STRUCTURING AND ANALYZING LOW-QUALITY LECTURE VIDEOS

Yu-Tzu Lin[1]
Bai-Jang Yen[2]
Greg C. Lee[2]
[1]Information Technology Center
National Taiwan Normal University
[2]Department of Computer Science and Information Engineering
National Taiwan Normal University

ABSTRACT

This paper presents a lecture video structuring and analysis scheme to provide students an efficient way to access the lecture content. Instead of using color-based or histogram-based methodologies, we propose a new edge-based shot boundary detection algorithm to accurately rebuild the slide structure. The proposed approach can successfully resist the unwanted influences induced from the variant illumination condition and occlusions. Besides, original slide content can be extracted excluding any obstruction by using human removing techniques. Furthermore, the teaching focus is analyzed so that this system becomes more useful for learning.

Index Terms— ***video structuring***, shot change detection, moving object detection, distance learning

1. INTRODUCTION

In recent years, videos have become immensely popular. They are being generated at an enormous rate everyday by a variety of sources such as satellites, medicine imaging, news, home entertainment systems, digital learning systems etc. This large amount of video data makes it a tedious and hard job to browse and annotate them by just fast forward and rewind, besides, transmitting all video data is not practical because the network bandwidth is limited and what users need is only part of the data. Therefore, organizing this information into well structured data is of crucial importance for using these videos in a meaningful way. The user can then readily retrieve those sections of the video that he is interested in without having to go through the all the videos involved. The structuring, indexing and retrieval of video data are essential tasks in many applications.

Traditional distance learning systems [1] considered slides as a video and compressed it with a low-bit rate coding to save network bandwidth. However, slide content is severely redundant in a video and only contains new information whenever there is a slide flipping, the instructor points at the slide or writes something in it. Therefore, in this paper, we try to analysis the lecture video by extracting the slide structure and finding the instructors' emphasis so that the students could read the lecture content much more efficiently and easily, and besides, the amount of lecture video data can be considerably decreased. Scene change detection plays an important role in video retrieval applications. These issues are intensively studied [2,3] and several working systems have been developed [4]. However, there are differences between lecture video and other ones: the variation of the lecture video is only in the slide content (text, graphics, etc.), and is relatively small compared to the slide background, so that the shot boundary detection becomes more difficult than that of general videos because the traditional color-based [3] or histogram-based algorithms will fail. [5] also structured the lecture video by scene transition detection, but the performance highly depends on the text recognition result.

Moreover, keeping the instructors away from the screen is not reasonably, so part of the slide may be obstructed by the instructor or even the students. In addition, the color of lecture slides is critical to illumination conditions because that of the lecture recoding environment are variant. So, the challenge of this problem is to handle a low quality video which is recorded in an unconstraint environment.

In the rest of this paper, we will briefly introduce the framework of the proposed system in Section 2. An edge-based shot boundary detection algorithm and teaching focus analysis scheme are provided in Section 3 and Section 4, respectively. Section 5 shows the experimental results and then Section 6 gives the concluding remarks.

2. SYSTEM OVERVIEW

The system framework is illustrated as Fig. 1. Firstly, the human area in the lecture video is detected, and at the same time, the video is preprocessed to be illumination-independent. According to the produced information in the first step, the slide shot change is then detected and finally the teaching focus is analyzed.

