Automatic Camera Control System with Both Image and Sound Processing

Suganuma, Akira
Department of Intelligent Systems, Kyushu University

Ashikawa, Taira
Department of Intelligent Systems, Kyushu University

出版情報：Proceedings of the IASTED International Conference Computers and Advanced Technology in Education. 2003, pp.722-727, 2003-06. International Association of Science and Technology for Development
バージョン：
権利関係：

Proceedings of the IASTED International Conference
COMPUTERS AND ADVANCED TECHNOLOGY IN EDUCATION
June 30 - July 2, 2003, Rhodes, Greece

AUTOMATIC CAMERA CONTROL SYSTEM WITH BOTH IMAGE AND SOUND PROCESSING

Akira Suganuma
Depertment of Intelligent Systems
Kyushu University
6–1 Kasuga-koen, Kasuga, 816–8580, Japan
email: suga@limu.is.kyushu-u.ac.jp

Taira Ashikawa
Depertment of Intelligent Systems
Kyushu University
6–1 Kasuga-koen, Kasuga, 816–8580, Japan
email: ashikawa@limu.is.kyushu-u.ac.jp

ABSTRACT
We are developing an Automatic Camera control system for Education: ACE, which captures a traditional lecture. ACE focuses on the latest object that a teacher writes on a blackboard. When this recording strategy is realized, it is necessary for ACE to detect beginning and ending of the teacher's writing. The early version of ACE mis-judges ending of his writing according to the region occluded by a teacher because it detects the region only using image processing. In this paper, we describe a method of detecting the beginning and ending with both image and sound processing. We have applied our method to recording a real lecture to validate it.

KEY WORDS
Distance lecture, Detecting a chalking sound, Image processing, Estimating teacher's writing, Recording strategy

1 Introduction

The growth of a communication network technology enables people to take part in a distant lecture. There are mainly two kinds of method to held a distant lecture. One is an web-based method, the other is a method of sending video and audio of lecture scenes. We are studying some supporting systems for a distant lecture. For the web-based method, we have designed and developed two supporting system: Computer Aided Cooperative Classroom Environment (CACCE)[1] and Automatic Exercise Generator based on the Intelligence of Students (AEGIS)[2, 3]. On the other hand, we are developing a camera control system for the distant lecture held with the second method[4, 5]. We call our system "ACE": Automatic Camera control system for Education.

The traditional style lecture delivered in a normal classroom is the target of ACE. Nowadays, a teacher teaches his students with an overhead projector and/or other visual facilities. Indeed many lectures in the information technology or the programming are frequently held by using visual facilities and/or computers in many universities but there are still many traditional-style lectures, in which a teacher explains something with a blackboard. When we deliver the lecture with the visual facilities, the progress of the lecture is accelerated. It is prone to be too fast for students to understand the detail of the teacher's explanation. On the other hand, in the lecture with a blackboard, the progress is not accelerated more than the pace of the teacher's writing. That means the progress of the lecture with a blackboard is not so fast that many students keep understanding their teacher's explanation. It seems then that such a traditional-style lecture will not completely disappear in the future although they will deliver by combining the blackboard and some visual facilities such as an overhead projector and/or Power Point software. We are consequently developing ACE.

When a lecture scene is recorded, a camera-person usually controls a camera to take suitable shots; alternatively, the camera is static and captures the same location all the time. It is not easy, however, to employ a camera-person for every occasion. If possible, he is required to be adept in recording a lecture. On the other hand, the scene captured by a steady camera hardly gives students a feeling of a live lecture. Moreover, some characters in the scene seem to be too small for students to read because the camera has to keep capturing a whole blackboard. It is necessary, consequently, to control a camera automatically. ACE enables people to do it by taking suitable shots for a distant lecture. ACE analyses a scene sent from a camera and recognizes the complexion on the lecture. ACE judges what is important in the scene and controls the camera to focus on it.

