Camera Control for a Distant Lecture Based on Estimation of a Teacher's Behavior

Atsushi Shimada[1], Akira Suganuma[1], and Rin-ichiro Taniguchi[1]

[1] Department of Intelligent Systems, Kyushu University
6-1 Kasuga-koen, Kasuga, 816-8580, Japan

Abstract. The growth of communication network technology enables people to take part in a distant lecture. We are developing a supporting system for a distant lecture named ACE (Automatic Camera control system for Education). The previous version of ACE captures the lecture focusing on the latest object written on a blackboard because a teacher frequently explains it. However, a teacher not always explains the latest object. We have designed, therefore, ACE to take a suitable shot according to a teacher's behavior. This paper describes our system, our camera control strategy, the algorithm to estimate a teacher's behavior in a lecture and experiment of our system.

1 Introduction

The growth of communication network technology enables people to take part in distant lectures. When a lecture scene is captured, a camera-person usually controls a camera to take suitable shots. (Alternatively, the camera is static and captures the same location all the time.) However, it is not easy to employ a camera-person for every occasion. However the scene captured by a fixed camera hardly gives us a feeling of the live lecture. It is necessary to control a camera automatically. We are developing a supporting system for a distant lecture. We call it "ACE" (Automatic Camera control system for Education).

There are some methods to support distant lectures. For example, Kameda et al. [1][2] and Onishi et al. [3][4] have been studying some supporting systems. They use multiple cameras and switch them according to the situation of the lecture. Their both systems track a teacher and capture him/her mainly. It is important to capture a teacher, but it is more important for students to see objects (a character, a sentence, a figure, a table, or so) explained by him/her.

ACE captures important scenes in a lecture. Both objects written by a teacher on a blackboard and an area that he/she is explaining are important. The previous version of ACE[5] [6] mainly captured the objects written on the blackboard because a teacher frequently explains them. It, however, could not take a suitable shot if a teacher explained an object written before. Moreover, the video captured by the previous version of ACE is fatiguing to student's eyes because the video scene is often changed owing to zoom in, zoom out, pan-tilt and so on. We have designed, therefore, ACE to take an

Shimada A., Suganuma A. and Taniguchi R. (2004).
Camera Control for a Distant Lecture Based on Estimation of a Teacher's Behavior.
In *Proceedings of the 4th International Workshop on Pattern Recognition in Information Systems*, pages 15-24
DOI: 10.5220/0002677600150024

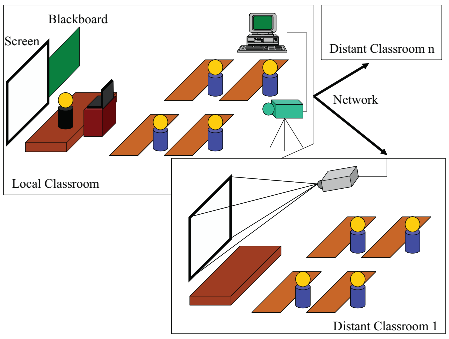

Fig. 1. Lecture style assumed in ACE

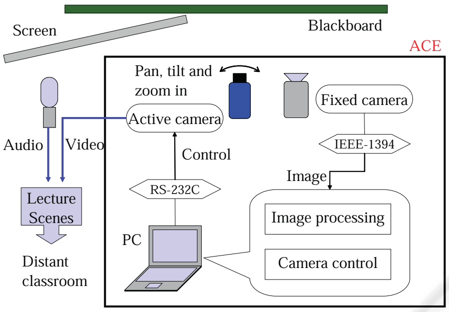

Fig. 2. Framework of ACE

area explained by a teacher. ACE need to estimate the teacher's behavior for this design. When the teacher is writing an object on a blackboard, ACE captures the teacher and the object mainly. When he/she is explaining to his/her students, ACE captures the latest object and, if necessary, the objects written on the blackboard before. In this study, we contrived the estimation method of the teacher's behavior and made ACE find the target area to capture according to his/her behavior.

