A MULTI-MODAL-FEATURE BASED ALGORITHM FOR PARSING NEWS PROGRAM VIDEOS

Hsuan-Wei Chen, Jin-Hau Kuo, Jen-Hao Yeh and Ja-Ling Wu, ***Senior Member****, IEEE*
Communication and Multimedia Laboratory, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, R.O.C.

ABSTRACT

In this paper, a multi-modal-feature based scene change detection algorithm (which can be viewed as a mid-stage solution between the single-modal-feature- and the semantic-based approaches) is proposed to parse the news programs, effectively.

1. INTRODUCTION

Along with the great success of the Internet and the steady growth of broadband technology and computing power, the roles of the content provider have shifted from mass media to general-information end users. Because of the huge amount of data volume, it is difficult to search and index the video database, directly. Moreover, segmenting videos into semantic meaningful units is very difficult and is still a big challenge to the video segmentation society. Generally agreed, the first step towards video indexing is to detect boundaries between consecutive camera shots so as to organize videos into elementary indexing units.

In this paper, a multi-modal-feature based scene change detection algorithm (which can be viewed as a mid-stage solution between the single-modal-feature- and the semantic-based approaches) is proposed to parse the news programs, effectively.

It is believed that reasonable and effectively solutions of semantic video parsing should take users' intervention into account; in other words, the system must learn to adjust itself according to users' feedbacks. To meet this requirement, a user-friendly interface is a must for the system. Since our target is to analyze the recorded television news program, we perform the so-called two-pass analysis of videos and audios. The first pass generates rough low-level (non-semantic) shot changes using multiple-descriptors, based on the MPEG-7 visual standard. All the adopted descriptors can be extracted from MPEG-compressed bitstreams. In the second stage, we do an audio analysis to assist the detection of (more-semantic) scene changes. The scene or shot changes caused by some visual transition effects using cinematic techniques are also detected in this second stage. Some of these transition effects are still very hard to be accurately detected by using visual clues, only. But they will be obvious in the audio domain. After the two-pass analysis of audiovisual data and merging with the user's feedback, the system performs another analysis again so as to get more accurate results.

2. THE INHERENT SCENE CHANGE INFORMATION AND MACROBLOCK TYPE STATISTICS

A shot change, or shot cut, is defined as an image content switch between two consecutive frames with different scenes. From the above definition, the foundation of shot change detection is grounded on the similarity comparison between two consecutive frames, regardless of whether the video is uncompressed or compressed. However, based on the GOP structure of MPEG standard, the hint of shot change information obtained directly from compressed video sequences should not only consider the two consecutive frames, but also take the relationship between frames within a GOP into account.

Abrupt shot changes may occur at I frame, P frame, front B frame or rear B frame, as illustrated in Fig. 1. For most of the cases, the macroblock types in a B-frame are composed of Intra macroblocks, Forward Motion Compensation macroblocks, Backward Motion Compensation macroblocks and Bidirection Motion Compensation macroblocks. Let *Intra*, *FW*, *BW* and *BI* denote the percentages of these four types of macroblocks, with respective to the total macroblock count in a frame. Generally,

$$Intra + FW + BW + BI \approx 1 \quad (1)$$

In the case of shot change occurred at rear P- or I-frame (c.f. Fig. 1(a)), most of the macroblocks in the front and the rear B-frames (denoted as *Bf* and *Br*, respectively) are inclined to be of forward motion prediction type, because they are much more similar to the front P- or I-frame than the rear P- or I-frame. The confidence value of this shot change situation can be defined as:

$$FW(Bf) * FW(Br), \quad (2)$$

In the case of shot change that occurred at frame *Bf* (c.f. Fig. 1(b)), owing to the large difference between the content of front P-frame and the following two B-frames, most of the macroblocks in the front and the rear B-frames are expected to be of backward motion prediction type. The confidence value of this shot change situation can be defined as:

$$BW(Bf) * BW(Br), \quad (3)$$

Similarly, in the case of the shot change that occurred at frame *Br* (c.f. Fig. 1(c)), most of the macroblocks in the front B-frame are forward motion predicted and most of the macroblocks in the rear B-frame are backward motion predicted. The confidence value of this shot change situation can be defined as:

$$FW(Bf) * BW(Br) \quad (4)$$

By these equations, we can detect shot changes at frame level effectively. Further details can be found in [1]. Fig. 2 shows the snapshots of the tested advertisement program video and the corresponding P_b's from frame 101 to frame 300, where 'P_b' denotes the confidence value of shot change for each frame. It is obvious that taking P_b equal to 0.5 will detect almost all shot change locations.

