Development of a Supporting System for a Distant Lecture Using Visual Facilities

Shimada, Atsushi
Department of Intelligent Systems, Kyushu University

Suganuma, Akira
Department of Intelligent Systems, Kyushu University

Taniguchi, Rin-ichiro
Department of Intelligent Systems, Kyushu University

出版情報：Proceedings of the International Conference on Information Science and Electrical Engineering 2003, pp.581-584, 2003-11
バージョン：
権利関係：

Proceedings of the International Symposium on Information
Science and Electrical Engineering 2003 (ISEE 2003)
ACROS Fukuoka, Fukuoka, Japan, November 13-14, 2003

DEVELOPMENT OF A SUPPORTING SYSTEM FOR A DISTANT LECTURE USING VISUAL FACILITIES

Atsushi Shimada, Akira Suganuma, and Rin-ichiro Taniguchi

Department of Intelligent Systems, Kyushu University

ABSTRACT

The growth of a communication network technology enables people to take part in a distant lecture. In recent years, a lecture that a teacher uses both a screen and a blackboard has been increasing as compared with a traditional one using a blackboard. We are developing a supporting system for a distant lecture, in which a teacher teaches his/her students with both a blackboard and a screen, which enables the teacher to hold a distant lecture without changing his/her usual style.

1. INTRODUCTION

The growth of a communication network technology enables people to take part in a distant lecture. In recent years, a lecture that a teacher uses both a screen and a blackboard is increasing as compared with a traditional one using a blackboard. In case a teacher gives a lecture using a screen, what is projected on the screen is an image of the material that the teacher precomposed by using a presentation tool such as Power Point software. When a lecture scene is videoed, 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. In addition, the scene videoed by a steady camera hardly gives us a feeling of the live lecture. Therefore, we are developing a supporting system for a distant lecture, in which a teacher teaches his/her students with both of a blackboard and a screen, which enables the teacher to hold a distant lecture without changing his/her usual style.

Our system mainly captures the area that a teacher is explaining. When the teacher writes something on a blackboard, he explains often the latest object written on the blackboard. The system needs to recognize the teacher's behaviors such as writing, explaining and moving in order to capture appropriate areas. Our system can determine the area and control a camera automatically according to the result of recognizing the teacher's behavior by image processing. On the other hand, when the teacher uses both a projector and a screen, the teacher mainly explains something projected on the screen. The teacher usually lectures pointing out an explained part using either a physical pointer or a laser pointer. Our system gets the slide image as a still image from the teacher's computer directly and transmits and project it as it were a paper picture show. Our system also recognizes the place pointed by the teacher on the slide image. The system projects the still image received in the remote classroom onto a screen and shows a mark on it according to the result of recognizing the teacher's pointing position.

In this paper, we describe the design of our system for a distant lecture using both a blackboard and a screen. We also describe image processing technique to have used for our system to recognize something explained by a teacher.

2. OVERVIEW

Our system consists of two subsystems. One is the system which works in case a teacher uses a screen and the other is the system which works in case he/she uses a blackboard. In this section, we describe the detail of each subsystem.

2.1. Design of the Screen Side

A teacher projects the display image of his/her computer on a screen. We assumed the following things at the design of our subsystem which works on the screen side.

- A teacher uses a laser pointer or a physical pointer when doing an explanation.
- A teacher is not reflected in the image transmitted to the remote classroom.

The first assumption makes our system easy to extract a teacher's explaining part on a screen by the image processing. It is often difficult even for students in the local classroom to find out the explaining part when the teacher explains something on the screen with no pointer. Then almost all teachers probably point the explaining part by a kind of pointer.

The second assumption means that the image that transmitted to the remote classrooms by this subsystem is same

as one projected onto the screen in the local classroom Although an animation can be built in a slide of a teaching material, almost all images projected by a presentation tool are still. We decided that the video image needed not be transmitted in such a case. Although this subsystem captures a lecture scene on a screen with a video camera, the scene is only used for the image processing and is not directly transmitted to the remote classrooms.

