GENERATING A TIME SHRUNK LECTURE VIDEO BY EVENT DETECTION

Takao Yokoi, Hironobu Fujiyoshi

Department of Computer Science, Chubu University
Email: {taka, hf}@vision.cs.chubu.ac.jp

ABSTRACT

Streaming a lecture video via the Internet is important for E-learning. We have developed a system that generates a lecture video using virtual camerawork based on shooting techniques of broadcast cameramen. However, viewing a full-length video takes time for students. In this paper, we propose a method for generating a time shrunk lecture video using event detection. We detect two kinds of events: a speech period and a chalkboard writing period. A speech period is detected by voice activity detection with LPC cepstrum and classified into speech or non-speech using Mahalanobis distance. To detect chalkboard writing periods, we use a graph cuts technique to segment a precise region of interests such as an instructor. By deleting content-free periods, i.e, period without the events of speech and writing, and fast-forwarding writing periods, our method can generate a time shrunk lecture video automatically. The resulting generated video is about 20%∼30% shorter than the original video in time. This is almost the same as the results of manual editing by a human operator.

1. INTRODUCTION

Recently, E-learning such as Web Based Training (WBT) has become a popular method used in higher education. In particular, archiving the lecture by videotaping and broadcasting the archived lecture video through the Internet can help remote learning. However, video recording by a cameraman and video editing of the archived lectures takes a long time and costs a great deal.

To solve this problem, we have developed a system that generates dynamic lecture video using virtual camerawork from the high resolution images recorded by a HDV camcorder [1]. This system generates a lecture video by cropping from the high resolution image to track the region of interest (ROI) such as the instructor. Since the system uses virtual camerawork that is modeled on actual broadcast cameramen's techniques [1], the generated video is very similar to video shot by a broadcast cameraman. However, viewing a full-length lecture video takes time for students. Therefore, it is necessary to generate a time shrunk video without losing any of the contents of the lecture.

Various approaches [3, 4] for video skimming have been proposed. Camera motion analysis is used to detect scene changes by characterizing the flow throughout the entire image. Since these approaches are aimed at general video and TV programs, they can not be applied to skimming a lecture video. Ishizuka et al. [5] reported a method for indexing based on state prediction of the lecture by the detecting instructor's position. To generate a time shrunk video, indexing of a lecture video is not sufficient. We have to detect content-free periods of the lecture by analyzing audio and images.

In this paper, we propose a method for generating a time shrunk lecture video using event detection. We detect two kinds of events: a speech period and a chalkboard writing period. A speech period is detected by voice activity detection with LPC cepstrum and classified into speech or non-speech using Mahalanobis distance. To detect a chalkboard writing period, we use a graph cuts technique to segment precise ROI such as an instructor in order to detect a change of characters on the chalkboard. By deleting content-free periods and fast-forwarding writing periods, our method can generate a time shrunk lecture video automatically.

2. GENERATING LECTURE VIDEO USING VIRTUAL CAMERAWORK

A HDV (1080*i*) camcorder is located at the back of the classroom to videotape images with high resolution (1,400 × 810 pixels), which contain the whole area of the chalkboard, so that students can read the handwritten characters on the chalkboard. However, it is impossible to display the high resolution image on the small screen of a general notebook PC (XGA).

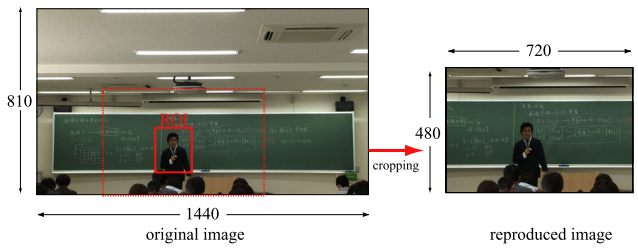


Fig. 1. Cropped image from high resolution image.

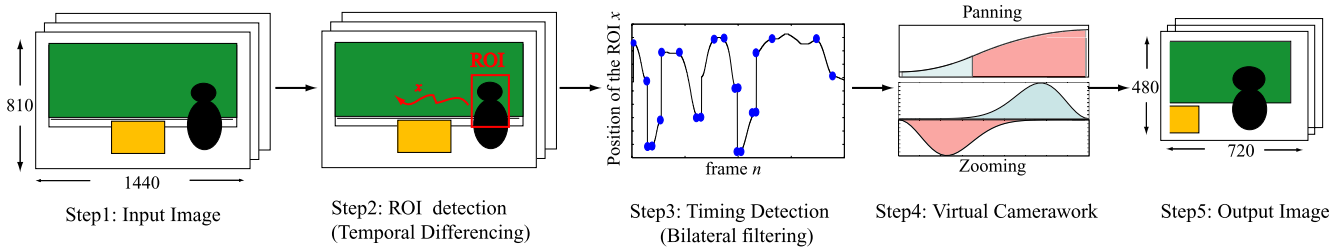


Fig. 2. Generating lecture video using virtual camerawork.