3. THE MPEG-7 DESCRIPTORS RELATED TO SHAPE AND COLOR INFORMATION

3.1. The Shape Information

As suggested in the MPEG-7 visual part of experimentation Model Version 11.1 [2], the edge orientation types can be roughly classified into horizontal, vertical, diagonal (dia45), anti-diagonal (dia135) and non-directional edges in each local region (the so-called *sub-image* in this document), as shown in Fig. 3. More specifically, there are N non-overlapped sub-images and for each sub-image we generate a local edge histogram with 5 bins (horizontal, vertical, dia45, dia135 and non-directional edges). Since there are five types of edges for each sub-image, we have a total of N*5 histogram bins. As shown in Fig. 4, by further dividing the sub-image into *image-blocks*, edge type information can be extracted for each image-block.

The DC coefficient is, in particular, the average of the luminance or chrominance signal of all pixels within an 8 x 8 block. Because of the GOP structure of MPEG, we can get the DC coefficients free only in the I-frame. Yeo [3] proposed an approach which extracting DC coefficients from P- or B-frames to form a DC-image and a DC-sequence [3]. These DC coefficients, which represent the luminance mean values of 8 x 8 blocks can be applied to edge detection. We select the *image-blocks* of size 16 x 16 pixels so as to meet the macroblock size defined in the MPEG standard. Fig. 5 shows the *image-blocks* and the corresponding labeling of the four blocks. The DC coefficients of the four blocks within a macroblock can be defined as follows:

$$8*D_0(i,j) = \frac{4}{\text{MB\_size} \times \text{MB\_size}} \sum_{m=0}^{\text{MB\_size}/2\ -1} \sum_{n=0}^{\text{MB\_size}/2\ -1} I_{ij}(m,n) \quad (7)$$

$$8*D_1(i,j) = \frac{4}{\text{MB\_size} \times \text{MB\_size}} \sum_{m=MB\_size/2}^{\text{MB\_size}\ -1} \sum_{n=0}^{\text{MB\_size}\ -1} I_{ij}(m,n) \quad (8)$$

$$8*D_2(i,j) = \frac{4}{\text{MB\_size} \times \text{MB\_size}} \sum_{m=0}^{\text{MB\_size}/2\ -1} \sum_{n=MB\_size/2}^{\text{MB\_size}\ -1} I_{ij}(m,n) \quad (9)$$

$$8*D_3(i,j) = \frac{4}{\text{MB\_size} \times \text{MB\_size}} \sum_{m=MB\_size/2}^{\text{MB\_size}\ -1} \sum_{n=MB\_size/2}^{\text{MB\_size}\ -1} I_{ij}(m,n) \quad (10)$$

where $I_{ij}(m,n)$ is the $(m, n)^{th}$ gray level in the $(i,j)^{th}$ macoblock. $D_k(i,j)$ denotes the DC coefficient of the k^{th} block of the image block at position (i,j).

We have four mean values for four blocks and they are then convolved with the filter coefficients as shown in the following equations to obtain the edge magnitudes.

$$ver\_edge\_stg(i,j) = \sum_{k=0}^{3} \left| D_k(i,j) \times ver\_edge\_filter(k) \right| \quad (11)$$

$$hor\_edge\_stg(i,j) = \sum_{k=0}^{3} \left| D_k(i,j) \times hor\_edge\_filter(k) \right| \quad (12)$$

$$dia45\_edge\_stg(i,j) = \sum_{k=0}^{3} \left| D_k(i,j) \times dia45\_edge\_filter(k) \right| \quad (13)$$

$$dia135\_edge\_stg(i,j) = \sum_{k=0}^{3} \left| D_k(i,j) \times dia135\_edge\_filter(k) \right| \quad (14)$$

$$nond\_edge\_stg(i,j) = \sum_{k=0}^{3} \left| D_k(i,j) \times nond\_edge\_filter(k) \right| \quad (15)$$

where 'A_edge_stg(i,j)' denotes the edge strengths of the corresponding edge orientation types and 'A_edge_filter(k)' denotes the filters of the corresponding edge orientation types. The coefficients of each filter are specified in matrix form and shown in Fig. 6. Fig. 7 shows the edge-images of the test sequences **Chaire** and **Tennis,** with threshold 40. It appears that the rough shape of the original image is shown in the edge-image. For the edge-image of each frame, we generate two levels of representation of localization, namely Global and Local edge histograms [2], respectively. The Global edge histogram is composed of 5 bins. In Fig. 8, both the Global and Local edge histograms are presented. We observed that the prescribed two histograms are very similar to each other within a shot, but are quite different for two different shots.