2 Automatic camera control

2.1 Distant Lecture We Envisage

We envisage that a lecture delivered in a normal classroom are recorded by a video camera and students in a distant classroom take part in the lecture by watching the scene projected on a screen. Figure 1 illustrates a form of the distant lecture by recording a traditional classroom. A teacher teaches his students in a normal classroom, which has a blackboard. The teacher writes and explains something on the blackboard. Some video cameras are setting in the room and captures a lecture scene in order that the scene is transmitted to distant classrooms. The students in the dis-

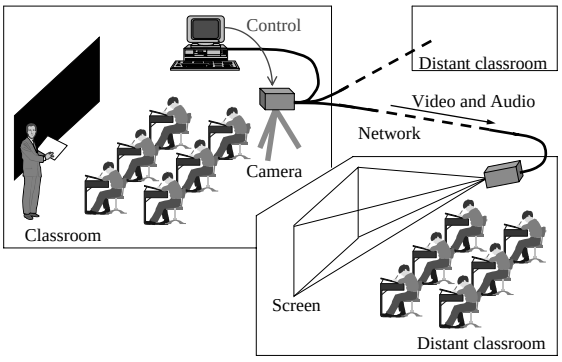
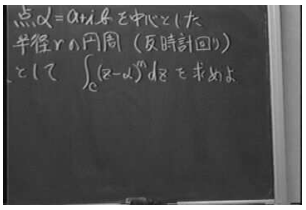
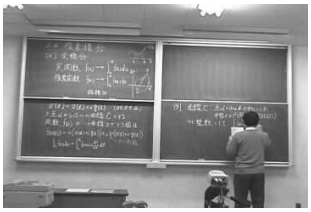


Figure 1. A form of the distant lecture by recording the normal classroom

(a) Sample scene of ordinary shot



(b) Sample scene of zooming-in shot



(c) Sample scene captured by a steady camera

Figure 2. Sample scene of a lecture

tant classroom take part in the lecture by watching a scene reflected on a screen. The teacher seems to write something anywhere on the blackboard so that the camera is required to capture all written on the blackboard. If the camera captures the whole blackboard, the character projected onto the screen is too small for students to read because the blackboard in the local classroom is usually wider than the screen in the remote classroom. This recording method is improper.

2.2 Camera Control Strategy

ACE focuses on an important thing for all students. It seems to be teacher's gesture and something written on a blackboard. A teacher usually explains something by writing contents and pointing them. The scene transmitted to distant classroom is required to contain the explained objects such as words, sentences, formulas and so on. We adopted, therefore, a strategy that ACE captured something explained by the teacher. It is very difficult, however, to extract directly what a teacher explains. Consequently, we made an assumption that the latest object written on the blackboard closely parallel something explained by a teacher because he frequently explains the latest object. Therefore, ACE captures the latest object written on the blackboard[6].

When a lecturing scene is recorded, both a constantly changing shot and an over-rendering shot are not suitable because they tire the students. A change-less shot is, if anything, more appropriate than those shots. It is very important that students can easily read contents on the blackboard. ACE usually takes a shot containing the latest object and a region near it in a discernible size like figure 2-(a). The blackboard often consists of four or six small boards like figure 2-(c). A teacher frequently writes relational objects within a small board. ACE captures, then, the blackboard by the small board in an ordinary time. If ACE does not detects any new latest objects for a long time, ACE changes a shot near the teacher. On the other hand, when once ACE detects something on the blackboard as the new latest object, ACE takes a shot which consists of the neighborhood of that, and zooms in on it after the teacher has written it. Figure 2-(b) shows a sample of the zooming-in shot. After a-few-second zooming, ACE takes an ordinary shot again.

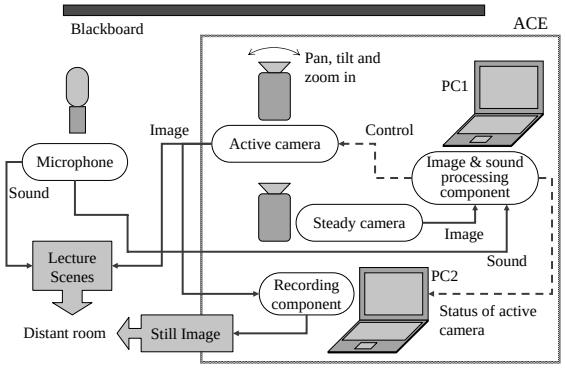


Figure 3. Overview of the recording function of ACE

2.3 Overview of ACE

Figure 3 illustrates an overview of ACE. ACE requires two cameras. One is a steady camera and the other is a active one. The steady camera captures a whole blackboard at a constant angle for image processing. The captured image is sent to the image and sound processing component of ACE running on PC1 over an IEEE-1394 protocol. And a sound picked up by the microphone is also sent to the component. The component analyses both of the image and the sound, and decide how to control the active camera according to

a camera control strategy shown in section 2.2. The control signals are sent to the active camera over an RS-232C protocol. The active camera, hereby, takes a suitable shot.

