CARMUL : CONCURRENT AUTOMATIC RECORDING FOR MULTIMEDIA LECTURE

Y. Kameda[1]*, S. Nishiguchi*[2] *and M. Minoh*[1]

Academic Center for Computing and Media Studies, Kyoto University[1]
Graduate School of Law, Kyoto University[2]
Sakyoku, Kyoto, 606-8501, Japan

ABSTRACT

An advanced multimedia lecture recording system is presented in this paper. The purpose of our system is to capture multimodal information that can be received only when lectures are being held in a classroom where a teacher and students share the same time and space. The system captures not only hand-writings and slide switching intervals but also audio and video of the people with their spatial location information. These recorded media will be served to users in distance learning process. We designed the system so that it does not interfere classes. It can archive lectures while they hold classes on regular basis and generate multimedia archive automatically. Teachers are neither asked to remain at a certain place nor wired with devices they would have to put on. We have implemented our approach and our lecture archive system currently works for six classes per week, since October, 2002.

1. INTRODUCTION

A number of smart methods and systems such as [1] have been proposed that can provide lecture materials including audio, video, and graphical slides for distance learning. An upcoming problem therefore is how to collect media archives of attractive lectures automatically. Especially it is needed to obtain good video that represents the atmosphere of the lectures, because it can attract attention of users into their content.

We think there are two approaches to solve this problem.

One is to record lecture videos at a special studio such as broadcast station. This approach is acceptable when creators have enough time and plenty of power to prepare high-level lecture videos. However, the videos may be less attractive and sometimes very boring unless the creators can spend time and power. In such a case, it may result in a video from a fixed camera with only one teacher and no students.

The other approach is to record a real lecture that is held daily on campus. Since teachers always have their regular classes, so they usually attract students and let them learn. It is very valuable to capture the scene of the live class together with its atmosphere, as it goes. We have adopted this approach because we have many interesting classes on our campus.

Our goal in this paper is to capture most of information that can be obtained only when real lectures are being held in a classroom where a teacher and students share the same time and space. It includes not only hand-writings and graphical slide intervals but also voices and videos of the people with their 3D location information. All of these will help users to feel the atmosphere of the class and make them devote their attention for learning.

During the recording phase, we have to consider two problems: what should be captured and how it should be captured.

Ideally, the best way to archive the class with its atmosphere is to record all the spatio-temporal features in the classroom and reconstruct 3D classroom virtually at distance learning site. However, since it is almost impossible to reconstruct complete 3D shape and colors of teachers and students on-line in fine resolution, multiple videos from multiple pan-tilt-zoom cameras together with other media such as audio and teaching materials are recorded and used to convey the atmosphere instead of using virtual 3D scene reconstruction. As video can image only a part of 3D scene, the problem is how to control the pan-tilt-zoom cameras so that they extract appropriate subspace of the classroom. The system needs to estimate where people are and who is interesting from educational viewpoint at each time. Some image processing approaches have been proposed[2][3] for camera control, but image clues are sometimes subtle and may be missed.

In order not to miss information that should be recorded, we have developed a sensor fusion method on our system that uses three cues: audio analysis by microphone array, image analysis by CCD cameras, and location estimation by ultrasonic equipment.

Approaches that are designed for distance video conferencing may be similar to ours, but they usually assume that people are not moving around, while we cannot assume that teachers would stand still during their classes.

An audio and vision analysis approach[4] has been proposed which can be applied to estimate the location of speakers, but it is not designed for a noisy circumstance although their method can estimate the location of multiple speakers precisely. Our method is designed to eliminate noise.

We also pay much attention that the recording system should not require any efforts of teachers to prepare recording, and be transparent for both teachers and students during the class. It does not disturb or interfere the classes in their regular fashion during operation. The teachers are neither asked to remain at a certain place nor wired with the devices they would have to put on. They just put two small beacons on their shoulders and a wireless microphone.

