VIDEO SCENE CHANGE DETECTION METHOD USING UNSUPERVISED SEGMENTATION AND OBJECT TRACKING*

Shu-Ching Chen[1], Mei-Ling Shyu[2], Cheng-Cui Zhang[1], R. L. Kashyap[3]

[1]Distributed Multimedia Information System Laboratory
School of Computer Science, Florida International University, Miami, FL 33199

[2]Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33124

[3]School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907

ABSTRACT

In order to manage the growing amount of video information efficiently, a video scene change detection method is necessary. Many advanced video applications such as video on demand (VOD) and digital library also require the scene change detection to organize the video content. In this paper, we present an effective scene change detection method using an unsupervised segmentation algorithm and the technique of object tracking based on the results of the segmentation. Our results have shown that this method can perform not only accurate scene change detection, but also obtain object level information of the video frames, which is very useful for video content indexing and analysis.

1. INTRODUCTION

Recently, multimedia information has been made overwhelmingly accessible with the rapid advances in communication and multimedia computing technologies. The requirements for efficiently accessing mass amounts of multimedia data are becoming more and more important. Video scene change detection is a fundamental operation used in many multimedia applications such as digital libraries and video on demand (VOD), and it must be performed prior to all other processes [1,2].

Video data can be divided into different shots. A shot is a video sequence that consists of continuous video frames for one action. Scene change detection is an operation that divides video data into physical shots.

There are a number of methods for video scene change detection in the literature. The matching process between two consecutive frames is the essential part of it. Many of them use the low-level global features such as the luminance pixel-wise difference [3], luminance or color histogram difference [4] to compare two consecutive frames. However, since luminance or color is sensitive to small change, these low-level features cannot give a satisfactory answer to the problem of scene change detection. In Lee's work [5], he focused on scene change detection in the compressed video data such as MPEG-1. In his method, he used the binary edge maps as a representation of key frames. Two frames can then be compared by calculating a correlation between their edge maps.

In this paper, focusing on the uncompressed video data, we propose an innovative scene change detection method using an unsupervised segmentation algorithm and the object tracking technique. By using the segmentation algorithm, the segmentation mask map of each video frame can be automatically extracted. The segmentation mask map, in another word, can be deemed as the clustering feature map of each frame. In such a way, the pixels in each frame have been grouped into different classes (for example, 2 classes). Then two frames can be compared by checking the difference between their segmentation mask maps. In addition, in order to better handle the situation of scene rotation, the object tracking technique based on the segmentation results is used as an enhancement to the basic matching process. Since the segmentation results are already available, the cost for object tracking is almost trivial. Moreover, our key frame representation uses the information of the segmentation results such as the bounding boxes and the positions of the segments within that frame. The advantages of using unsupervised segmentation are:

- It is fully unsupervised, without any user interactions.
- The algorithm for comparing two frames is simple and fast.
- The object level segmentation results can be further used for video indexing and content analysis.

This paper is organized as follows. In Section 2, we explain the scene change detection method as well as the mechanism of the unsupervised segmentation algorithm and the object tracking technique. In Section 3, experimental results are analyzed to show the effectiveness of the proposed method. Finally, conclusions and future work are given in Section 4.

2. SCENE CHANGE DETECTION METHOD

In this section, we first explain how the unsupervised segmentation algorithm and object tracking work, and then give out the steps of the scene change detection method based on the discussion.

* This research was supported in part by NSF CDA-9711582.

0-7695-1198-8/01/$10.00 (C) 2001 IEEE

2001 IEEE International Conference on Multimedia and Expo
ISBN 0-7695-1198-8/01 $17.00 © 2001 IEEE

2.1 Segmentation Information Extraction

In this paper, we use an unsupervised segmentation algorithm to partition the video frames. First, the concepts of a class and a segment should be clarified. A class is characterized by a statistical description and consists of all the regions in a video frame that follows this description; while a segment is an instance of a class. In this algorithm, the partition and the class parameters are treated as random variables. The method for partitioning a video frame starts with a random partition and employs an iterative algorithm to estimate the partition and the class parameters jointly [6, 7, 8].