If transmitting the video image captured by a video camera through the network, the system requires the very high-speed network because we are going to transmit a lecture scene which the other subsystem described in section 2.2. If transmitting a still image sequence as it were a paper picture show, the system decreases the quantity of the transmitted data.

2.2. Design of the Blackboard Side

A teacher teaches his/her students by using a blackboard. The teacher sometimes uses the blackboard mainly, and sometimes uses it for supplementary explanation of the contents which are projected on a screen. According to the following assumptions, we designed the other subsystem which works on the blackboard side.

- Students are not reflected in the scenes captured by the camera.
- A teacher is not required to give the system a special cue.

The first assumption is made to decrease processing costs. If students are reflected in the scenes, our system has to distinguish a teacher and them. This processing is complex and takes much time. It is easy to satisfy this assumption if we take a scene from the ceiling.

The second assumption is very important for a teacher. If a teacher gives the system his/her special cue such as to press a button of a remote controller, the system may control a camera more easily. If the teacher, furthermore, put on a special cloth, on which some color markers are attached, it is easier to detect his/her position and/or action. The special cue and the special cloth, however, increase the load on the teacher. He/She may omit to give the system his/her cue. He/She ought to concentrate his/her attention on his/her explaining. We decided, consequently, we did not require him/her to give the system his/her cue.

2.3. System Overview

The overview of our system is shown in Fig. 1. Our system consists of three cameras and four computers. Two of the cameras are steady ones and the other is an active one. Two steady cameras are for the image processing.

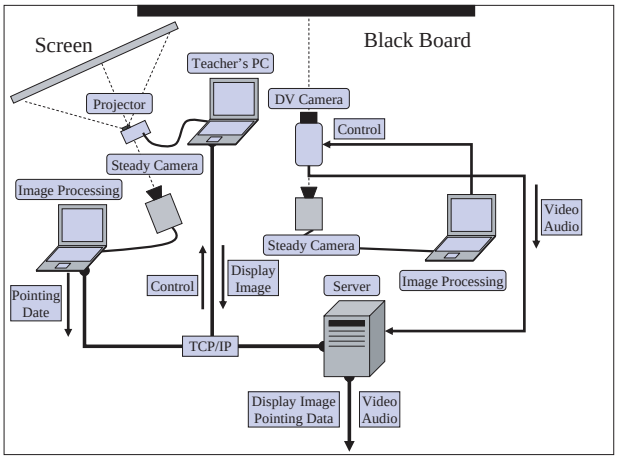


Fig. 1. An Overview of Our System

A subsystem capturing a shot of the blackboard side requires the steady camera, which captures a whole blackboard at a constant angle. The image captured by the camera is sent to PC over an IEEE-1394. This subsystem analyzes the image and decides what is the worth capturing. The subsystem also requires the active one, which is panned and tilted to capture a suitable shot. This subsystem controls the active one over an RS-232C. The visual and the audio are sent to the server PC by using DVTS (Digital Video Transport System) via the network.

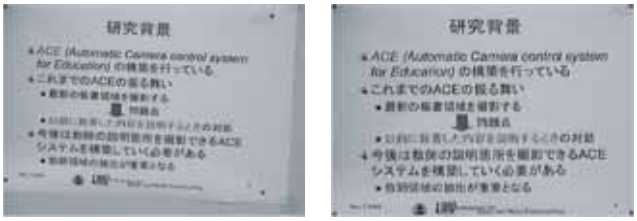
On the other hand, the other subsystem requires a steady camera which captures a whole screen. This subsystem recognizes something from the video image with the computer vision techniques and the results are sent to the server PC connected over a TCP/IP.

The server PC gets the projected image from the process on the teacher's PC. What the server PC sends to the remote classroom are the video, the audio and the image which is obtained from the teacher's PC.