A video captured by the active camera and an audio picked up by the microphone are sent to the distant classroom. Students in the distant classroom watch and listen to them. ACE also stores some shots as a still image. The recording component running on PC2 selects adequate ordinary shots in which a teacher doesn't occlude contents on the blackboard. The students can watch a requested previous scene as a still image[7].

In our study, we are interested in how to record a lecture delivered in the normal classroom. We are using a known method or product as a way transmitting the video and audio via the network.

3 Image Processing

We developed a method capturing the latest object on a blackboard with the computer vision techniques[6]. We need a detecting method of beginning and ending of a teacher's chalking and the latest object to implement the above camera control strategy. ACE detects a rectangle circumscribed the latest object using both the background subtraction technique and the method of subtracting between successive frames. The background subtraction technique is a method to detach the foreground image from the background image. The method can detect the object on the blackboard as the foreground image. However, the teacher is also contained in the foreground image. On the other hand, the method of subtracting between successive frames can detect the moving object. Basically, ACE detects a teacher as a moving objects in the foreground image and something written on the blackboard as a steady object in the foreground image. ACE extracts the latest object on the blackboard masking out the teacher's region.

It is important for ACE to detect the ending of a teacher's chalking as exactly as possible. Because ACE must take a scene in which the teacher occlude the object if it zooms in on the written object while a teacher keeps writing. The rectangle circumscribed the latest object usually change frame by frame when a teacher is writing something. On the other hand, the rectangle stops changing after the teacher has finished to write something because he usually clears the object to make his students watch it and the rectangle does not increase. ACE judges that a teacher stopped chalking, if the detected rectangle didn't change for a particular time.

We implemented the above method detecting a teacher's chalking as a prototype of the earlier version of ACE. The method misses detecting the ending teacher's chalking when the teacher walks near the latest object and talks something after he wrote. In this case, the rectangle circumscribed the latest object detected by this prototype keeps changing while the teacher walks. ACE then cannot detect the ending of the teacher's changing. The method also falsely detects the ending when the teacher stands still and writes something near his body. In this case, the rectangle has no change because a new object near the teacher's body is masked out and is not detected by our image processing method. These cases often occur in a lecture. We consequently have studied the detecting method of a teacher's chalking using sound processing.

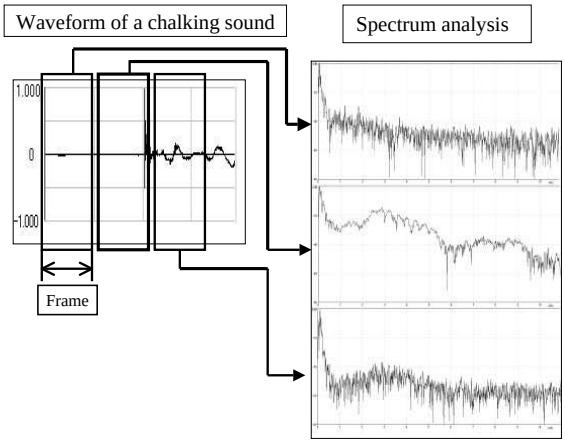


Figure 4. Spectrum analysis of a chalking sound

4 Sound Processing

4.1 Detecting a Chalking Sound

A chalking sound is a short sound made when a chalk hits a blackboard. It is a discontinuous sound and has a short interval between successive two chalking sounds. Any Raps except a chalking sound are seldom made in a classroom.

Figure 4 shows a spectrum analysis of a chalking sound. The left graph shows the waveform of a chalking sound and the others show the Discrete Fourier Transform(DFT) of each frame of the chalking sound. The DFT grows in almost all frequency bands when a chalking sound is made and it tumbles when the sound finishes. This means that the DFT varies dramatically at the beginning and ending of a chalking sound. The upsurge occurs in a very short time.

We modified the Spectrum Flux(SF)[8] to detect a chalking sound more accurately and developed the Band Pass Spectrum Flux(BPSF). SF is defined as the average variation value of spectrum between the adjacent two frames. SF becomes large when the short upsurge of sound spectrum occurred. It is then useful for detecting a rap.

