A REAL TIME ARCHIVING SYSTEM BASED ON AUDIO-VISUAL EVENTS

Xin Li, Luo Sun, Linmi Tao, Guangyou Xu and Ying Jia
Key Laboratory of Pervasive Computing, Ministry of Education
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
{x-li02, sunluo00}@mails.tsinghua.edu.cn {linmi, xgy-dcs}@tsinghua.edu.cn ying.jia@intel.com

ABSTRACT

It is of a great value to archive human activities such as meetings, lectures, and seminars in real time, so that people can quickly retrieve the information through important events or objects in the scene later. Traditional archiving methods usually index audio video data with constructive information, such as shot detection. In this paper we present a new archiving strategy based on real time audio-visual events, which are reasoned from rich spot information such as the speaker location captured by microphone arrays. To this end the information from microphone array and multiple cameras are fused under the frame of particle filter. The results then contain semantic information and become more reasonable and accurate.

1. INTRODUCTION

With the development of multimedia technology, the recording and archiving of human activities, such as meetings, lectures and seminars by computer have become prevalent. This archiving strategy has advances in at least two aspects. First, it enhances the traditional ways of meeting recording by grasping every detail of the meeting, rather than only taking simple notes. Second, it can add indexes to the recordings so that people can quickly retrieve important information which they are interested in.

Aiming at these advances, several systems have been developed in recent years: the Distributed Meetings system designed by Microsoft Research [1], the meeting capture system at University of Michigan [2] and so on. All these systems are good examples of capturing and recording meetings by computer. But they also have shortcomings in some aspect: first, they usually process the recordings offline, thus some valuable information, such as the Sound Source Localization obtained by Microphone Array, is lost. Second, they often index the recordings with constructive information, such as segmenting speakers by clustering of SSL angles in [1], or correlating time-stamping notes to multimedia data in [2]. This approach tends to ignore the semantic aspect of the meeting content, and sometimes results in indexes that are not reasonable or accurate.

The Interactive Systems Laboratories at CMU recently developed a system to record the meeting in a somewhat real time manner [4]. Their work emphasizes on the understanding of meeting contents by people identification and speech recognition. While in this paper we designed a system that emphasizes on recording and indexing meetings or seminars based on semantic cues (events) in real time. We define some commonly occurring and important activities in the meetings as events. Then we use real time information captured by camera and Microphone Array to detect these events and index the recording accordingly. As a result, the archiving will become more reasonable and accurate.

The rest of the paper is organized as follows: section 2 briefly describes our scenario. Section 3 discusses the system architecture and data flow. In section 4 we give some experiment results. Section 5 draws a brief conclusion and discusses on future work.

2. SCENARIO

In this section we briefly describe our meeting room setup.
A picture indicating a typical meeting room environment is shown below: There may be cameras, Microphone Arrays, platform, whiteboard, projectors and many other instruments in this room.

Fig.1. Meeting Room Picture

As to the focus of our application, the meeting room can be divided into two parts, the front platform and the audience seats, as shown in the figure below.

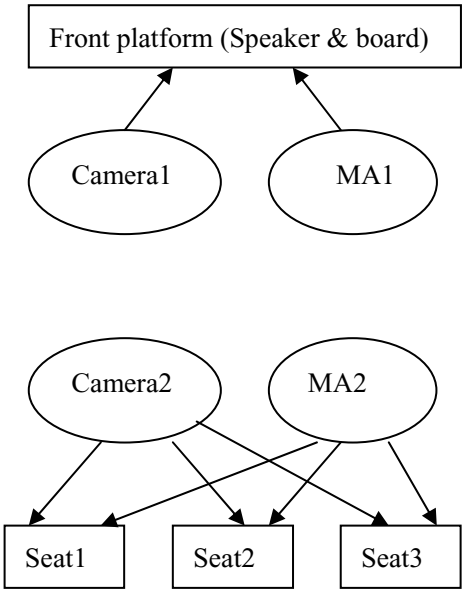

Fig.2. Meeting Room Setup

On the front platform, a speaker can walk around and talk. We place a camera and a Microphone Array (MA) pointing at the platform, these instruments are used to track the speaker in real time based on our previously developed audio visual fusion method for human tracking [3]. In the back of the meeting room there are several seats for the audience. Another camera and Microphone Array are pointing at them. The speaker in the audience can then be detected. Due to the limitation of the Microphone Array, we assume that only one person speaks at any time. The content on the whiteboard can be obtained by camera or Pen Input system.