2.4. Camera Control Strategy

What does our system capture? It is very important thing for the system. One solution is to take the scenes that students want to watch, but in this case, 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 developed ACE (Automatic Camera control system for Education) [1] to take lecture scenes automatically and efficiently with computer vision techniques. ACE deals with a lecture which uses only a blackboard. We assumed an object which is worth capturing is the latest one written by the teacher because he/she frequently explains the latest one. ACE mainly captures, hereby, the latest ob-



(a) Before Calibration (b) After Calibration

Fig. 2. Calibration of a Distorted Image

ject written on the blackboard as something explained by the teacher. However, the assumption is not perhaps satisfied when our system captures the new-style lecture with both a blackboard and a screen.

It is very important that the system categorizes the teacher's behavior. If the system categorizes the teacher's behavior into three categories such as "Writing", "Explaining" and "Moving", the system can capture the lecture scene according to his/her behavior. When the teacher is writing, the system should capture nearby the teacher. When the teacher is explaining, the system should capture the contents.

3. EXTRACTING THE EXPLAINED PART ON THE SCREEN

3.1. Calibration of a Distorted Image

An image projected onto a screen is often distorted because the projector is hard to set plumb in the face of the screen. Although an undistorted image can be projected, when a camera captures the image, it may be distorted because the camera is also hard to set plumb in the face of the screen. Our system has to transform the distorted image into the undistorted one by the calibration technique at the first process to take compatible in the coordinates of the image projected on the screen and the undistorted actual slide image. Using the principle of the perspective transformation, the coordinates in the calibrated image is calculated with the one in the image from camera by the following equation:

$$(x, y) = \left(\frac{p_1 X + p_2 Y + p_3}{p_7 X + p_8 Y + p_9}, \frac{p_4 X + p_5 Y + p_6}{p_7 X + p_8 Y + p_9} \right)$$
$$\vec{p} = (p_1, \cdots, p_9)^T, \quad |\vec{p}| = 1$$

where (X, Y) is the coordinates of the camera image and (x, y) is the coordinates of the transformed image which is undistorted. When looking for perspective transformation parameter $\vec{p}$, we can calibrate a distorted image[2]. If a user has clicked the four corners of the slide image in the captured image, our system computes the parameter $\vec{p}$ from the coordinates of these four points. Fig. 2 shows a projected image before and after calibration. We can confirm that the distorted image is transformed into undistorted one.

3.2. Extraction of the Explained Part

Our system extracts the laser pointer or the tip of the physical pointer by the image processing to estimate the part of the image on the screen which the teacher is explaining. Our system detects the position of the red laser pointer or the tip of the physical pointer using color processing. These colors are shifted by both the background color of the slide and the illumination condition. Especially, the color of the laser pointer is sensitive to the environment. So a range of intensity of RGB signal has to be wide to detect one in any environment. If the range expands widely, however, some pixels in something else are detected falsely. Our system detects, therefore, the position with the method combined with the background subtraction technique and the color processing. Our system applies the color processing to the foreground image extracted by subtracting the background image, which is updated every when the image projected onto the screen switched over. Our system can more accurately detect the position of the pointer with this method because it can drop red pixels which are belong to the projected image.

4. CATEGORIZATION OF THE TEACHER'S BEHAVIOR

4.1. Detecting Teacher's Region

First, the system has to segment the teacher to categorize his/her behavior in a frame. We use a background subtraction technique to detect objects in the image. The background image is captured before opening the lecture. The image contains only the blackboard on which written no object. After subtracting the background from the image captured by the same camera during the lecture, our system can get some foreground objects. They consist of something to write 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. The erosion is the filtering technique where it passes a mask over the image and puts the smallest value on the mask in the center pixel. The larger the mask size becomes, the more the center pixel on the mask is influenced by the pixels on the mask. We use a 5×5 mask because the object or the noise in the image is thinner and smaller than the teacher's body. After the erosion, our system makes the region circumscribed all highlight pixels loosely and considers it the teacher's region.