3.2. The Color Information

In [2], the color layout descriptor is proposed to specify a spatial distribution of colors for high-speed retrieval and browsing. Here, we take advantage of the color layout information in each frame to detect the shot change. As mentioned in [2], obtaining the color layout information is to partition the whole image into *blocks* and derives the local dominant colors of each *block*. Due to the speed consideration, we adopt the simplest method to acquire the dominant colors in each *block*. For our goal of using color layout is just to detect the differences between frames, each *block* represented as one dominant color is enough.

The simple, but useful, color layout information can be easily obtained from MPEG compressed bitstreams. As stated in section III-1, the MPEG standard is based on 8 x 8 DCT's. The DC coefficient is, in particular, the average of the luminance or chrominance signal of all pixels within the block. As pre-described, these DC coefficients, which represent the luminance mean values of the block can be applied to extract the simple color layout. As one dominant color is chosen for representing each block, it can be approximated by the average color value within the block. During decoding, using the color layout image difference to detect the shot change location requires extra computations only after Variable Length Decoding. The methods [3-4] have been evaluated and adopted as they providing the best results.

Shot detection results for the three individual shot-change hints: **color, shape and inherent information** (obtained in section 2), are then integrated through the fusion rule, shown in eqn. (16).

$$\text{Frame\_shot}(t) = \begin{cases} 1, & \text{shot-change-frame} \quad \text{if } (c(t)=1) \cup (s(t)=1) \cup (h(t)=1) \\ 0, & \text{non} \qquad\qquad\qquad\quad \text{otherwise,} \end{cases} \quad (16)$$

where $c(t)$, $s(t)$ and $h(t)$ denote the shot results detected by color, shape and inherent information, respectively. Each hint domain is set to 1 when it detects a shot in its own domain.

A Frame is classified as an shot change if at least one of its three shot-change hints is set to 1. Otherwise, it is classified as a non-shot-change frame and Frame_shot*(t)* is set to 0.

4. THE MPEG-7 DESCRIPTORS RELATED TO THE AUDIO INFORMATION

Since the audio information is used to assist in detecting scene change or the transitional visual effects in the news program, only the low level tools defined in MPEG-7 audio is adopted. The flowchart of this procedure is illustrated in Fig. 9. The audio input is the MPEG audio extracted at the same time of the visual analysis. Then, we convert the MPEG audio to PCM waveforms.

There are usually two kinds of audio features: short-term frame level and long-term clip level features. An audio frame is defined as a group of neighboring samples last for about 10 to 40ms. As to the clip level features, it is usually used to extract the more semantic meaning of an audio signal. Thus, it is necessary to analyze over a much longer period, usually from one second to several seconds. The definitions of the frame and the clip are shown in Fig. 10. In our work, we use the hybrid features of frame and clip levels to detect the shot and scene changes.

As shown in Fig. 9, after converting the MPEG audio to PCM waveforms, the audio information is adapted to the MPEG-7 *ScalableSeries Type*. The Scalable series are the data-types for series of values (scalars or vectors). They allow the series to be scaled (downsampled) in a well-defined fashion. Two types of them are available: *SeriesOfScalarType* and *SeriesOfVectorType*. They are useful in particular to build descriptors that contain time series of values. Here the clip lengths we used are 0.1, 0.5, 1 and 3secs. There are many frames in one clip, and one frame is of length 512 samples (with a frame shift = 128 samples).

The features we adopted are divided into six groups named ***Volume related***, ***Zero Crossing Rate related***, ***Silence Segment Related***, ***Frequency Centroid and Band width Related***, ***Sub-band energy Related and Modulation Energy***. There are total fifteen different features in these six groups. The detailed formulas and definitions of these features can be found in [5-6].

5. EXPERIMENTAL RESULTS

The snapshot of the proposed scene change detection system is shown in Fig. 11. The system allows for the video input of MPEG-1, 2 formats. It provides two detection- modes: automatic and semi-automatic, for selection by users. In the automatic mode, the system adopts the default parameter P_b for each frame and the distance of shape and color descriptors for detecting the shot. Alternately, if semi-automatic mode is chosen, the parameters of those features mentioned above can be changed interactively by the user (c.f. Fig. 11(b)). On the left side of the snapshot is a window displaying the detection results. It marks the number and the type (I-, P-, or B-) of the shot frame detected. On the bottom side of the snapshot is a window displaying the edge-image.

