Shooting the Lecture Scene Using Computer-controlled Cameras Based on Situation Understanding and Evaluation of Video Images

Masaki Onishi
Bio-Mimetic Control Research Center
RIKEN
Moriyama-ku, Nagoya, 463-0003 Japan
onishi@bmc.riken.jp

Kunio Fukunaga
Graduate School of Engineering
Osaka Prefecture University
Sakai-shi, Osaka, 599-8531 Japan
fukunaga@cs.osakafu-u.ac.jp

Abstract

In this paper, we propose a computer-controlled camera work that shoots object scenes to model the professional cameramen's work and selects the best image among plural video images as a switcher. We apply this system to a shooting of a lecture scene. In the first, our system estimates a teacher's action based on features of a teacher and a blackboard. In the next, each camera is directed to a shooting area based on the teacher's action, automatically. In the last, this system selects the best image among plural images under the evaluation rule. Moreover, we have tried experiments of shooting lecture scene and have confirmed the effectiveness of our approach.

1. Introduction

Recent years, distant learning systems have come into wide use in the universities, companies and so on. In general, the video image of distant learning is selected out of images from plural camera positions by cameramen and a director. Many operators such as professional cameramen and directors, however, are necessary in this method. Under these circumstances, there is strong request for the system which produces the vivid lecture scene automatically [1, 2].

In this paper, we propose a method of producing the most suitable lecture video image automatically by controlling camera pan/tilt/zoom and switching from plural camera images. In the first, to understand the situation of lecture, we estimate a teacher's action based on teacher's features and blackboard's features extracted by image understanding. In the next, each camera is directed to a shooting area based on the teacher's action. In the last, the most suitable lecture video image is selected out of the images from plural positions under the heuristic evaluation rules. And we confirm effectiveness of our approach by an experiment.

2. Extraction of events appeared on lecture

In the case of shooting a cooking show[3], sports and the other scene which have no scenario in advance, we need to extract a trigger for decision of a shooting area. If we design the camera work on the assumption that computer-controlled camera shoots the lecture scene, teacher's actions are important triggers. In order to check the triggers, we try to recognize the teacher's action using both the character on the blackboard and features of teacher's motion extracted from the fixed camera's image.

2.1. Separation of teacher and characters

There are important two major objects in lecture video images. One is the teacher and the other is characters on blackboard written by the teacher. (Here, we call 'letters and figures' 'characters'.) To extract them from video images, we use a method based on edge detection from spatiotemporal images of a video image sequence [4]. The method is based on edge directions from several cross sections of the spatiotemporal image, in order to divide the edges appeared on spatiotemporal images into two kinds of edges, edges caused by moving objects (we call *dynamic edges*.) and edges of stationary objects (*static edges*).

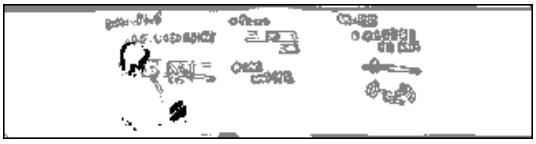
Figure 1 shows an input image (a) and *a static and dynamic edge* image (b). It is possible to assure that characters on blackboard are extracted as *static edges* (gray level) and a teacher is extracted as *dynamic edges* (black level) appeared on the lecture video images taken by a fixed camera.

2.2. Extraction of teacher's feature

In the lecture video, suppose that the image region of the teacher is comprised in the *dynamic edges*. To recognize the teacher's action, our system estimates the teacher's position, face direction and hands position using the *dynamic edges* and colors of pixels.

0-7695-2128-2/04 $20.00 (C) 2004 IEEE

(a) Input image.



(b) *Static edge* (gray) and *dynamic edge* (black).

(c) Extraction of teacher's feature.

(d) Extraction of skin color.

(e) Extraction of written region on the blackboard.

Figure 1. Result of situation understanding.

To estimate the teacher's head position, our system analyzes the *dynamic edges*. Under the assumption that the shape of the head is an ellipse, the teacher's head is extracted from a region of the elliptic distribution of the *dynamic edges*. Here the ratio of a minor axis to the major axis of the ellipse is 1 to 1.2. Figure 1 (c) shows the extraction result of the teacher's head.

Our system extracts the teacher's skin color to estimate the teacher's face direction and hands position. Using the extraction results of the teacher's head, our system estimates the teacher's head direction. In our approach, we use the CIE1976UCS color space that offers more precision in color measurement comparing with the conventional RGB color space. In advance using the man power we extract the N pixels which compose the face and hands under the various light conditions. These pixels are projected CIE1976UCS color space $\boldsymbol{C_i} = (u_i, v_i)$ $(i = 1, 2, \cdots, N)$, and our system calculates the mean value $\boldsymbol{\mu_S}$ and the covariance matrix $\boldsymbol{\Sigma_S}$ of the color parameters. Using $\boldsymbol{\mu_S}$ and $\boldsymbol{\Sigma_S}$, our system calculates the degree of a color similarity between input pixels $\boldsymbol{x}$ and the pixels of skin model using following equation,

$$P_S(\boldsymbol{C_x}) = \frac{1}{2\pi\,|\boldsymbol{\Sigma_S}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{C_x} - \boldsymbol{\mu_S})\boldsymbol{\Sigma_S}^{-1}(\boldsymbol{C_x} - \boldsymbol{\mu_S})^T\right\}. \quad (1)$$

Our system extracts the skin color using the suitable threshold and equation (1). Figure 1 (d) shows the result of skin color extraction. Dark pixels show the pixels of skin color of the face and hand region.