In this paper, section 2 presents design of ACE and our strategy of camera control. Section 4 describes the algorithm to estimate a teacher's behavior in a lecture, and section 5 describes an experiment that we applied ACE to video lectures. Finally, concluding remarks are given in section 6.

2 Overview of ACE

2.1 Design

A style of the distant lecture which we envisage is illustrated in Fig.1. A teacher teaches his/her students in a local classroom, and students in remote classrooms take part in the lecture by watching the video captured in the local classroom. ACE supports the lecture in which a teacher teaches his/her students by using both a blackboard and a screen.

2.2 Framework of ACE

Fig.2 shows the framework of ACE. ACE needs two cameras. One is a fixed camera, which captures whole lecture scene for image processing. The other is an active camera to capture a suitable shot, and its video is transmitted to remote classrooms. The image captured by the fixed camera is sent to PC (in Fig.2) over an IEEE-1394. The PC analyzes the image and controls the active camera via an RS-232C. The video and the audio picked up by a microphone are sent to remote classrooms via the network using DVTS (Digital Video Transport System)[8].

2.3 Camera Control Strategy

What does ACE capture? One solution for this problem is to take the scene that students want to watch, but many scenes are probably requested by many students at the same time. Although this solution needs the consensus of all students, it is very difficult to make it. We have decided, therefore, that ACE captures the most important thing from a teacher's point of view. When we designed the previous version of ACE, we assumed that the most important thing is the latest object written on the blackboard. The previous version of ACE took a shot zoomed in on the object after the teacher had written it on the blackboard. After a-few-second zooming, the previous version of ACE zoomed out and take a shot containing the latest object and a region near it. However, this strategy has some problems. The previous version of ACE cannot take a suitable shot when the teacher explains object written before. The lecture scene captured by the previous version of ACE changes at short intervals because the latest object is often found when he/she goes on writing objects on the blackboard for a long periods of time. Such a video is not appropriate for students who take part in the distant lecture.

We have adopted, therefore, the strategy that ACE captures an object explained by a teacher. When the teacher is writing object on a blackboard, ACE captures the teacher and the object. When he/she is explaining to his/her students, ACE captures the latest object. If he/she is explaining the objects written on the blackboard before, ACE also captures them. On the other hand, the image processing component on the PC checks whether the next target which ACE should capture is included in the current capturing area or not. If the next target is included, ACE need not move the active camera. This solves the problem that the lecture scene changes at short intervals. When he/she is explaining the objects on the screen, ACE captures whole of the screen even if the objects are anywhere on the screen.

3 Teacher's Behavior Model

3.1 Teacher's Behavior in a Lecture

We observed some lecture videos (Table 1) to find out what behavior a teacher is doing in his/her lecture. We found out that the behavior of the teacher could be categorized into three kinds: “Writing”, “Explaining” and “Moving”. When the teacher was writing some objects on the blackboard, we categorized his/her behavior as “Writing”. When the teacher was explaining objects on the blackboard or on the screen, we categorized his/her behavior as “Explaining”. We categorized the other kind of behavior as “Moving”. Table 1 shows the ratio of each behavior in the lectures.

3.2 Creating Teacher's Behavior Model

We got the position of the teacher's centroid $\boldsymbol{g}(t) = (g_x(t), g_y(t))$, face $\boldsymbol{f}(t) = (f_x(t), f_y(t))$ and hand $\boldsymbol{h}(t) = (h_x(t), h_y(t))$ from the Video A in Table 1 by the hand work in 2 fps. The positions may be represented as a time series $\boldsymbol{I}(0), \boldsymbol{I}(1), \cdots, \boldsymbol{I}(T)$, where $\boldsymbol{I}(t)$ denotes the position of the teacher's centroid, face, and hand at time t.