Suppose there are two classes -- *class1* and *class2*. Let the partition variable be $\boldsymbol{c} = \{\boldsymbol{c_1}, \boldsymbol{c_2}\}$, and the classes be parameterized by $\theta = \{\theta_1, \theta_2\}$. Also, suppose all the pixel values y_{ij} (in the image data Y) belonging to class k $(k=1,2)$ are put into a vector $\boldsymbol{Y_k}$. Each row of the matrix Φ is given by $(1, i, j, ij)$ and a_k is the vector of parameters $(a_{k0}, \ldots, a_{k3})^T$.

$$y_{ij} = a_{k0} + a_{k1} i + a_{k2} j + a_{k3} ij, \ \forall (i, j)\ y_{ij} \in \boldsymbol{c_k}$$
$$\boldsymbol{Y_k} = \Phi\, a_k$$
$$\hat{a}_k = (\Phi^T \Phi)^{-1} \Phi^T \boldsymbol{Y_k}$$

The best partition is estimated as that which maximizes the a posteriori probability (MAP) of the partition variable given the image data Y. Now, the MAP estimates of $\boldsymbol{c} = \{\boldsymbol{c_1}, \boldsymbol{c_2}\}$ and $\theta = \{\theta_1, \theta_2\}$ are given by

$$(\hat{c}, \hat{\theta}) = Arg \max_{(c,\theta)} P(c, \theta \mid Y)$$
$$= Arg \max_{(c,\theta)} P(Y \mid c, \theta) P(c, \theta)$$

Let $J(c, \theta)$ be the functional to be minimized. With the above assumptions, this joint estimation can be simplified to the following form:

$$(\hat{c}, \hat{\theta}) = Arg \min_{(c,\theta)} J(c_1, c_2, \theta_1, \theta_2)$$
$$J(c_1, c_2, \theta_1, \theta_2) = \sum_{y_{ij} \in c_1} -\ln p_1(y_{ij}; \theta_1) + \sum_{y_{ij} \in c_2} -\ln p_2(y_{ij}; \theta_2)$$

The algorithm starts with an arbitrary partition of the data in the first video frame and computes the corresponding class parameters. Using these class parameters and the data, a new partition is estimated. Both the partition and the class parameters are iteratively refined until there is no further change in them. Since the successive frames do not differ much due to the high temporal sampling rate, the partitions of adjacent frames do not differ significantly. The key idea is then to use the method successively on each frame of the video, incorporating the partition of the previous frame as the initial condition while partitioning the current frame, which can greatly reduce the computing cost.

2.2 Object Tracking

The first step in connecting the segments obtained in each frame is to extract the segments in each class. For example in Figure 1, the CDs belong to one segment. Then the bounding box and the centroid point for that segment are obtained. The next step is to connect the related segments in successive frames. The idea is to connect two segments that are spatially the closest in the adjacent frames [9]. In another word, the distances between the centroids of the segments in adjacent frames are used as the criteria to track the related segments. Besides, size restriction should be employed in determining the related segments.

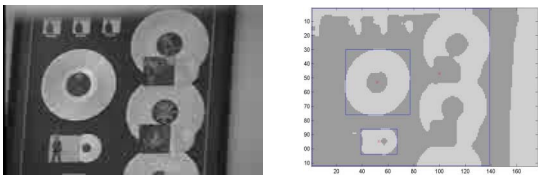


Figure 1: Examples of *classes* and *segments.* The left is the original video frame, and the right is the segmentation mask map of the left frame.

2.3 Scene Change Detection Method

The steps for our proposed scene change detection method are given in the following:

1. Start the segmentation on the first video frame using the randomly generated initial partition and get the segmentation mask map for it. Let the variable *pre_map* keep the value of the obtained segmentation mask map.
2. Do the segmentation on the next frame and get the current mask map. Let the variable *cur_map* keep the current mask map's value.
3. *diff* = ‖ *cur_map-pre_map* ‖;
   diff_num = the number of elements in *diff* which are nonzero;
4. If (*diff_num* / (total number of elements in *diff*)) < *Low_Th*
   then
   Not scene change, go to step 5;
   else
   If (*diff_num* / (total number of elements in *diff*)) < *Hi_Th*
   Do object tracking between the previous frame and the current frame, and let the number of unmatched segments in the previous frame be *num_unmatch*;
   If *num_unmatch* <= 2
   then
   Not scene change, go to step 5;
   else
   Mark the current frame as a *scene cut frame*;
   End;
   End;
   Mark the current frame as a *scene cut frame*;
   End;
5. *pre_map* = *cur_map*;
   Go to step 2 and repeat.