By clicking the small shot icon on the left side, the user can browse the content clip from the shot point to the next point detected. If the user is unsatisfied with the results (e.g. missing a shot point), he can set the parameters interactively and process the analysis, focusing on the unsatisfied clip, again

The scene change results detected by using the audio features are shown in Fig. 12(a). Between frames 3613~3650, there is a news dissolve clip miss-detected by the visual features. Even with the shape information, the dissolve scene is not obvious because of the increasing number of the non-directional edges (caused by the dissolve effect). However, the scene change is detected by using the audio features, because of the different background music (a news ending and an advertisement following) involved. Note that the unit used in the audio part is 1 sec which equals to 30 frames.

6. CONCLUSION

In this paper, a news program parser using both visual and audio clues, based on the MPEG-7 descriptors, is presented. The so-called two-pass analysis of video and audio is suitable for parsing the recorded television news program off-line. The provided user-friendly interface is for user intervention, which is believed to be reasonable and effective for achieving semi-semantic video segmentations. With the aid of audio features, a dissolve scene change (which is hard to be detected by using visual clues only) can be detected successfully.

Although our current work still for from the target: semantic meaningful video parser; however, the proposed multi-modal-feature based approach shows its superiority, at least in a particular application (the news video program parsing), to the commonly used single-modal-feature based approach.

7. REFERENCES

[1] J.-H. Kuo and J.-L. Wu, "An Efficient Algorithm for Scene Change Detection and Camera Motion Characterization Using the approach of Heterogeneous Video Transcoding on MPEG Compressed Videos," *IEEE Conf. On Information, Communications & Signal Processing,* Oct 2001.

[2] ISO/IEC JTC1/SC29/WG11, CODING OF MOVING PICTURES AND AUDIO, m7691, MPEG-7 Visual part of experimentation Model Version 11.1

[3] B. Yeo, "Efficient processing of compressed images and video," Ph.D dissertation, Elect. Eng. Dept., Princeton Univ., Princeton, NJ, January 1996.

[4] Ullas Gargi, Rangachar Kasturi, and S. H. Strayer. " Performance characterization of video-shot-change detection methods," *IEEE Trans. on Circuits and Systems for Video Technol,*, Vol. 10, No. 1, pp. 1-13, Feb. 2000.

[5] Y. Wang, J. Huang, Z. Liu, and T. Chen, "Multimedia Content Classification using Motion and Audio Information," *Proc. of IEEE ISCAS' 97*, Vol. 2, pp.1488-1491, 1997.

[6] Y. Wang, J. Huang, and Z. Liu, "Multimedia Content Analysis using both audio and visual clues," *IEEE Signal Magazine*, pp. 12-36, 2000.

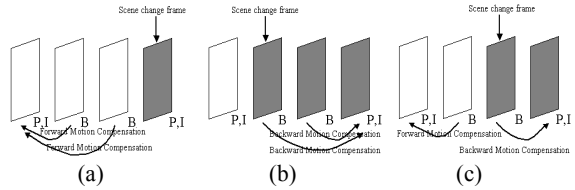

Fig. 1: (a) Scene change at P or I frame, (b) Scene change at front B frame, and (c) Scene change at rear B frame.

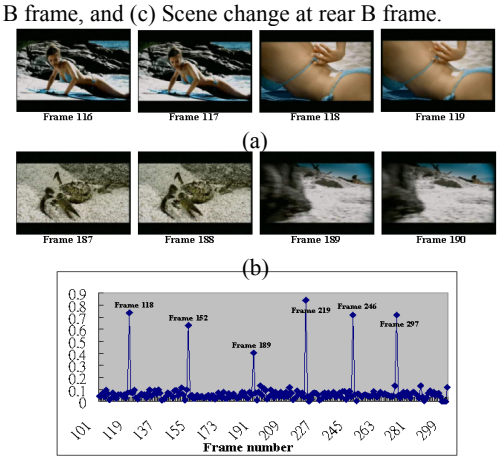

Fig. 2 : (a) Snapshots of a shot change duration from frames 116 to 119, (b) snapshots of scene change duration, with similar backgrounds, from frame 187 to 190, and (c) P_b's from frames 101 to 300.