$$\boldsymbol{I}(t) = (g_x(t), g_y(t), f_x(t), f_y(t), h_x(t), h_y(t)) \tag{1}$$

Table 1. The lecture videos we observed and the ratio of each behavior

	Teacher	Length (min)	Lecture Style	Writing	Explaining	Moving
Video A	A	40	Use Only Blackboard	47.9%	44.6%	7.5%
Video B	B	60	Use Only Blackboard	36.8%	54.3%	8.9%
Video C	C	25	Use Blackboard and Screen	10.8%	85.9%	3.3%
Video D	D	60	Use Blackboard and Screen	13.2%	81.7%	5.1%

We got 1,463 data for the behavior “Writing”, 1,248 data for “Explaining” and 857 data for “Moving”. We made 5-dimensional feature vector $\boldsymbol{v} = (v_1(t), v_2(t), v_3(t), v_4(t), v_5(t))^T$ by using $\boldsymbol{I}(t)$ and $\boldsymbol{I}(t-1)$. The feature vector has following five elements.

$$v_1(t) = |f_x(t) - f_x(t-1)| \tag{2}$$

$$v_2(t) = |f_y(t) - h_y(t)| \tag{3}$$

$$v_3(t) = \sqrt{(f_x(t) - h_x(t))^2 + (f_y(t) - h_y(t))^2} \tag{4}$$

$$v_4(t) = |g_x(t) - h_x(t)| \tag{5}$$

$$v_5(t) = |g_y(t) - h_y(t)| \tag{6}$$

The v_1 is the horizontal movement of the teacher’s face. The v_2 is the vertical difference between his/her face and hand. When a teacher writes an object on the blackboard, the position of his/her face changes a little. In addition, his/her hand is close to his/her face. According to above features, we chose v_1 and v_2 as the elements of the feature vector. When a teacher explains to his/her students, he/she use his/her body to physically express words. Therefore, the positional relationship among his/her centroid, face and hands are very important. The distance between his/her face and hand (v_3), the lateral difference (v_4) and the vertical difference (v_5) between the centroid of his/her body and hand are picked out for such occasions.

Fig. 3 shows a scatter plot of 300 samples projected onto the (v_1, v_2) subspace. The circles in Fig. 3 show “Writing”, the squares show “Explaining”, and the triangles show “Moving”. We can regard the feature vectors as the distribution of the points over the vector space. We use the Gaussian mixture model in order to approximate the distribution. The Gaussian mixture modeling approximates a probability density function by a weighted sum of multivariate Gaussian densities[7]. We applied EM-algorithm to estimate the parameter set of the Gaussian mixture model. We got three stochastic models (“Writing”, “Explaining” and “Moving”) in 5-dimensional vector space.

4 Processing of ACE

In this section, we describe the processing of ACE. ACE’s process consists of three steps: estimating of the teacher’s behavior, looking for the area explained by him/her, and finding a target area to take a suitable shot.

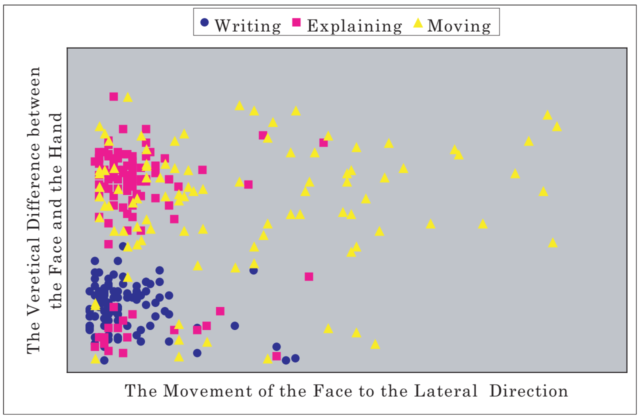

Fig. 3. The plot of samples projected onto (v_1, v_2) subspace

4.1 Estimation Teacher's Behavior

ACE analyzes the image captured by the fixed camera to estimate the teacher's behavior. ACE detects the teacher's region at the first step. Next, it extracts features to estimate his/her behavior. Finally, it estimates his/her behavior.