3. SYSTEM DESCRIPTION

Our system architecture and data flow can be shown in the following diagram:

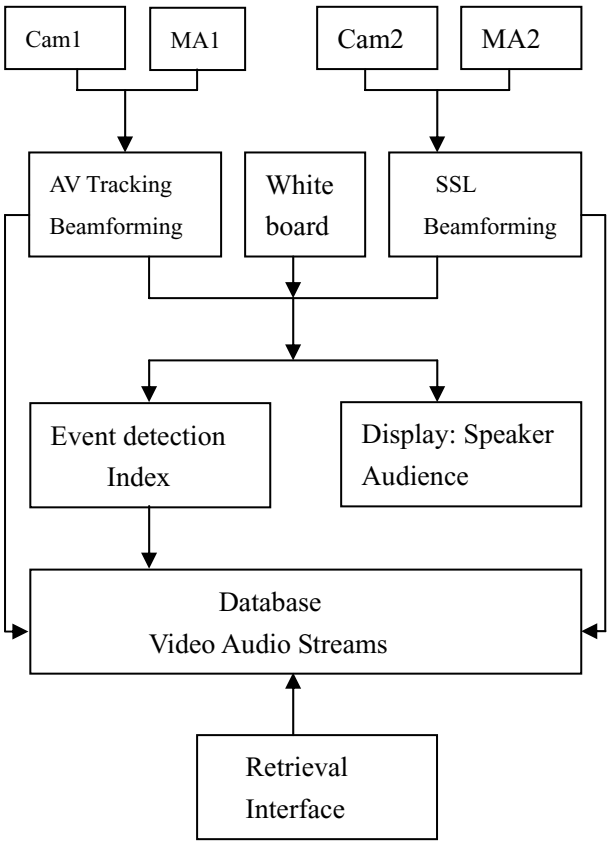

Fig.3. System Architecture

The information of the speaker on the front platform captured by Camera1 and MA1 are sent to the AV Tracking module, the speaker's activity is tracked and his voice is gathered by beamforming of the Microphone Array. Similarly, the speaker in the audience is detected by SSL and his voice is also gathered by beamforming. The video audio streams are then stored in the database and at the same time sent to the event diction module and display module. The display module simply displays speakers. In the event detection module, events are detected based on the information of audio video streams and content of the white board. The indexes are then added to the database. The retrieval interface module is used for

retrieving information in the database.

We now discuss the AV Tracking module, the SSL & beamforming module in some details and then emphasize on the Event Detection & Index module.

3.1. The AV Tracking Module

The AV Tracking module uses both the camera and MA to track the speaker, the camera and the MA are first calibrated, then Sound Source Localization of the speaker by audio information and a kernel-based color tracking of the speaker by visual information are carried out simultaneously. A particle filter strategy that can automatically adapt the weights of audio and visual information is used to fuse the information, making the tracking result more accurate and robust. The details are presented in our previous paper [3]. Beamforming is then conducted according to the tracking result.

3.2. The SSL and Beamforming Module

Similar as above, the camera and the Microphone Array are first calibrated. When one of the three persons in the audience speaks, the MA can detect him and return an angle, which can be used to detect the speaker of the audience in video frames.

3.3. The Event Detection and Index Module

The event detection module detects important events in the meeting and indexes them in the video audio stream.

Our event detection method has two main different characters versus traditional methods.

First, rather than processing the information totally offline, a real-time strategy is adopted in our system. In our AV Tracking module for platform speaker and the SSL & Beamforming module for audience, we both use the Microphone Array to localize the sound source. The sound source locations are hard to save and process offline. On the other hand, they are very important and valuable. In the AV Tracking module, SSL can compensate for weakness in the visual tracker—easily influenced by light condition and occlusion, and provide an overall more robust tracking result. In the SSL & beamforming module for audience, SSL angle is used to detect different speakers, which is usually difficult to judge only by video information. Therefore we choose to process the information in real time. In this way, visual and audio information are fused together and compensate for each other. We also use beamforming to gather speaker's voice simultaneously.

Second, we index the meeting content with pre-defined events detected by our system. These events contain semantic information rather than constructive information. As a result, the indexes tend to be more accordant with human sense.

The events we define by now are the switch of speakers and the change of meeting topic. We detect the switch of speaker by several cues. First, the speaker on the platform may change. This change can be detected according to the information provided by the AV Tracking module. We track the speaker in the front throughout his whole activity and can get notified when he steps out of the platform. Then when a new person steps in, we can see that the speaker has changed. Furthermore, we can use face recognition methods to decide whether the speaker has changed because the AV Tracking module provides the exact picture of the speaker's face.