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

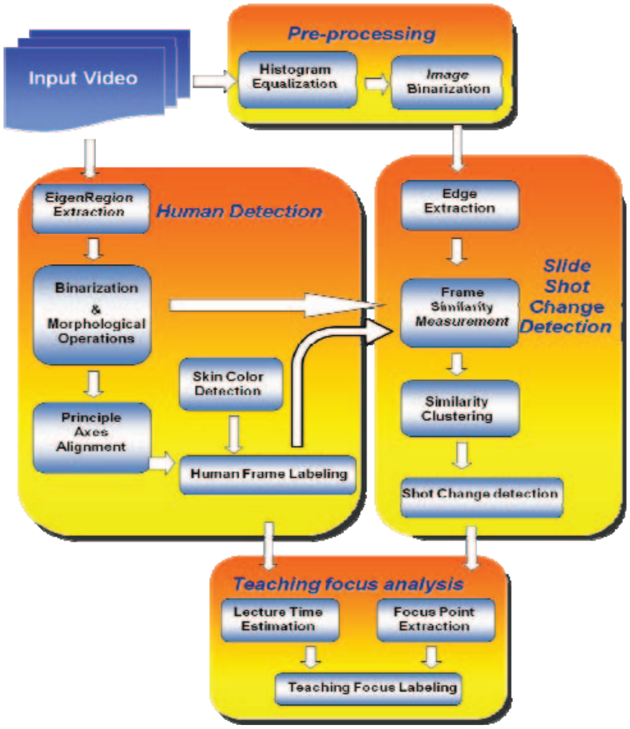


Figure 1. The system framework

3. EDGE-BASED SLIDE SHOT BOUNDARY DETECTION

There may be two illumination sources while the lecture is recoded: sunlight and lamplight. Instructors may turn on and off the light or open and close the windows according to teaching situation and the environment. Consequently, it will be impractical to use the histogram-based algorithms of shot change detection because the color variation of the same slide will be induced from irregular illumination. Moreover, in the lecture slides video, differences between slides often occurs only in the lecture content, the slide background is invariant. Thus, both color-based and histogram-based methods will fail because the content change is relatively very small compared to the constant background. In order to solve this problem, we propose an edge-based shot boundary detection method

3.1 Pre-processing

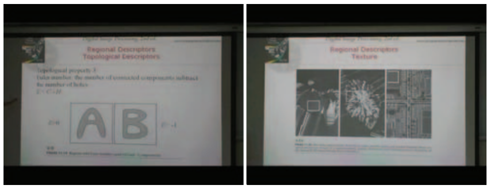
Although edge-based methods are much more insensible to illumination conditions, the image contrast is still influenced by the same and the produced edges will be instable. Therefore, the video frames are histogram equalized at first. After the histogram equalization, we also binarize the image by Otsu's method [6] to prevent the influence of slight illumination change.

3.2 Human Detection

In the lecture video, there are two acute change cases which may be detected as shot transitions: the first one is the real slide shot change (Fig. 2), and the second one occurs when something apart from the slide content (e.g., the instructor or students) enters the scene (Fig. 3), which will cause detection errors. Therefore, the area of human body should be ignored in shot change detection.

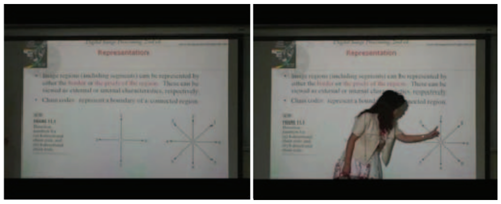
We extract the human area by detecting the moving object in the video frames, which is carried out by finding the eigenregions in the frames. That is, the moving objects can be distinguished from the still objects by methods of classification. The PCA (Principal Component Analysis)-based approach is used in this paper, which is detailed in the following. Three successive frames F_{i-1}, F_i, and F_{i+1} are firstly transformed into a matrix $X=[F_{i-1}\ F_i\ F_{i+1}]$, then the covariance C is computed as $C=X^TX$. Finally, each frame is aligned with the first two principle vectors (which are the eigenvectors of C associated with the two largest eigenvalues. Thus, the area with higher values represents that with higher variances, i.e., the moving object.

After the eigenregion is extracted, the produced image (Fig. 4 (d)) is binarized and applied by morphological operators to fill and smooth the region in order to obtain a more stable mask. Fig. 4 shows one example of moving object detection, in which (a), (b), and (c) are three successive frames, (d) is the corresponding eigenregion, (e) is the binarized one, and (f) is the final mask after morphological operations.



(a) (b)

Figure 2. Shot change: (a) the current frame, and (b) the next frame.



(a) (b)

Figure 3. An acute frame transition without slide change: (a) the original slide, and (b) the frame in which some human enters.

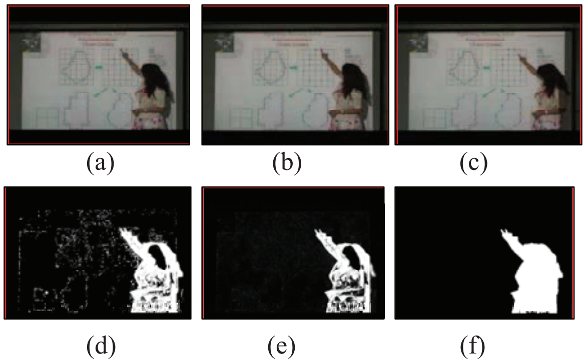

(a) (b) (c)
(d) (e) (f)

Figure 4. Human detection: (a), (b), and (c) are successive frames, (d) Eigenregion (lighter area), (e) binarized eigenregion, and (f) the resulting mask.

In addition, skin color information is employed to assure correctness of the human area. We then use this mask (the obtained area) to mask out the motion area and perform the following step, that is, only the inhuman areas are used in the shot boundary detection.