To solve this problem, our approach is to generate a lecture video by cropping from the high resolution image to track the ROI such as the instructor as shown in Fig.1. Fig.2 shows the procedure of generating a video by virtual camerawork. First, the system detects a moving object by temporal differencing. Next, the timing for virtual camerawork is detected using bilateral filtering and zero crossing. If the ROI has a large movement, this period of the video is classified into panning, and if the ROI has no motion but has voice activity, this period is classified into zooming. Finally, virtual camerawork is calculated based on the shooting technique of a broadcast cameraman [2].

3. EVENT DETECTION

To generate a time shrunk video, our approach is to detect important events such as speech periods and chalkboard writing periods by analyzing audio and images.

.

3.1. Voice Activity Detection

Our voice activity detection (VAD) computes 16 dimensional LPC cepstrum, and classifies whether a frame contains speech or not by calculating a Mahalanobis distance. Conditions of acoustic analysis are summarized in Table 1.

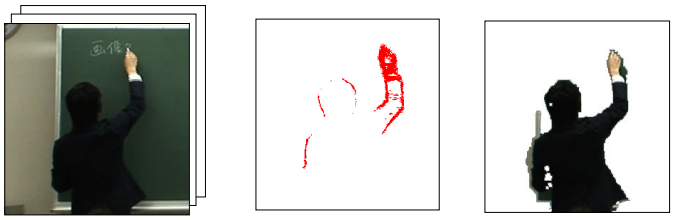
Table 1. Conditions of acoustic analysis for VAD.

Sampling frequency	11kHz
Analysis window	48 msec Hamming window
Frame shift	18 msec
Feature parameters	16 LPC cepstrum Power spectrum(150-900Hz)

3.2. Chalkboard Writing Detection

Extracting a precise object region is needed for detecting periods of writing characters on the chalkboard. Nishiguchi et al. [6] proposed a method for detecting characters from a mosaic image of the chalkboard. However, any object detection such as background subtraction can not be employed to extract all foreground pixels due to lighting changes, etc. Our approach uses a combination of object detection and segmentation (graph cuts technique) to extract precise foreground pixels.

Step 1: Object Detection and Segmentation Temporal differencing is robust to lighting change. So, we apply temporal differencing to detect a ROI from the high resolution images. Since the temporal differencing can not extract all foreground pixels of moving objects, we use a graph cuts technique [7].



(a) Input images (b) Detected pixels(seed) (c) Segmentation example

Fig. 3. Segmantation.

The Detected pixels by temporal differencing are set as seeds for segmentation by graph cuts as shown in Fig.3. Finally, an object mask at frame t O_M^t is obtained from the segmentation results of the graph cuts as follows:

$$O_M^t = \begin{cases} 1 & : object \\ 0 & : background \end{cases} \quad (1)$$

Step 2: Generation of Current Chalkboard Image A current chalkboard image I_C^t , which does not contain any foreground objects such as the instructor, at frame t is obtained by

$$I_C^t = I^t \cdot \bar{O}_M^t + I_C^{t-1} \cdot O_M^t \quad (2)$$

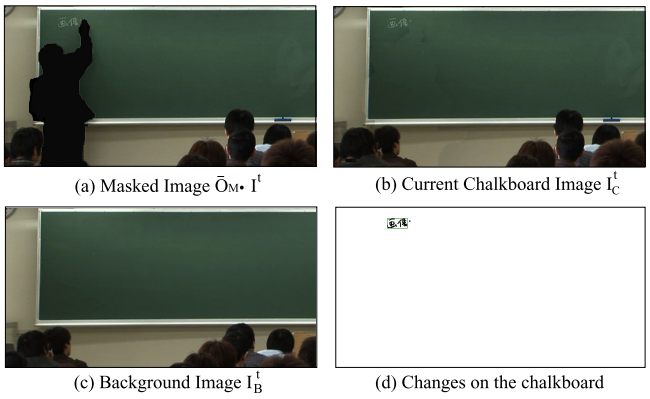
where I^t is a current image which may contain the instructor, and I_C^0 is prepared in advance as a background image (initial chalkboard image). Fig. 4(b) shows an example of a current chalkboard image generated from a current image I^t.