Our prototype system records voice of a teacher, voices of students, 8 video streams each of which is taken by pan-tilt-zoom camera, slides with their switching intervals on the screen in the classroom, handwritings on an electric-whiteboard, trajectory of a teacher, and location of speaking students simultaneously.

The rest of the paper is organized as follows. In Section 2., we discuss information to be recorded and used in our approach. In Section 3., we describe the method of recording information that can be obtained directly from sensor outputs. In Section 4., we describe the method of recording videos with our camera control method. We explain our prototype system in Section 4. and show recorded results by our system. In the last, Section 6. concludes the paper.

2. INFORMATION PRESENTATION MODE

It is very important to carefully select what information should be recorded into lecture archive because it makes serious influence on the design of total archiving system. We first discuss lecture style which is related to the selection of information, then we describe information presentation modes that will be recorded in our approach.

We have observed what kind of lecture style is common on the campus and we selected the way of giving lectures, we assume here, as follows. The teacher stands in front of his/her students and he/she walks freely. The teacher gives a talk and may use slide-show prepared in advance, and may also use whiteboard to write down formulas. The students have their seats and can make questions, but only one person can talk at once. Two pictures in Fig. 1. show two typical lectures conducted in a classroom. On the left, a teacher uses slide-show and one whiteboard and another teacher uses two whiteboards on the right.

Information presentation mode is defined as knowledge that emerges while lectures are being held. In other words, it can only be observed during the lectures. For example, switching interval of slides is the mode because it can be measured only in the class. On the other hand, contents

Fig. 1. Snapshot of a lecture.

of the slides are not considered as information presentation mode but just knowledge because the contents themselves can be seen anytime. Table 1. shows the list of information presentation modes. Mode No.1-6 marked by level-1 are captured by direct devices whereas we need intelligent process for obtaining Mode No.7 at level-2.

Table 1. Information Presentation Modes.

No.	Description	Level
1	voice of a teacher	1
2	voice of students	1
3	switching interval of slides	1
4	handwritings on whiteboard	1
5	trajectory of a teacher	1
6	location of a speaking student	1
7	video streams with wide variation	2

3. FIRST LEVEL RECORDING

In our recording system, first six information presentation modes in Table 1. at level 1 are recorded by the following techniques. Note that each of them does not ask the teacher and students for adopting in a classroom because we believe that the recording system should neither interfere classes nor restrict their behaviors.

1: Voice of a teacher: A small wireless pin microphone is used. As it is often used in ordinary classrooms, this is not a burden to teachers.

2: Voice of students: A handy wireless microphone is used. Students must obtain a microphone before they start to talk in our current system. We are going to introduce a microphone array to get voice of the students so that they do not need to pass a handy microphone to each other in the near future.

3: Switching interval of slides: We have developed a small plug-in of Microsoft PowerPoint that records the time of changing each slide in slide-show and sends that time stamp to our recording server together with the snapshot of the slide. Teachers are simply asked to bring their PowerPoint file and upload to our PC, or bring their own PC and install the plug-in.

4: Handwritings on whiteboard: We installed electric whiteboard that can record 4 color drawings by stroke style with time stamp. Teachers can use it in the same manner of an old fashioned whiteboard. At every stroke, our process send its location, color and its time stamp to the recording server.

5: Trajectory of a teacher: Since we assume a teacher walks around in the front area of a classroom, the trajectory of the teacher is not only a clue to estimate the status of a lecture but also essential information to film him/her properly with pan-tilt-zoom camera. In order to estimate precise trajectory of the teacher, we integrate three location estimation methods. One is a computer vision method that uses multiple cameras to estimate 3D location of the teacher in wide area[5]. Another method is based on acoustic analysis. As a classroom has noise sources such as fans of AV equipment, we assume there is at most only one voice source at a certain time window and we have developed multiple Cross-power Spectrum Phase (M-CSP) analysis method[6] based on CSP methods[7][8][9]. Although these acoustic and vision methods are device free, the resolution of the 3D location estimation is not quite accurate. Therefore, we introduce ultrasonic 3D measurement system in addition to them. We use two ultrasonic beacons, size of which is $2.5 \times 2.5 \times 1.7$ cm, and the teacher just puts these small beacons on his/her shoulders. The ultrasonic equipment can estimate the location within 5 cm. As it sometimes does not work due to head occlusion or shoulder movement, we need acoustic/visual sensing together with ultrasonic one.