Next, using the distribution of skin pixels in the head region (estimated by ellipse), the face direction is estimated approximately. Our approach tries to guess roughly the face direction. Under this condition, we decides the face direction to be one of the four directions, "front", "back", "left" and "right", by examining the number of skin color pixels appeared on the left side or the right side of the head region.

Using the head region and the skin region, the hands position is estimated. In general, the skin region extracted from the lecture video images expresses teacher's face and hands. The remainder of skin color pixels after removing the face region results in the hands region. The clusters of these pixels are decided to be the region of the teacher's hands. Using k-means algorithm, our system extracts two largest clusters which represents the both hands. In Figure 1 (c), we extract only one hand, because the centers of two clusters are close to each other.

2.3. Extraction of blackboard's feature

To extract teacher's action completely, it is not enough to extract only the regions of face and hands as the teacher's features. For example, the actions such as "write on blackboard" or "Erase blackboard" could not decide only using above features, also using the face direction toward blackboard and moving of the hands. Here we introduce the change of number of characters appeared on blackboard. In the case, when the teacher "writes on blackboard", the number of character on the blackboard increases and the case when the teacher "erases blackboard", the character on the blackboard decreases. So these actions are discriminated by using increase or decrease of the characters on the blackboard. In the lecture scene, characters on the blackboard are given by the *static edges*, so amount of characters is supposed to be proportional to the number of *static edges*.

To recognize the action "erase blackboard" accurately, the position of the blackboard eraser is estimated in the same way as hands position using the color value of eraser. If the system could not find it at the rail of the blackboard, the teacher uses it to "erase the blackboard".

Table 1. Rule of action estimation.

event(teacher's action)	conditions				
	POSI	FACE	HAND	CHAR	ERAS
1. "Write on blackboard"	stop	back	move	increase	∗
2. "Erase blackboard"	stop	back	move	decrease	absence
3. "Explain using blackboard"	stop	except front	move	no change	∗
4. "Explain to students"	stop	front	move	no change	∗
5. "Move to left"	left	∗	∗	no change	∗
6. "Move to right"	right	∗	∗	no change	∗

(∗ : don't care)

2.4. Recognition of teacher's action

We recognize the teacher's action based on his/her movements and blackboards features. Using the features from the current image and previous image, our system estimates the change of teacher's position (POSI), face direction (FACE), change of teacher's hands position (HAND), amount of characters on the blackboard (CHAR) and presence or absence of the eraser (ERAS). Moreover, the teacher's action is estimated using the if-then rule for POSI, FACE, HAND, CHAR and ERAS. Table 1 shows the if-then rule to recognize the action under our experience.

If teacher's position does not change, direction of the face look toward the blackboard, teacher moves the hands and the amount of the characters on the blackboard increase, the teacher "write on blackboard".

3. Decision of shooting area based on situation understanding

In general, the change of the view direction of students depends on the teacher's action. For example, if the teacher "explains using blackboard", many students look at the teacher and characters which the teacher points out. At this time not only characters which teacher points out but also written region around the pointed characters are focused by many students. In this case, shooting area of the lecture scene is not only pointed characters but also a block area which denotes the letters and figures of explaining its. We call this block area a *written block*. The *written block* is extracted by using a method for segmentation of written region on a blackboard [4]. Figure 1 (e) shows extracted *written block* surrounded by black lines. We regard a *written block* near the teacher's hands to be a used *written block*, and we can decide the shooting area to take a *written block* in the case of "write on blackboard" and "explain using blackboard" into consideration.

In this paper, we make an automatic shooting rule in the lecture scene supposing the objects many students pay attention to. In this shooting rule, events (teacher's action) trigger requests of shooting area. Table 2 shows our shooting rule. For example, if the teacher "writes on blackboard", shooting area sets to be teacher and *written block* used by teacher. In the case when the teacher "erases blackboard", shooting area set to be whole blackboard, because there is no particular area to pay attention to.

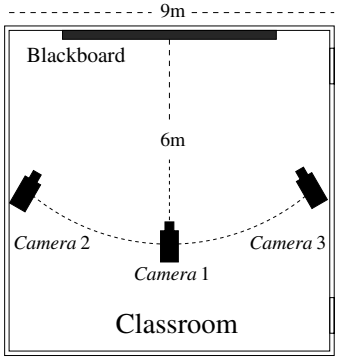


Figure 2. Layout of three cameras.