Step 3: Chalkboard Writing Detection In order to detect a change on the chalkboard, a temporal changes Δ_t is computed by

$$\Delta_t = \sum_{(i,j)\in I} |I_C^t(i,j) - I_C^{t-1}(i,j)| \quad (3)$$

The value of Δ_t becomes large during the writing of characters on the chalkboard. Once the writing is finished, the

Δ_t becomes 0 even if the instructor is moving in front of the chalkboard. By thresholding the value of Δ_t, our method enables precise detection of the writing period. Fig. 4(d) shows example of chalkboard writing detection.



(a) Masked Image $\bar{O}_M \cdot I^t$ (b) Current Chalkboard Image I^t_C

(c) Background Image I^t_B (d) Changes on the chalkboard

Fig. 4. Examples of chalkboard writing detection.

3.3. Generating A Time Shrunk Video

Even though the voice activity and the chalkboard writing are detected on a frame by frame basis, final decision is made by checking the temporal continuity which is a voting from a frame buffer (40 frames) centered on the current frame. Finally, our method outputs speech periods and chalkboard writing periods as shown in Fig. 5. To generate a time shrunk lecture video, we delete content-free periods such as block A and D (silence) in Fig.5. Writing periods without voice activity are important to understand the content of the lecture, but it can be shrunk by fast-forwarding the video. In our implementation, the writing period is fast-forwarded to 3 times faster than the original video by down-sampling.

After deleting the content-free periods from the original lecture video, a transition from one period to the next period occurs as a sudden scene change. This makes the viewer feel uncomfortable. To solve this problem, we apply a cross fade effect for the transition by alpha blending (See Fig.6(b)).

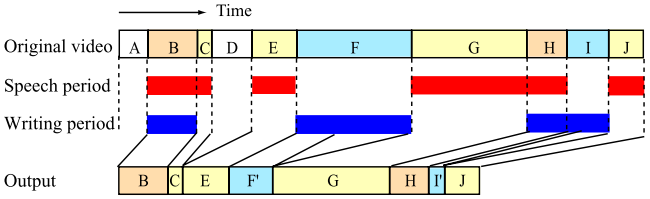


Fig. 5. Generating a time shrunk video.

4. EXPERIMENTAL RESULTS

4.1. Evaluation

We videotaped 3 lectures (each video is 90 min long) by HDV camcorder. These video sequences are reproduced as time shrunk lecture videos by our method. Some periods of "writing" and "erasing" the characters on the chalkboard are contained in these lecture video sequences.

In this evaluation, we use "recall" and "precision" for determining effectiveness of detection result. The "recall" is the ratio of the number of true positives retrieved to the sum of total number of true positives and false negatives in the data. The "precision" is the ratio of the number of true positives retrieved to the sum of the number of true positives and false positives as in the following equations. A human operator evaluates video footage to determine speech and writing periods.

$$recall = \frac{True\ positives}{True\ positives + False\ negatives} \quad (4)$$

$$precision = \frac{True\ positives}{True\ positives + False\ positives}$$

4.2. Results

Table 2 shows the experimental results of voice activity detection compared to [6]. We see that the average of recall and precision for voice activity detection is about 96%. Sometimes, talking in undertones causes losses of the speech period. There are 200 false positive frames in a 90 min video because of the students' voices.

Table 2. Results of voice activity detection

	Recall [%]	Precision [%]
Movie1	95.6	97.5
Movie2	97.5	96.0
Movie3	97.0	94.6
Ave.	96.7	96.0

Table 3 shows the experimental results of chalkboard writing detection. We see that recall was no less than 82%, and precision was no less than 93%. It is clear that our method has higher performance compared to the writing detection method in [6]. This is because our method segments more precise foreground pixels by a combination of object detection and segmentation. When characters on the chalkboard are obstructed by the instructor, our method can not detect the writing period. However, each undetected period by our method was only 5 seconds, which is not a problem in our application.

4.3. Generating a time shrunk lecture video result

Fig.7 shows the compaction rates of a generated video by human operator editing, our method, and [6]. We can see that, compared to the method (C) in [6], the compaction achieved by our method (B) is closer to the compaction achieved by human editor. This is because our method uses object detection