(Here, *Low_Th* and *Hi_Th* are two threshold values of the *diff_num* that are derived from the experiential values.)

3. EXPERIMENTAL RESULTS

We have performed a series of experiments on TV news videos (in MPEG-1 format) that include FOX 25 LIVE NEWS and ABC 7 NEWS. The average length of each news video is about 2120 frames long, and each frame is of size 117*176. The experimental

2001 IEEE International Conference on Multimedia and Expo
ISBN 0-7695-1198-8/01 $17.00 © 2001 IEEE
0-7695-1198-8/01/$10.00 (C) 2001 IEEE

results demonstrate the effectiveness of the proposed scene change detection algorithm.

Figure 2 gives an example scene cuts detection results for an ABC 7 NEWS video. Figure 2(a) shows the original video frames that have been detected as the scene cuts, and Figure 2(b) is the example segmentation mask maps for the scene cuts in Figure 2(a). Since the segmentation mask maps are binary data, it is very simple and fast to compare the two mask maps of the successive frames. The performance is given in terms of *precision* and *recall* parameters. N_C means the number of correct scene change detections, N_E means the number of incorrect scene change detections, and N_M means the number of missed scene detections.

$$precision = \frac{N_C}{N_C + N_E}$$

$$recall = \frac{N_C}{N_C + N_M}$$

In our experiments, the *recall* and the *precision* values are both above ninety percent. As mentioned before, the method of using low-level features is very sensitive to luminance and color change, but our segmentation-based method is not. Another thing should be mentioned here is that even it is efficient to simply compare the segmentation mask maps, the employment of the object tracking technique is very useful in case of scene rotation. It helps to reduce the number of incorrectly identified scene cuts.

Moreover, the process produces not only the scene cuts, but also the object level segmentation results. As can be seen from Figure 2(b), each detected scene cut is selected as a key frame and has been modeled by the features of its segments such as the bounding boxes and centroids. Based on this information, we can further structure the video content using some existing multimedia semantic model such as the multimedia augmented transition network (MATN) model [8].

4. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed an innovative scene change detection method using the unsupervised segmentation algorithm and object tracking technique, and showed the performance using the sample MPEG-1 news video data. The key idea of the matching process in scene change detection is to compare the segmentation mask maps between two successive video frames. In addition, the object tracking technique is employed as a complement to handle the situations of scene rotation without any extra overhead. Unlike many methods using the low-level features of the video frames, the proposed method is not sensitive to the small change in luminance or color. Moreover, it has high precision and recall values as shown in our experiments.

As mentioned in Section 3, the proposed method together with its segmentation results is very useful in modeling a video-based system. We are now investigating the possibilities of developing a more complete video-indexing framework to structure the video hierarchy for video database queries.

5. REFERENCES

[1] H. Zhang and S. W. Smoliar, "Developing power tools for video indexing and retrieval," in *Proc. SPIE'94, Storage and Retrieval for Image and video Databases II*, vol. 2185, San Jose, CA, 1994.

[2] B. Shahraray, "Scene change detection and content-based sampling of video sequences," in *Proc. SPIE'95, Digital Video Compression: Algorithm and Technologies*, vol. 2419, San Jose, CA, 1995.

[3] H. Zhang, A. Kankanhalli, and S. W. Smoliar, " Automatic partitioning of full-motion video," *Multimedia System*, vol. 1, 1993.

[4] D. Swanberg, C. F. Shu, and R. Jain, "Knowledge guided parsing in video database," in *Proc. SPIE'93, Storage and Retrieval for Image and video Databases, vol. 1908, San Jose, CA, 1993.*

[5] S.-W. Lee, Y.-M. Kim, and S.-W. Choi, "Fast Scene Change Detection using Direct Feature Extraction from MPEG compressed Videos," *IEEE Trans. on Multimedia*, vol. 2, No. 4, Dec. 2000.

[6] S.-C. Chen, S. Sista, M.-L. Shyu, and R. L. Kashyap, "Augmented Transition Networks as Video Browsing Models for Multimedia Databases and Multimedia Information Systems," *11th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'99),* pp. 175-182, November 9-11, 1999.

