, (2003) 457-464.

[Son11] Song, M. S., Zhang, C., Florencio, D., and Kang, H. G. An Interactive 3-D Audio System With LoudSpeakers. *IEEE Transactions on Multimedia* (2011), Vol. 13(5), 844-855.

[Sua12] Suau, X., Ruiz-Hidalgo, J., and Casas, J. R. Real-Time Head and Hand Tracking Based on 2.5D Data. *IEEE Transactions on Multimedia* (2012), Vol. 14(3), 575-585.

[Tak13] Takahashi, M., Fujii, M., Naemura, M., and Satoh, S. Human Gesture Recognition System for TV Viewing Using Time-of-Flight Camera. *Multimedia Tools and Applications* (2013), Vol. 62, 761-783.

[Tan12] Tang, J. C., Marlow, J., Hoff, A., Roseway, A., Inkpen, K., Zhao, C., and Cao, X. Time Travel Proxy: Using Lightweight Video Recordings to Create Asynchronous, Interactive Meetings. *Proc. CHI (2012)*, 3111-3120.

[Wan07] Wang, F., Ngo, C. W., and Pong, T. C. Lecture Video Enhancement and Editing by Integrating Posture, Gesture, and Text. *IEEE Transactions on Multimedia* (2007), Vol. 9(2), 397-409.

[Wan08] Wang, F., Ngo, C. W., and Pong, T. C. Simulating A Smartboard by Real-time Gesture Detection in Lecture Videos. *IEEE Transactions on Multimedia* (2008), Vol. 10(5), 926-935.

[Wil12] Williamson, B., LaViola, J., Roberts, T., and Garrity, P. Multi-Kinect Tracking for Dismounted Soldier Training. *Proc. Interservice/Industry Training, Simulation, and Education Conference* (2012), 1727-1735.

[Yu10] Yu, Z. and Nakamura, Y. Smart Meeting Systems: A Survey of State-of-the-art and Open Issues. *ACM Computing Surveys*, (2010), *Vol. 42, No. 2*, Article 8.

[Zha12] Zhang, J. R. Upper Body Gestures in Lecture Videos: Indexing and Correlating to Pedagogical Significance. *Proc. MM* (2012), 1389-1392.