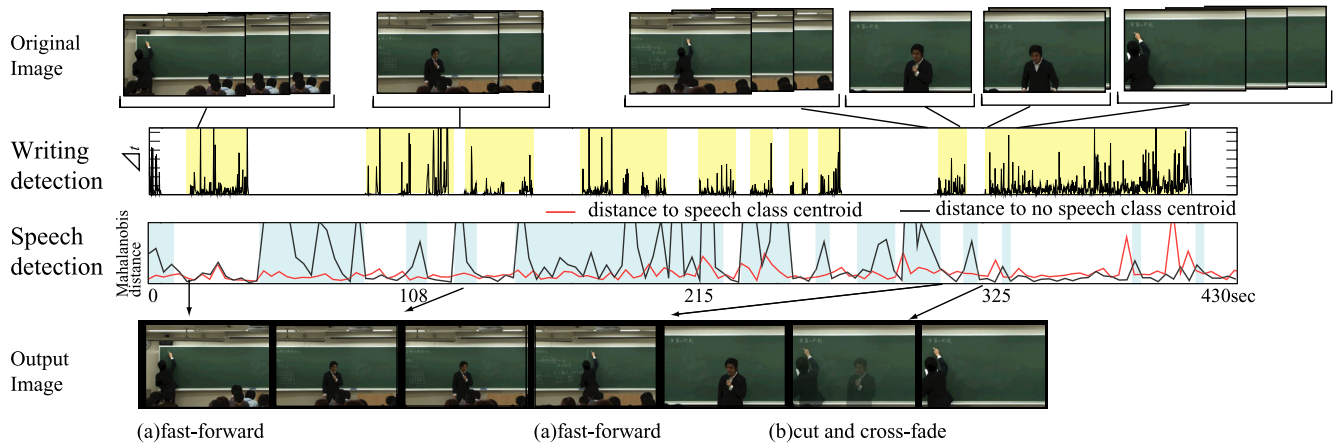


Fig. 6. Example of time shrunk lecture video by event detection.

Table 3. Results of chalkboard writing detection.

	Recall [%]		Precision [%]	
	Our method	Method [6]	Our method	Method [6]
Movie1	84.8	53.9	96.7	84.6
Movie2	82.4	54.7	93.9	74.9
Movie3	88.1	57.0	96.4	87.5
Ave.	85.1	55.2	95.7	82.3

and segmentation to detect the timing of the writing period very accurately.

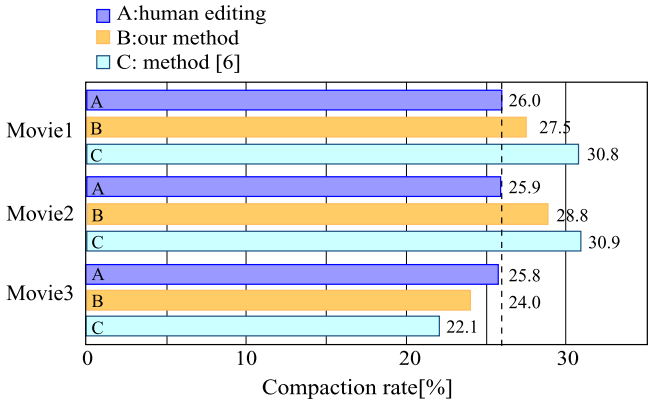


Fig. 7. Compaction rate.

5. CONCLUSION

This paper presents a novel approach for generating a time shrunk lecture video using event detection. Our method detects speech periods by voice activity detection and chalk board writing periods by a combination of object detection and segmentation techniques. By deleting the content-free periods and fast-forwarding the chalkboard writing periods, our method can generate a time shrunk lecture video automatically. The resulting generated video is about 20%∼30% shorter than the original video in time. This is almost the same as the results of manual editing by a human operator.

Indexing of lecture contents is left as future work.

6. REFERENCES

[1] T. Yokoi and H. Fujiyoshi, "Virtual Camerawork for Generating Lecture Video from High Resolution Images", Proc. of IEEE ICME 2005, July, 2005.

[2] D. Kato, M. Yamada, and K. Abe, "Analysis of the Work and Eye Movement of Broadcasting-Studio Cameramen," The Journal of Institute of Television Engineers of Japan, Vol. 49, No. 8, pp. 1023-1031, 1995.

[3] M Smith and T Kanade, "Video Skimming and Characterization through the Combination of lmage and Language Understanding Techniques", Proc. of CVPR, 1997.

[4] K. Miura, R. Hamada, I. Ide, S.i Sakai, H. Tanaka, "Motion based automatic abstraction of cooking videos",Proc. of ACM Multimedia 2002 Workshop on Multimedia Information Retrieval, 2002.

[5] K. Ishizuka, Y. Kameda, M. Minoh, "Speech and Video Indexing on Automatic Lecture Recording System", IEICE technical report. PRMU99-258, pp. 91 - 98, 2000.

[6] S. Nishiguchi, S. Senda, M. Minoh, K. Ikeda "A Recording System of Text on a Blackboard using an Active Camera", Proc. of the IEICE General Conference, pp. 37-42, 1996.

[7] Y. Boykov, M. Jolly, "Interactive Graph Cuts for Optimal Boundary & Region Segmentation of Objects in N-D Images", Proc. of ICCV, vol. I, pp. 105-112, 2001.