Detecting Teacher's Region
ACE has to segment a teacher. We use a background subtraction technique to detect objects in the image. The background image is captured before opening the lecture. The image contains the screen and the blackboard on which no object was written. After subtracting the background from the image captured by the fixed camera during the lecture, ACE can get a foreground image. It consists of objects written on the blackboard, the teacher and so on. We would like to detect only the teacher. We apply, therefore, the erosion to the foreground image. We use a 5×5 mask because the objects or the noise in the image are thinner and smaller than the teacher's body. After the erosion, our system makes the histogram of all highlight pixels in the filtered image because some noises are still remained in the image. ACE extracts the teacher's region from the histogram by setting an appropriate threshold. A sample of the teacher's region is shown in Fig. 4.

Extraction of Features
After detection of the teacher's region, ACE extracts feature points $\boldsymbol{I}(\mathrm{t})$. ACE acquires the centroid of the teacher as the center of the highlight pixels in the teacher's region. The teacher's face and hands are detected by extracting skin color pixels in the teacher's region. We used HSV color space to extract skin color pixels. However, the teacher's hand is often hidden by his/her own body, so our system sometimes detects his/her both hands and sometimes detects only one hand or no hand. Our system categorizes, therefore, the skin color area into at most three clusters, because the skin color area consists of three parts, by applying the k-means clustering method.

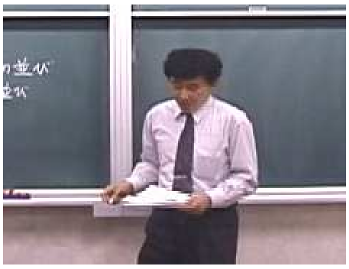

Fig. 4. The teacher's region

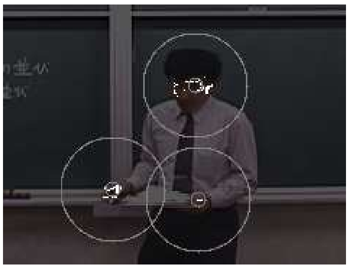

Fig. 5. The result of clustering

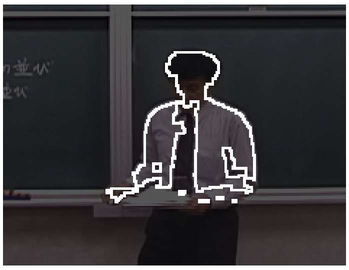

Fig. 6. The edge pixels in the teacher's region

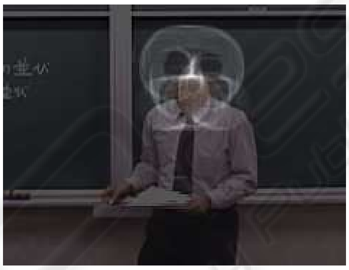

Fig. 7. The Hough transform

Fig. 5 is the result of clustering. The white colored circles show the clusters. ACE needs to categorize the clusters into his/her face and hands. We use the Hough transform to detect a face because the shape of the face can be approximated as a circle. Fig. 6 shows the edge pixels in the teacher's region. ACE votes the central point of circle, which has an edge pixel on the circumference in upper half of the teacher's region. Fig. 7 shows the result of voting. The white dots in the figure are the pixels which were voted as the central point of circle. Generally, the Hough transform takes much time. However, ACE can the image processing in 2 fps because it applies the Hough transform in the very narrow range. ACE extracts the pixel which was voted at the greatest numbers as the center of the teacher's face. The face position, which was detected by the Hough transform, is compared with each cluster. ACE categorizes the nearest cluster as a facial area. The other clusters are regarded as teacher's hands.

Estimation

ACE calculate the feature vectors by using the feature points which were gotten in subsection 4.1. ACE puts the feature vector into the Gaussian mixtures which indicate each behavior model, and acquire probability of each behavior. The behavior whose probability is highest is regarded as the teacher's behavior.

4.2 Looking for Explained Area

A teacher explains to his/her students the latest object written on the blackboard or the object written before. ACE acquires the position of the object written on the blackboard by applying the background subtraction and noise reduction to the image sequence except for teacher's region. A teacher generally tends to explain newer object then older one. ACE labels the objects to record the latest object. ACE refers to labels and decides the area explained by him/her.