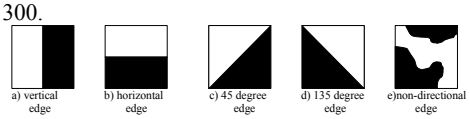

Fig. 3 : The five types of edges

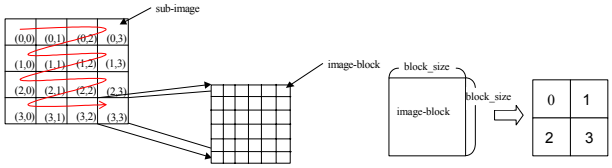

Fig. 4:Definitions of sub-image and image-block. Fig. 5: A Macroblock and its labelings.

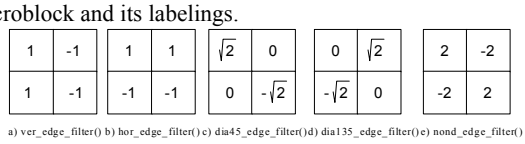

Fig. 6 : Filter coefficients for various edge detections.

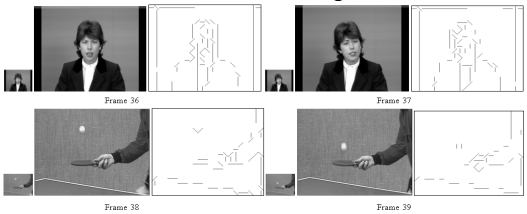

Fig. 7: The DC-images , original images, and the corresponding edge-images of the Chaire (upper) and the Tennis (lower) sequences.

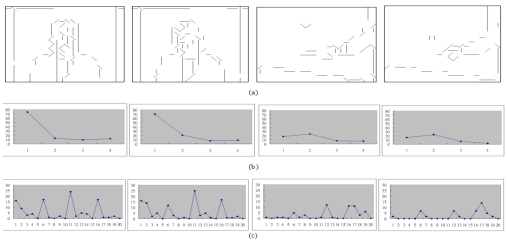

Fig. 8 : Examples of edge histogram (a) edge-image, (b) Global edge histogram, and (c) Local edge histogram.

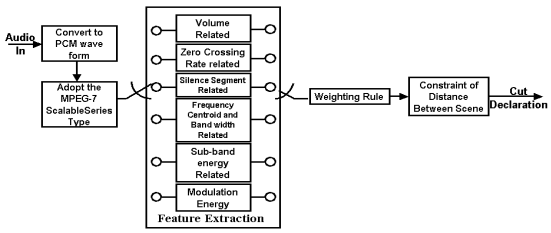

Fig. 9 : The block diagram of the audio analysis.

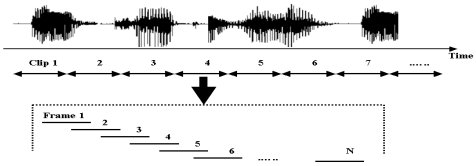

Fig. 10 : The definitions of clip and frame used in the audio feature analysis.

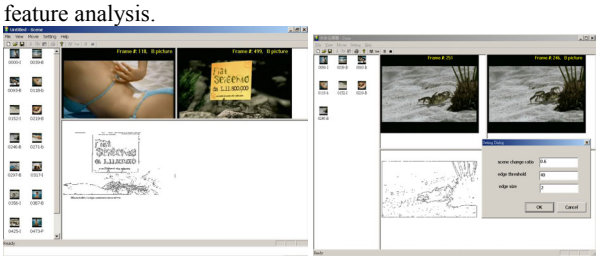

Fig. 11 : The snapshots of the scene change detection system in (a) automatic mode, and (b) semi-automatic mode, with the user-interactive feedback interface.

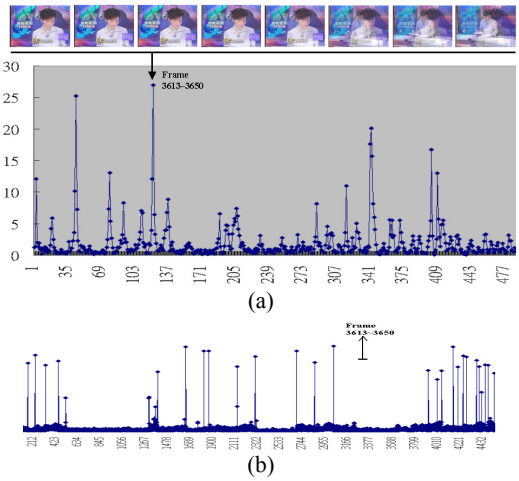

Fig. 12 : The snapshots of a news dissolve clip and the corresponding scene detections using (a) the audio and (b) the visual features, respectively.