BPSF is defined by the following formula:

$$BPSF = \frac{\sum_{n=1}^{N-1} \sum_{k=K_1}^{K_2} \left[\log \frac{(A(n,k)+\delta)}{(A(n-1,k)+\delta)} \right]^2}{(N-1)(K_2-K_1)}$$

where $A(n,k)$ is the Discrete Fourier Transform of the n-th frame, k-th spectrum of input signal. δ is a very small

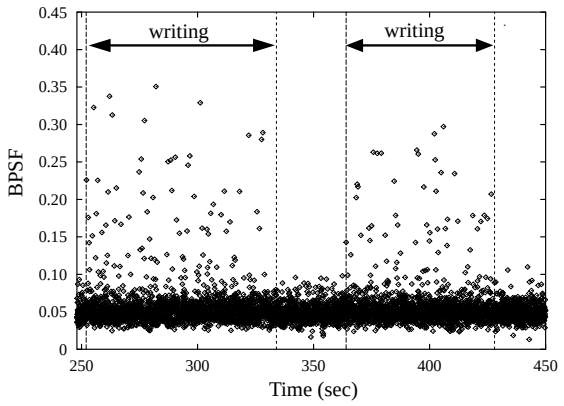


Figure 5. BPSF when a teacher writes something on a blackboard

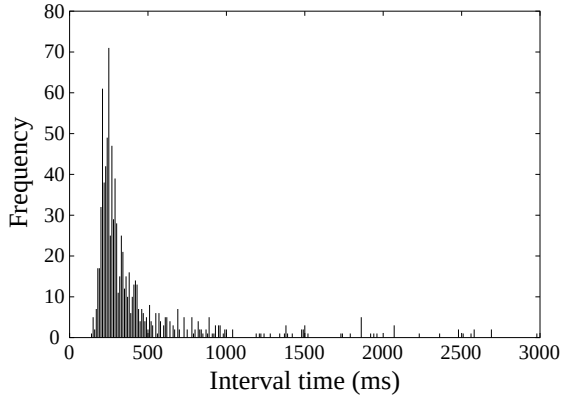


Figure 6. Frequency distribution of interval time between successive two chalking sounds

value to avoid calculation overflow and N denotes the total number of frames. K_1 and K_2 are the lower bound and upper bound orders of DFT respectively. BPSF narrows the spectrum down to the bandwidth between 1 kHz and 4 kHz which often reflects the characteristics of the chalking sound. This is hard to cause the increase of BPSF according to any sounds else.

Figure 5 shows BPSF of sound in a particular lecture. It was calculated with the condition that the frame length was 10ms and the total number of frames N was 4. A teacher was writing in the two segments from 252 to 334 second and from 364 to 428 second. There are some points of a large BPSF value in these periods and there are no points of a large one in other periods. We can find a threshold to detect a chalking sound.

4.2 Method detecting teacher's writing

The following is our basic method detecting a teacher's writing by sound processing.

- If β chalking sounds are detected within α seconds, it judges that a teacher starts writing something.
- If any chalking sounds are not detected for γ seconds, it judges that a teacher ends writing something.

The parameters, α, β, and γ seems to be individual characteristics. We investigated intervals between successive two chalking sounds made when a particular teacher wrote something. The result of our investigate is plotted in figure 6. The mode of the intervals is 250ms, the average of the intervals is 380ms, and the maximum of the intervals is 2960ms. We assumed, therefore, that these parameters, α, β, and γ were 1.0, 3, and 3.0 respectively. It takes more than γ second to detect the ending the teacher's writing only by the above method. The direct use of this detecting method is undesirable for ACE.

5 Estimating method of a teacher's writing by both image and sound processing

As mentioned in section 2.2, ACE focuses on the latest object written on the blackboard. A detection of beginning and ending a teacher's writing is necessary to implement ACE. The method which estimates the latest object as soon as possible is also essential. Our image processing mentioned in section 3 extracts the latest object on the blackboard. It sometimes fails, however, to extract the latest object if the teacher occludes the latest object when he is writing or after he wrote. We have designed a new extracting method for the latest object with both image processing and sound processing. ACE judges more accurately whether a teacher is writing or not. And ACE can extract the latest objects more quickly.

5.1 Probability that a teacher is writing something

When a chalking sound is made, our method may guess that a teacher is writing something. When no chalking sound is made, it may guess that he is doing another thing. However, that is not always true. If no chalking sound is made for a short time after one chalking sound was made, this situation may have to be guessed that he is writing something, because the chalking sound is a sudden, short sound. The probability that a teacher is writing something becomes less if the short interval gets longer.