When a teacher explains to his/her students the object written before, he/she moves near the object, or points his/her finger at the object. In the former case, the explained area is near the teacher's region. In the latter case, ACE calculate the vector from the centroid of the teacher to his/her hand. If the magnitude of the vector is larger than the threshold which we set, ACE assumes he/she is pointing /his/her finger at the object. If there is an object written on the blackboard on the straight line which is stretched from the vector, the object and a region near it are explained area.

4.3 Finding Target

ACE finds a target area according to a parameter set $\theta = \{p_1, p_2, p_3, p_4, p_5\}$ described in Table 2.

Table 2. The parameter for the camera control

p_1	the teacher's behavior
p_2	the position of the objects written on the blackboard
p_3	the positon of the teacher
p_4	the area explained by the teacher
p_5	the time when ACE controled the zoom rate of the active camera at last@

ACE changes the zoom rate of the active camera according to the teacher's behavior (p_1). If the interval between the times when ACE changed the zoom rate is short, it doesn't change the zoom rate. In the case of changing the zoom rate, ACE records the time (p_5).

When the teacher's behavior is "Writing", ACE zoom in and takes a shot with a focus on the latest object and the teacher. ACE calculates the target area by using p_2 and p_3. When it is "Explaining", ACE zooms out a little and takes a shot with a focus on the objects explained by him/her. p_4 is used to decide the target area.

5 Experiment

5.1 Preparation

We evaluated our image processing by using lecture videos which had been captured previously. We prepared the following lecture videos in Table 3.

Table 3. The detail of the lecture video

	Teacher	Lecture Style	Length (min)
Video 1	A	Use Blackboard Only	20
Video 2	C	Use Blackboard Only	19
Video 3	C	Use Blackboard and Screen	25

The Video 1 is the lecture video which the same teacher appears on the Video A (in Table 1). The teacher on the Video 2, 3 is another one. We apply the image processing of ACE to the Video 1, 2, 3. First, we investigated whether the position of the teacher's face and hands extracted by ACE is correct or not. Next, we evaluated the estimation result of teacher's behavior. Finally, we verified how accurately the ACE find the target to capture the lecture scene. ACE calculates the target area from the parameters in Table 2, and display the rectangle on the video window as the target area.

5.2 Result of Experiment

How accurately did ACE extracted the feature points?
Table 4 shows the precision ratio and the recall ratio. Precision ratio means how much correct the positions which ACE extracted are, is defined by formula (7). ACE sometimes extracts two hands although only one hand is visible and vice versa. In such a case, we regarded ACE could not extract the positions correctly. The position of the teacher's face was detected correctly in over 90% of frames, and the position of the hand was detected correctly in nearly 90% of frames.

Recall ratio means how much ACE extracted the position when we wanted ACE to extract it in the teacher's region, is defined by formula (8). The position of the teacher's face and hand were caught in over 90% of frames. The experimental results show that the position of the teacher's face and hand are probably detected correctly.

$$Precision = \frac{\textit{number of correct positions}}{\textit{total number of positions extracted by ACE}} \tag{7}$$

$$Recall = \frac{\textit{number of correct positions}}{\textit{total number of positions we wanted ACE to extract}} \tag{8}$$

Result of estimation of teacher's behavior
Table 5 shows the result of estimation of teacher's behavior. The behaviors of both "Writing" and "Explaining" are estimated in over 70%. However, ACE could not estimate the behavior of "Moving" at a high ratio. This is because the behaviors of "Moving" and "Explaining" are similar in the image sequence, we couldn't distinguish two behaviors without the sounds.

The ratio of estimation for the Video 1 is higher than the other videos. This is because the teacher in the Video 1 is same in the Video A which was used to generate each model of behavior. Despite of different teacher, the results for the Video 2 and Video 3 are not very low compared to the result of the Video 1.