6: Location of a speaking student: Localization of a speaking student is also a challenging topic. A simple source localization approach does not work in this situation. So we have developed hybrid location estimation method[6] which uses both acoustic and vision analysis. It uses fish-eye camera for detecting students in a classroom and a microphone array which consists of 8 microphones to estimate planner location of sound source.

Fig. 2. shows the outlook of a teacher at our recording system. The small black cubes on his shoulders inside the circles are the ultrasonic beacons. The whiteboard in the picture is used to record hand-writings. A pin microphone is shown in center circle. Teachers can behave freely because these devices are small.

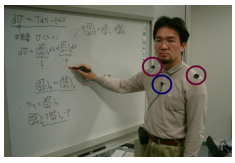


Fig. 2. A teacher with ultrasonic beacons and a microphone.

4. SECOND LEVEL RECORDING

The most difficult part of lecture recording process is to extract the subspace of a classroom that a camera films as information presentation mode of a lecture. Since the subject to be filmed changes according to the status of the lecture, the status should be estimated first. It is done by utilizing the information obtained at the first level of the recording. After the subject is determined based on the estimated status, the recording system assigns different pan, tilt, and zoom value to each pan-tilt-zoom camera so that it can retain various video streams of the same subject from different views. It is valuable to record multiple views of the same subject because we plan to let users of the lecture archive prescribe their favorite way of watching a lecture[10] so that the system can edit video stream by estimating the match score of the videos in the archive to their prescription.

We proposed an automatic filming method using multiple pan-tilt-zoom cameras in a classroom[5]. We have improved the approach so that it uses not only vision processing results but also acoustic analysis to estimate the status of the lecture more precisely and can control the cameras in more sophisticated way[1].

In our recording system, a teacher is filmed by at most four pan-tilt-zoom cameras and students by at most four other cameras. It estimates the current status of the lecture and then control all the filming cameras so that each camera assigned to the same subject produces different view of the subject, such as its size on the image, direction of the subject, and so on. The cameras are controlled on-line accordingly as the status changes.

5. PROTOTYPE SYSTEM: "CARMUL"

We have implemented the recording system named CARMUL in a classroom the size of which is about 15m by 8m. Fig. 3. shows the layout of the classroom. There are 14 pan-tilt-zoom cameras and 8 of them are used as filming cameras. We use Intersense IS600mk2 as ultrasonic 3D location estimation equipment and two beacons are set on the shoulders of the teacher (Fig. 2.).

Fig.4. shows a snapshot of browsing the archive taken by CARMUL. The video displayed on the top left is an edited video stream. The bottom left region displays drawings on an electric whiteboard, and the right region is used for displaying slides. These media are provided in SMIL format from a RealNetworks server.

Fig.5 shows a snapshot of 8 videos taken simultaneously. Although our system can record all the 8 videos to the archive, we currently store an edited video stream only due to the HDD capacity limitation.

[1]We do not explain the method in detail due to the limited space of the paper.

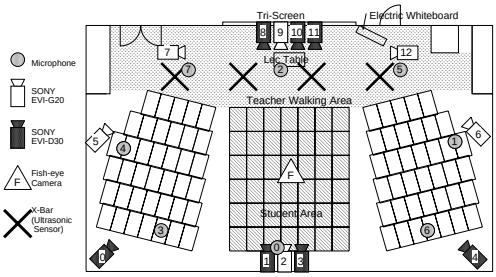


Fig. 3. Classroom layout.