We defined the probability P_{chalk} as table 1 based on the method mentioned in section 4.2. If a chalking sound was detected, P_{chalk} becomes $\frac{1}{\beta}$. When once the sound was detected, the probability is kept at the same value for m seconds and decreases steadily after that. On the other hand, it increases with each chalking sound was detected within α seconds. The probability becomes 1 if β chalking sounds are detected within α seconds. After the probability was 1, if no chalking sound is detected, the probability

Table 1. How to calculate P_{chalk}

```
Flag←FALSE;
x=0;
while in-lecture do begin
  if BPSF ≥ TH then x ← x + 1;
  if Flag=FALSE then begin
    if (x > 0) or (t − t_chalk < m) then
      P_chalk ← x/β;
    else P_chalk ← x/β − t∗x/((α−m)∗β);
    if P_chalk ≤ 0 then x ← 0;
    if P_chalk ≥ 1 then Flag←TRUE
  end
  else begin
    if (BPSF ≥ TH) or (t − t_chalk ≤ m) then
      P_chalk ← 1;
    else P_chalk ← 1 − t/(γ−m);
    if P_chalk ≤ 0 then begin
      Flag←FALSE;
      x ← 0
    end
  end
end
```

decreases steadily and becomes 0 γ seconds later. We assumed the parameter, α, β, γ, and m, were 1.0, 3, 3.0, and 0.25 respectively based on our investigation mentioned section 4.2.

5.2 Extraction of the new latest object on the blackboard

When the probability P_{chalk} is greater than a particular threshold TH_c, ACE judges that a teacher began writing something. ACE doesn't zoom in, however, on the teacher's writing area as soon as P_{chalk} gets grater than TH_c because he may occlude something written on the blackboard or because some students may still need the previous latest object. After a new object written on the blackboard was detected near the teacher, ACE decides zooming in on it at brief intervals.

While ACE judges that a teacher is writing something, it keeps checking that all of the new latest object is in the region captured by the active camera, because the region of the latest object is growing. If a part of the object protrude beyond the region, ACE zooms out a little and/or pans/tilts the camera.

On the other hand, when the probability P_{chalk} is less than the threshold TH_c, ACE begins searching the latest object on the blackboard by the image processing. If ACE extracts the latest object, ACE judges that a teacher finished writing something and zooms in on the extracted latest object.

Table 2. Results of the questionnaire

No.	Score	Steady	ACE	Old ACE
(1)	Average	2.60	3.25	3.19
	5: Excellent	0.0%	2.1%	
	4: Good	23.4%	39.6%	
	3: Satisfactory	31.4%	41.7%	
	2: Unsatisfactory	25.5%	14.6%	
	1: Poor	19.1%	2.1%	
(2)	Average	1.45	3.73	2.75
	5: Excellent	0.0%	12.5%	
	4: Good	2.1%	56.3%	
	3: Satisfactory	2.1%	25.0%	
	2: Unsatisfactory	34.0%	4.2%	
	1: Poor	61.7%	2.1%	
(3)	Average	1.53	3.38	2.49
	5: Excellent	0.0%	2.1%	
	4: Good	0.0%	45.8%	
	3: Satisfactory	10.6%	39.6%	
	2: Unsatisfactory	31.9%	12.5%	
	1: Poor	57.4%	0.0%	
(4)	Average	2.47	3.06	2.86
	5: Excellent	0.0%	0.0%	
	4: Good	8.5%	29.2%	
	3: Satisfactory	40.4%	50.0%	
	2: Unsatisfactory	40.4%	18.8%	
	1: Poor	10.6%	2.1%	
(5)	Average	1.81	3.17	2.49
	5: Excellent	0.0%	2.1%	
	4: Good	0.0%	27.1%	
	3: Satisfactory	8.5%	56.3%	
	2: Unsatisfactory	63.8%	14.6%	
	1: Poor	27.7%	0.0%	

6 Applying ACE to a real lecture

We have developed ACE and done an experiment of applying ACE to a real lecture. We delivered two 25-minutes lectures for 50 undergraduates. A teacher who was one of the authors taught them by only using a blackboard. Although the teacher knows the detecting algorithm of ACE, he taught by usual style. Sort of thing, he behaved as disadvantage for ACE. We took the lecture scene with ACE and with a steady camera in order to compare these shots. In our experiment, these shots were not transmitted over the network but were recorded on video cassettes and played in the classroom with VCR.

After each video lecture, we had the students fill in a questionnaire which consists of following five questions; (1) Could you watch the teacher's action well? (2) Could you watch the objects on the blackboard well? (3) Could you watch the object you wanted? (4) Were you given a feeling of the live lecture? (5) Could you give the scene a overall score as a lecture one? They scored each scene from 1 to 5. The distribution of the scores of each question

is shown in Table 2. In this table, the scores in "Steady" column are ones of the scene captured by the steady camera, the scores in "ACE" column are ones of the scene captured by ACE and the scores in "Old ACE" column are ones of the scene captured by the earlier version of ACE which analysis the lecture scene by only image processing. Although we investigated the score in the last column before[6], we gave a column to the score of the earlier version of ACE for comparison.