[7] S.-C. Chen, S. Sista, M.-L. Shyu, and R. L. Kashyap, "An Indexing and Searching Structure for Multimedia Database Systems," *IS&T/SPIE conference on Storage and Retrieval for Media Databases 2000*, pp. 262-270, January 23-28, 2000.

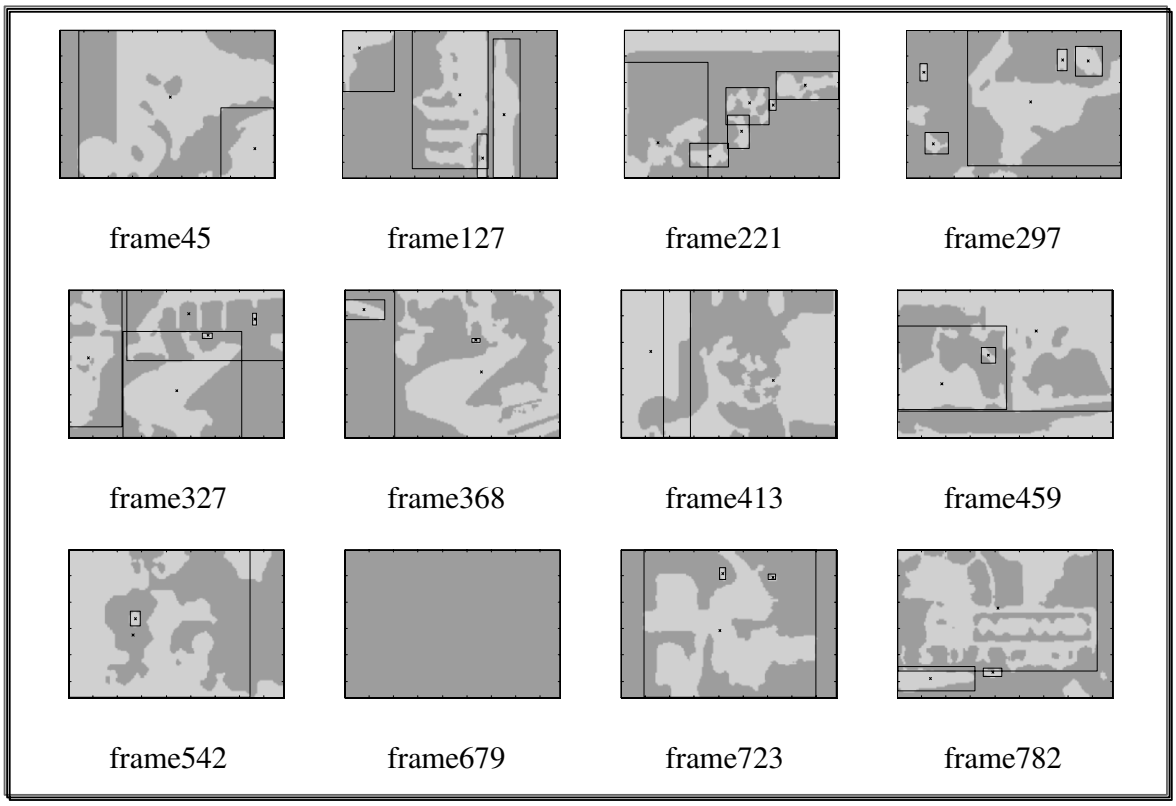
[8] S.-C. Chen, M.-L. Shyu, and R. L. Kashyap, "Augmented Transition Network as a Semantic Model for Video Data," accepted for publication, *International Journal of Networking and Information Systems*, Special Issue on Video Data, 2000.

[9] S. Sista and R. L. Kashyap, "Unsupervised Video Segmentation and Object Tracking," *IEEE Int'l Conf. on Image Processing*, 1999.

2001 IEEE International Conference on Multimedia and Expo
ISBN 0-7695-1198-8/01 $17.00 © 2001 IEEE

0-7695-1198-8/01/$10.00 (C) 2001 IEEE

a). The example scene cuts for the ABC 7 NEWS video sequence.



b). The example segmentation mask maps for the scene cuts in a).

Figure 2: The example scene cuts and their segmentation mask maps.

0-7695-1198-8/01/$10.00 (C) 2001 IEEE

2001 IEEE International Conference on Multimedia and Expo
ISBN 0-7695-1198-8/01 $17.00 © 2001 IEEE