4.2. Find the Teacher's Face and Hands

After detection of the teacher's region, we need to extract features. However, the features gotten from teacher's region are limited. We decided, therefore, to find teacher's face and hands. 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.

The teacher's face and hands are detected by searching skin color in the teacher's region. The teacher's face and hands need to be distinguished. Our system categorizes, therefore, the skin color area into at most three clusters by applying the k-means clustering method. On the other hand, the system also uses the Hough transform to detect a circle because the shape of the head is approximately a circular form. The head position detected by circle the Hough transform is compared with each cluster and the nearest cluster is categorized as a facial area. Our system resolves the other clusters are hands.

4.3. Extraction Result of Teacher's Face and Hands

To ascertain the propriety of the technique mentioned above, we examined whether the position of the teacher's face and hands extracted by our system is correct or not in about a 20-minute lecture.

Table 1. The Accuracy of the Position of the Teacher's Face and Hands Estimated by Our System

	Precision (%)	Recall (%)
Face	92.46	98.20
Hand	89.70	93.05

Table 1 shows the precision ratio and the recall ratio. Precision ratio means how much correct the positions which our system extracted are. Our system sometimes extracts two hands although only one hand is visible and vice versa. In such a case, we regarded our system is not extracting 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 our system extracted the position when we wanted our system to extract it in the teacher's region. The position of the teacher's face and hand were caught in over 90% of frames. The experimental result shows that the position of the teacher's face and hand are probably detected correctly.

4.4. The Spread of the Future

We will categorize the teacher's behavior by the stochastic approach. We manually extracted the position of the teacher's face and hands at each frame from the actual lecture video. We calculated the feature vector by using the positions, the displacements and so on according to the teacher's behavior ("Writing", "Explaining" and "Moving") at each frame. We can regard the feature vectors as the distribution of the points over the vector space. We will use the Gaussian mixture in order to approximate the distribution. The Gaussian mixture modeling approximates a probability density function by a weighted sum of multivariate Gaussian densities[3]. After approximating the distribution, we will calculate the probability of each behavior ("Writing", "Explaining" and "Moving") by using the features gotten in the teacher's region at a current frame. We will make our system output the highest probability as the categorization result of the teacher's behavior.

5. CONCLUSION AND FUTURE WORKS

We described our system which supports a distant lecture, in which a teacher teaches his/her students with both a blackboard and a screen. Our system transmits the contents on the screen as the still images. The video and the audio captured by an active camera which is videoing on the blackboard side is also transmitted to the remote classroom via the network.

Our camera control strategy is to capture the contents explained by the teacher. In order to realize our control strategy, we decided to categorize the teacher's behavior. We tried to extract the position of his/her face and hand. The teacher's behavior is output as the probability by the Gaussian mixture. We will calculate the probability and make our system control the active camera according to the teacher's behavior. We will also apply our system to the actual lecture and evaluate our system.

6. ACKNOWLEDGMENTS

This research was partly supported by the 21st Century COE Program 'Reconstruction of Social Infrastructure Related to Information Science and Electrical Engineering' and a Grant-in-Aid for Scientific Research (Category C) from Japan Society for the Promotion of Science(JSPS) No.14580224, 2002–2004.

7. REFERENCES

[1] Akira Suganuma and Shuichiro Nishigori, "Automatic Camera Control System for a Distant Lecture with Videoing a Normal Classroom", Proc. World Conference on Educational Multimedia, Hypermedia & Telecommunications, pp.1892–1897, 2002.

[2] Rahul Sukthankar, "Smarter Presentations: Exploiting Homography in Camera-Projector Systems", 8th International Conference on Computer Vision, 2001.

[3] Neil Johnson and David Hogg, "Representation and synthesis of behavior using Gaussian mixtures", Image and Vision Computing 20, pp.889–894, 2002