4. Switching based on evaluation

In this section, let us consider a switching rule. Our system can shoot lecture videos from plural positions, students watch ,however, only one monitor. Therefore, it is necessary to select the best image among plural images. In general, a switcher selects the best images among plural camera by comparison preview monitors. Here, we emulate the following general switching rule.

For example, in the case of taking a lecture video, a switcher selects the best camera position considering the characters on the blackboard and direction of teacher's eyes. This importance is changed by the teacher's action. If the teacher "writes on the blackboard" and "explains using blackboard", the best camera angle has a good view of the characters on the blackboard. If the teacher "explains to students", the best camera direction is the direction to the line of teacher's sight.

Table 2. Rule of camera work and evaluation.

event	request of shooting area	method of evaluation
1. "Write on blackboard"	teacher and used *written block*	have a better view of the written region on a blackboard
2. "Erase blackboard"	whole blackboard	a front of a blackboard (*Camera1*)
3. "Explain using blackboard"	teacher and used *written block*	same as 1
4. "Explain to students"	teacher (center)	meet the line of teacher's sight
5. "Move to left"	teacher (make left space)	near to the teacher
6. "Move to right"	teacher (make right space)	same as 5

On the assumption of producing a lecture video, we compose the rule of evaluation to select the best image. The right hand side of Table 2 shows the rule of evaluation. The best image is selected from among three images by using this rule. In our experiments, placement of three cameras in the classroom is given by Figure 2, and the evaluation value is set to be from 0 to 1.

5. Experimental results

Using our method, we produced a lecture video image by controlling three cameras and switching. The resolution of input images is 320×80 pixels as Figure 1 (a). Figure 3 shows some results of lecture video image. These images are shot by *Camera 1*, *Camera 2* and *Camera 3* from the top row to the bottom row. And from left to right, shows scenes of "Write on blackboard", "Explain to students" and "Erase blackboard" respectively. The value written at the bottom of each image, shows the evaluation value. The evaluation value means the degree of best video image. The images surrounded by a thick line have the largest evaluation valuation at the same time. The students take lessons using the bottom selected lecture images produced by our method. We asked eight students to push a button for evaluating the satisfaction of the lecture video. 97.6% of total time of the lecture video (about 20 min) taken by our method satisfies the majority of eight students who participated as the subjects.

6. Conclusions

In this paper, we proposed a method of shooting the lecture scene using computer-controlled cameras. To understand the situation of the lecture, the movement of the teacher and the conditions of the blackboard are extracted in the first. Then the system estimates the action of the teacher on the basis of these conditions. In the last, our system sets the best camera angle and selects the best video images. According to the experiments, our approach shows quite useful to produce lecture videos without the cameramen.

Acnowledgments: We are grateful to Mr. Masashi Murakami for helping experiments.

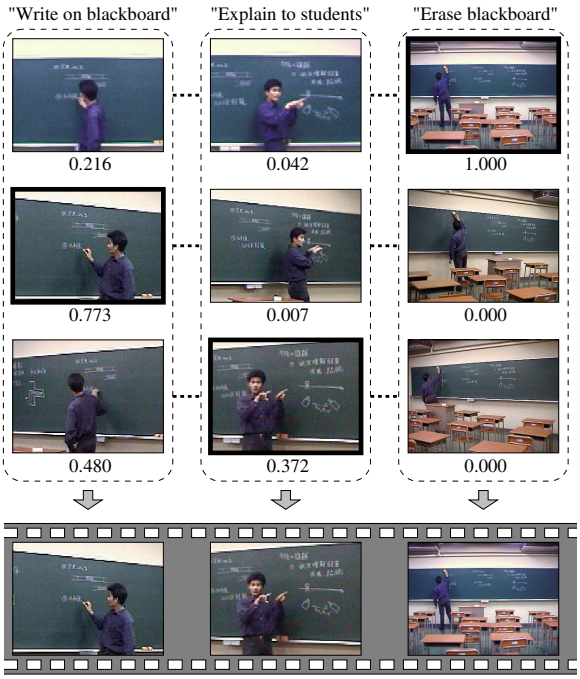


Figure 3. Example of generated lecture image.

References

[1] M. Minoh and Y. Kameda, "Imaging A 3D Lecture Room by Interpreting Its Dynamic Situation," *Proc. of Int. Workshop on Cooperative Distributed Vision*, pp.371–412, Mar. 2001.

[2] M. Onishi, M. Izumi and K. Fukunaga, "Production of Video Images by Computer Controlled Camera Operation Based on Distribution of Spatiotemporal Mutual Information," *Proc. 15th ICPR*, vol. 4, pp.102–105, Sep. 2000.

[3] C.S.Pinhanez and A.F.Bobick, "Approximate world models: Incorporating qualitative and linguistic information into vision systems, " *Proc. AAAI '96*, pp. 1116–1123, Aug. 1996.

[4] M. Onishi, M. Izumi and K. Fukunaga, "Blackboard Segmentation Using Video Image of Lecture and Its Applications," *Proc. 15th ICPR*, vol. 4, pp.615–618, Sep. 2000.