Second, the speaker change of the audience can also be detected. Since the positions of the audience are fixed, any dramatic change in the SSL angle of the Microphone Array can be an indication of a speaker change. If this angle change lasts for some time and a human face is found near the new SSL position, we can decide that a speaker switch has really occurred.

When the system detects a switch of speaker, it then adds indexes to the video audio stream. By recording the beginning and ending frame of the speaker change, talks of different speakers can then be distinguished.

Further analysis can be made based on the content of the whiteboard. For example, slides may be shown on the whiteboard, by analyzing the similarity between consecutive slides, we can detect the change of presentations (different color, style of slides, etc). Then we know a change of meeting topic happens，and we can divide the whole meeting into several periods with different topics.

Our system also has a good scalability. New information can be easily added to the event detection module. We are currently

integrating pen input module in our system so that writings of participants can be analyzed. New events can then be defined and detected, resulting in more abundant indexes of the meeting.

4. EXPEIMENTAL RESULT

At present our system is still under development. Some first step experimental results are given here

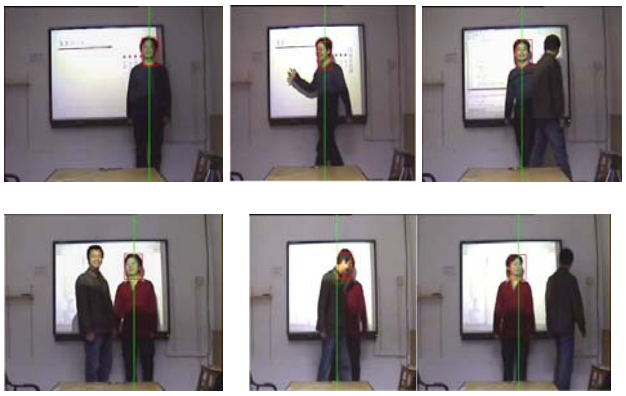

Fig.4. Results of AV Tracking

These pictures show the results of our AV Tracking module. The red rectangle represents the speaker's face detected, and the green line represents the speaker's position detected by the Microphone Array. We can see that SSL can indeed help to track the speaker, especially under conditions of light change and occlusion. And the speaker's face is obtained exactly. Based on this information, a speaker change can then be detected.

Presentation1 Presentation2 Presentation3

Fig.5. Different Presentations on Whiteboard

The above pictures show how we detect different presentations by analyzing slides shown on the whiteboard. Different presentations have slides of different colors and styles, by comparing consecutive slides, a meeting topic change can be detected.

5. CONCLUSION AND FUTURE WORK

In this paper, we presented a real-time archiving system based on audio-visual events. This system detects semantic events by fusing audio and visual information in real time. The meeting can then be archived according to these events. The real time strategy preserves some valuable information which is hard to save and process offline, and the events we define and detect contain semantic information, which is more accordant with human sense than constructive information. As a result, the archiving becomes more reasonable and accurate. Therefore people can retrieve information more quickly and conveniently. In future work, we may add other modules in the system. For example, pen input module is now being added to help obtain the writings of meeting participants. Other semantic events can then be defined and detected by analyzing these writings.

Acknowledgement: This work is supported by NSFC 60273005 project.

6. REFERENCES

[1] Ross Cutler, Yong Rui, Anoop Gupta, JJ Cadiz Ivan Tashev, Li-wei He, Alex Colburn, Zhengyou Zhang, Zicheng Liu, Steve Silverberg, **Distributed Meetings: A Meeting Capture and Broadcasting System，Proceedings of ACM Multimedia, 2002**

[2] Patrick Chiu, Ashutosh Kapuskar, Lynn Wilcox, **Meeting Capture in a Media Enriched Conference Room，Proceedings of ACM Multimedia '99. ACM, New York, pp. 149-158**

[3] Xin Li, Luo Sun, Linmi Tao, Guangyou Xu and Ying Jia**, A Speaker Tracking Algorithm Based on Audio and Visual Information Fusion Using Particle Filter，Proceedings of International Conference on Image Analysis and Recognition 2004**

[4] Michael Bett, Ralph Gross, Hua Yu, Xiaojin Zhu, Yue Pan, Jie Yang, and Alex Waibel，**Multimodal Meeting Tracker, RIAO 2000**