Table 4. The Accuracy of the Position of the Teacher's Face and Hands Estimated by Our System

		Precision (%)	Recall (%)
Video 1	Face	94.3	96.0
	Hand	86.2	90.6
Video 2	Face	92.5	98.2
	Hand	89.7	93.1
Video 3	Face	93.9	97.5
	Hand	87.1	90.5

Table 5. The Result of estimation of teacher's behavior for each video

	Writing		Explaining		Moving	
	Precision	Recall	Precision	Recall	Precision	Recall
Video 1	81.5	79.2	78.3	76.8	20.5	30.1
Video 2	76.8	79.9	75.7	77.8	17.9	11.6
Video 3	77.9	72.4	80.4	81.7	16.5	13.2

How correctly the ACE find the target to capture the lecture scene?
We show how correctly the ACE find the target to capture the lecture scene in Table 6. If the area which ACE should focus on at the frame is included in the target area which ACE found at each frame, we regard the target as appropriate area.

ACE could find the appropriate target at each lecture video even though ACE couldn't estimate teacher's behavior of "Moving" at a high ratio (Table 5). We guess that the behavior of "Moving" is less important than that of "Writing" and "Explaining". The ratio of the behavior "Moving" is also very low campared with the others "Writing", "Explaining" (Table 1). We have designed ACE's camera work strategy with a focus on the area explained by the teacher. The area is closely related to the teacher's behavior of "Writing" and "Explaining".

We think that the area which ACE should focus on is consequently included in the area which ACE found since ACE could find the target area by using the other parameters $p_2 \sim p_5$ even if the result of estimation of the teacher's behavior is wrong at a frame. In addition, because ACE didn't change the target area if the estimation result changes extemporaneously, it could find the target area successfully.

6 Conclusion

We have designed and developed ACE which is a supporting system for a distant lecture. ACE estimates teacher's behavior and controls the active camera to take a suitable shot. We have evaluated ACE with applying it to video lectures. We consequently make sure that ACE can take a suitable shot for the most part even in a real lecture.

ACE estimate teacher's behavior ("Writing", "Explaining" or "Moving"), and find the target to capture. It cannot distinguish, however, between "Explaining" and "Moving" proficiently because it is difficult to distinguish even if we see the video without

Table 6. How correctly the ACE find the target to capture the lecture scene?

	Success ratio (%)
Video 1	90.7
Video 2	94.3
Video 3	92.1

sounds. Using the sound information of the teacher, ACE could distinguish their two behaviors and capture more suitable scene. We will make ACE interpret the teacher's voice.

References

1. H.Miyazaki, Y.Kameda, M.Minoh: A Real-time Method of Generating Lecture Video for Multiple Users Using Multiple Cameras. IEICE, J82-D-II, No.10, pp.1684–1692, 1999
2. Y.Kameda, K.Ishizuka, and M.Minoh: A Real-time Image Method for Distant Learning Based on Dynamic Situation Understanding. CVIM, Vol.2000, No.121-11, pp.81–88, 2000
3. M.Onishi, M.Izumi, and K.Fukunaga: Automatic Production of Video Images for Distance Learning System Based on Distributed Information. IEICE, J82-D-II, No.10, pp.1590–1597, 1999
4. M.Onishi, M.Murakami, and K.Fukunaga: Computer-Controlled Camera Work at Lecture Scene Considering Situation Understanding and Evaluation of Video Imaages. CVIM, Vol.2001, No.125-5, pp39–46, 2001
5. A.Suganuma, S.Kuranari, N.Tsuruta, and R.Taniguchi: An Automatic Camera System for Distant Lecturing System. Conference on Image Processing and Its Applications, Vol.2, pp.566–570, 1997.
6. A.Suganuma and S.Nishigori: Automatic Camera Control System for a Distant Lecture with Videoing a Normal Classroom. World Conference on Educational Multimedia, Hypermedia & Telecommunications, pp.1892–1897, 2002.
7. N.Johnson and D.Hogg: Representation and synthesis of behavior using Gaussian mixtures. Image and Vision Computing 20, pp.889–894, (2002)
8. http://www.sfc.wide.ad.jp/DVTS/index.html