The score of the scene captured by ACE is clearly better than that of the scene captured by the steady camera. Some students evaluated ACE very highly except the fourth question. ACE selectively captures the object on the blackboard so that it is important to select the object timely. We satisfied that ACE can capture the adequate object timely because all averages of the question was greater than 3(satisfactory). And less than 20% students said unfavorable evaluation in all questions. On the other hand, although the scene captured by the steady camera included all objects on the blackboard, the score of the scene was low because the character in the scene was too small for the students to read.

We applied the t-test to compare the scores of the steady camera and ACE. Our null hypothesis is "*The scene captured by ACE is as good as the scene captured by the steady camera.*" This null hypothesis is rejected with 1% level of significance in all questions. In all questions, the score of ACE is better than that of the steady camera. The evaluation of ACE is, therefore, superior to that of the steady camera. The shots captured by ACE is, consequently, good enough to record a lecture.

The score of the scene captured by ACE is also better than that of the scene captured by the earlier version of ACE. We applied the t-test with these scores. Our second null hypothesis is "*The sound processing does not make ACE better.*" This hypothesis is rejected with 1% level of significance in the second question, the third question and the last question.

We asked the students the sixth question: Did you understand the content of this video lecture rather than that of the normal lecture? 62.5% students evaluated that they understood the contents of the video lecture as good as that of the normal lecture. Moreover, 20.8% students evaluated that they understood the contents better than that of the normal lecture.

7 Conclusion

We have designed a camera control strategy for recording a lecture and developed a prototype of ACE. We have implemented our system using both image processing and sound processing. The scene captured by the new version of ACE is more suitable than that captured by old one. We evaluated, moreover, ACE with applying it to a real lecture. As a consequence, we make sure that ACE is a useful tool for recording a traditional lecture.

ACE takes a suitable shot if the teacher explains the object as soon as he writes on the board. It cannot take, however, a suitable shot when he explains something written before. He usually teaches his students pointing the objects which he wants them to look at. Interpreting teacher's action and/or posture, or recognizing teacher's voice, ACE could capture more suitable scene. We will get ACE interpret and recognize it. We assume that a teacher teaches his students with a blackboard. But some teachers sometimes also use with an overhead projector. We will also make ACE be applied to such a situation.

References

[1] A. Suganuma, R. Fujimoto, and Y. Tsutsumi, An WWW-based Supporting System Realizing Cooperative Environment for Classroom Teaching, *Proc. World Conference on the WWW and Internet*, 2000, 830–831.

[2] T. Mine, A. Suganuma, and T. Shoudai, The Design and Implementation of Automatic Exercise Generator with Tagged Documents based on the Intelligence of Students: AEGIS, *Proc. International Conference on Computers in Education*, 2000, 651–658.

[3] A. Suganuma, T. Mine, and T. Shoudai, Automatic Generating Appropriate Exercises Based on Dynamic Evaluating both Students' and Questions' Levels, *Proc. World Conference on Educational Multimedia, Hypermedia & Telecommunications*, 2002, 1898–1903.

[4] A. Suganuma, S. Kuranari, N. Tsuruta, and R. Taniguchi, An Automatic Camera System for Distant Lecturing System, *Proc. Conference on Image Processing and Its Applications*, Vol.2, 1997, 566–570.

[5] A. Suganuma, S. Kuranari, N. Tsuruta, and R. Taniguchi, Examination of an Automatic Camera Control System for Lecturing Scenes with CV Techniques, *Proc. Korea-Japan Joint Workshop on Computer Vision*, 1997, 172–177.

[6] A. Suganuma and S. Nishigori, Automatic Camera Control System for a Distant Lecture with Videoing a Normal Classroom, *Proc. World Conference on Educational Multimedia, Hypermedia & Telecommunications*, 2002, 1892–1897.

[7] A. Suganuma, Development of an Automatic Camera Control System for Videoing a Normal Classroom to Realize a Distant Lecture, *Proc. International Conference on Information Technology & Applications*, 2002, CD-ROM.

[8] L. Lu, H. Jiang and H. J. Zhang, A Robust Audio Classification and Segmentation Method, *Microsoft-Research-TR-2001-79*, 2001.