3.3 Shot Boundary Detection

This paper proposed a new edge-based algorithm to detect shot boundary. First, Kirsch masks of 8 directions (Fig. 5) are applied to approximate the gradient values of the image.

$$NW=\begin{bmatrix}5&5&-3\\5&0&-3\\-3&-3&-3\end{bmatrix}\quad N=\begin{bmatrix}5&5&5\\-3&0&-3\\-3&-3&-3\end{bmatrix}\quad NE=\begin{bmatrix}-3&5&5\\-3&0&5\\-3&-3&-3\end{bmatrix}\quad E=\begin{bmatrix}-3&-3&5\\-3&0&5\\-3&-3&5\end{bmatrix}$$

$$W=\begin{bmatrix}5&-3&-3\\5&0&-3\\5&-3&-3\end{bmatrix}\quad SW=\begin{bmatrix}-3&-3&-3\\5&0&-3\\5&5&-3\end{bmatrix}\quad S=\begin{bmatrix}-3&-3&-3\\-3&0&-3\\5&5&5\end{bmatrix}\quad SE=\begin{bmatrix}-3&-3&-3\\-3&0&5\\-3&5&5\end{bmatrix}$$

Figure 5. Kirsch masks (N: north E: east W: west S: south).

Then, frame similarities are computed by:

$$S_{ij}=\frac{\sum_{x,y}[(F_i(x,y)\wedge F_j(x,y))\wedge(D_i(x,y)==D_j(x,y))-P_{i,j}(x,y)]}{\sum_{x,y}[F_i(x,y)\vee F_j(x,y)]} \tag{1}$$

Where

$$F_i(x,y)=\begin{cases}1, & \text{if pixel } (x,y) \text{ in the } i\text{th frame is the edge point,}\\ 0, & \text{otherwise.}\end{cases} \tag{2}$$

$D_i(x,y)$ is the edge direction of the pixel (x,y) in the ith frame, and

$$P_{ij}(x,y)=F_i(x,y)\oplus F_j(x,y),\quad \oplus \text{ is the } xor \text{ operator,} \tag{3}$$

which is the penalty function evaluating the difference between $F_i(x,y)$ and $F_j(x,y)$. As shown in Fig. 6, suppose (a) and (b) are current frame and the frame with the instructor, respectively. The area for similarity computation should be filtered by the human mask (as shown in Fig. 6 (c)), and the penalty area (P=1 in white area) is as Fig. 6 (d).

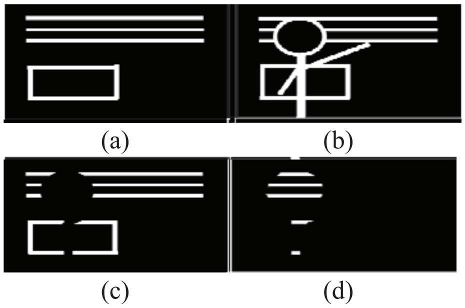
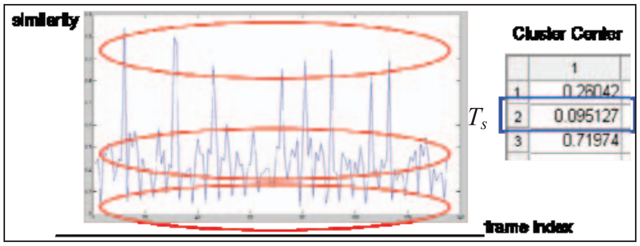

(a) (b)
(c) (d)

Figure 6. The area for similarity computation: (a) the original frame image, (b) the frame with the human, (c) the area after masking out the humans, and (d) the penalty area.

Since the differences between different slides are relatively small compared to the background pixels, the proposed similarity evaluation function is especially suitable for slide change detection because the background area is also filtered out by this equation.

After similarity values of all frames are computed, we classify the similarity values into 3 clusters by FCM (fuzzy *c*-means), and each cluster center is computed. The value of the median cluster center is then used as the threshold T_S (see Fig.7). Similarity values are computed for every two successive frames. Slides with higher similarity values than T_S are shot change points. The detection results then can be used to synchronize with the original slide file.



	1
1	0.26042
2	0.095127
3	0.71974

Figure 7. Similarity clustering for the threshold evaluation.

4. TEACHING FOCUS ANALYSIS

In this section, a teaching focus extraction method will be presented. We try to understand part of the lecturer's opinions in current slide. The most important information that we could mine from the lecture video is the teaching focus, which is meaningful for students. After structuring the lecture slides, the teaching focus is labeled according to the teaching time of each slide and the location where the instructor wants to emphasize with his fingers.