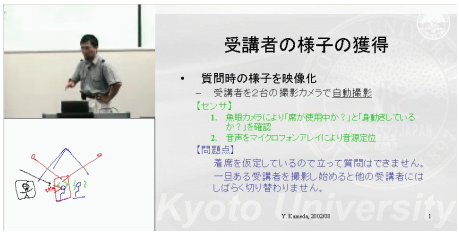


Fig. 4. An distance learning overview.

A trajectory of a teacher recorded by ultrasonic equipment in an actual class for 90 minutes is shown in Fig.6. The trajectory is used to estimate the status of the class and to control the cameras.

We conducted archive experiment on six regular courses at Kyoto University since October 2002. As each course has one class per week, we run the system six times per week for 15 weeks.

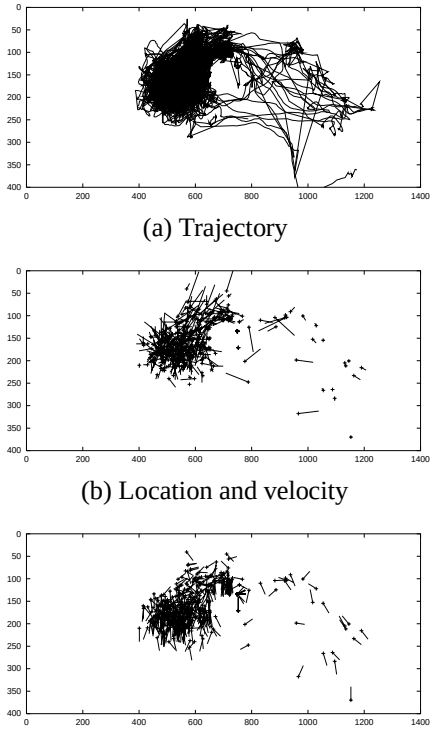
6. CONCLUSION

We have proposed a new lecture recording system named CARMUL. It can record automatically various kinds of information that can be obtained only when classes are being held. It takes multiple videos and can generate one video stream by estimating the status of the class based on vision, audio, and ultrasonic sensors.

Our advanced recording system has been working on recording six classes per week without asking teachers for special preparation effort to cooperate with our archiving project.

We are planning to extract tendency of teachers from number of their trajectories and improve camera control to film them.

Fig. 5. 8 video streams in the archive.



(a) Trajectory

(b) Location and velocity

(a) Location and direction

Fig. 6. Trajectory of a teacher [unit:cm].

7. REFERENCES

[1] A. Feinstein et al, "Teaching and learning as multimedia authoring: The classroom 2000 project," *ACM Multimedia*, pp. 187–198, 1996.

[2] M. Bianchi, "Autoauditorium: A fully automatic, multi-camera system to televise auditorium presentations," *Proc. of Joint DARPA/NIST Smart Spaces Tech. Workshop*, 1998.

[3] S. Mukhopadhyay and B. Smith, "Passive capture and structuring of lectures," *ACM Multimedia*, pp. 477–487, 1999.

[4] D. Zotkin, R. Duraiswami, H. Nanda, and L. S. Davis, "Multimodal tracking for smart videoconferencing," *ICME*, pp. 37–40, 2001.

[5] Y. Kameda, K. Ishizuka, and M. Minoh, "A live video imaging method for capturing presentation information in distance learning," *ICME*, pp. 1237–1240, 2000.

[6] S. Nishiguchi, K. Higashi, Y. Kameda, and M. Minoh, "A sensor-fusion method of detecting a speaking student," *ICME*, 2003.

[7] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. ASSP*, vol. 24, no. 4, pp. 320–327, 1976.

[8] M. Brandstein, J. Adcock, and H. Silverman, "A closed-form method for finding source locations from microphone-array time-delay estimates," *ICASSP*, pp. 3109–3022, 1995.

[9] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. SAP*, pp. 228–292, 1997.

[10] Y. Kameda, H. Miyazaki, and M. Minoh, "A live video imaging for multiple users," *ICMCS*, pp. 897–902, 1999.