Therefore, besides the analyzing of the statistics of teaching time for each slide, the fingers' location should be detected. We use HSV color space to find the areas of skin color [6], which includes the human face, hands and arms. For each skin color area A_i, its center C_{Ai} is firstly computed, then the human location H is determined by computing the center of all C_{Ai}. After finding the human location, the focus area F is decided by the rule:

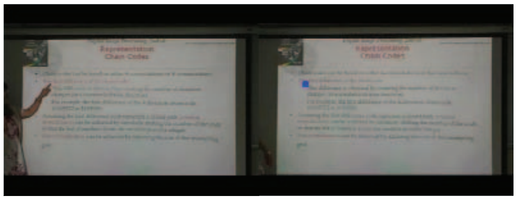
if (H>=$image\_width$/2) then

$$image\_width/2 \le F \le image\_width,$$

Else

$$0 \le F \le image\_width/2.$$

Then, for pixels of all A_i in F, the distances from H are calculated, and the pixel with a highest distance is considered as the pointer location, i.e. the teaching focus point (as shown in Fig. 8). Finally, the important content of the slide is labeled according to the location of the teaching focus.



(a) (b)

Figure 8. Teaching focus extraction: (a) Original frame, and (b) Teaching focus point (blue point).

5. EXPERIMENT RESULTS

In the experiments, there is no constraint on the lecture recoding, that is, the illumination condition is variant, and the humans may enter the slide region. The frame rate in this experiment is only 1 fps, if it increases, then the performance will become higher. The total length of the test video is 53 minutes and the resolution is 640×480. Fig. 9 illustrates the similarity values of for every two successive frames in a lecture video. The obtained recall and precision are 0.98 and 0.85, respectively.

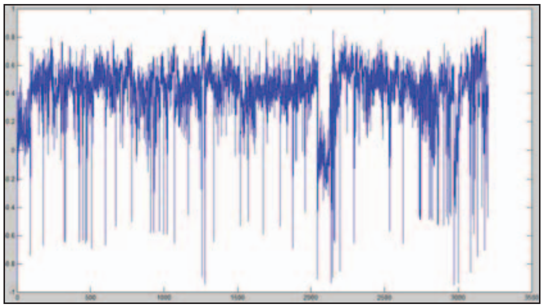
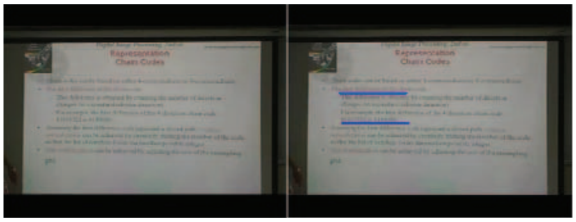


Figure 9. Similarity values of all frames in a lecture video.

In Fig. 10, one example of the teaching focus extraction is presented: Fig. 10 (a) is the original frame image and (b) is the frame in which the important content is highlighted with blue lines.



(a) (b)

Figure 10. Teaching focus labeling: (a) the original frame, and (b) the resulting frame in which the teaching focus content is highlighted with blue lines.

6. CONCLUSIONS

We propose a video structuring and analysis scheme for lecture videos recorded in an unconstraint environment. Experiment results show the feasibility of the proposed method, that is, the slide structure can be correctly reconstructed even if the illumination conditions is variant or the slides are obstructed by humans. Besides, the teaching focus is analyzed according to the location at which the instructor want to point and the lecture time of some slide. This system can really provide students a convenient and efficient way to access the lecture video.

7. ACKNOWLEDGMENTS

This work was partially supported by the National Science Council and the Ministry of Education of China under contact no. NSC97-2631-H-003-001 and NSC97-2218-E-003-001.

8. REFERENCES

[1] S.G. Deshpande, J.-N. Hwang, "A real-time interactive virtual classroom multimedia distance learning system," *IEEE Trans. Multimedia*, Vol. 3, No. 4, 2001.
432--444.

[2] D. Zhang, W. Qi and H. J. Zhang, "A New Shot Boundary Detection Algorithm," *Proc. of IEEE Pacific Rim Conference on Multimedia,* Beijing, China, pp.63-70, 2001

[3] A. Nagasaka and Y. Tanaka, "Automatic video indexing and full-video search for object appearances," P*rc. of IFIP Second Workshop Conf. on Visual Database System II, Budapest, Hunary*, pp.113-127, 1992

[4] R. Zabith, J. Miler, and K. Mai, "A Feature-Based Algorithm for Detecting and Classifying Scene Breaks," *Proc. of ACM Multimedia, San Francisco, CA, USA*, pp.189-200, 1995

[5] F. Wang, C. W. Ngo, and T. C. Pong, "Structuring Low-quality Videotaped Lectures for Cross-Reference Browsing by Video Text Analysis," *Pattern Recognition*, Vol. 41, No. 10, pp. 3257-3269, Oct 2008.

[6] N. OTSU, "A Threshold Selection Method from Gray-Level Histogram," *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS*, Vol. SMC-9, No. 1, JANUARY 1979.