COMBINED AUDIO AND VISUAL STREAMS ANALYSIS FOR VIDEO SEQUENCE SEGMENTATION

Jeho Nam and Ahmed H. Tewfik

Department of Electrical Engineering
University of Minnesota, Minneapolis, MN 55455
e-mail : jnam@ee.umn.edu, tewfik@ee.umn.edu

ABSTRACT

We present a new approach to video sequence segmentation into individual shots. Unlike previous approaches, our technique segments the video sequence by combining two streams of information extracted from the visual track with audio track segmentation information. The visual streams of information are computed from the coarse data in a 3-D wavelet decomposition of the video track. They consist of (i) information derived from temporal edges detected along the time evolution of the intensity of each pixel in temporally sub-sampled spatially filtered coarse frames, and (ii) information derived from the coarse spatio-temporal evolution of intra-frame edges in the spatially filtered coarse frames. Our approach is particularly matched to progressively transmitted video.

1. INTRODUCTION

Video sequence segmentation into individual shots is fundamental for content-based indexing, browsing and retrievals [1]. Most previous approaches to video segmentation [2], [3], [4] rely on differences of pixel intensities, frame-based histograms or motion discontinuities. Such techniques are not effective in dealing with shots that involve significant camera or object movements. In particular, they tend to detect false shot boundaries within such shots. Special effects, such as strobe lights that appear in a small number of frames, can also lead to false shot boundaries with such methods. In contrast, we found in extensive experiments that our multi-stream combined approach minimizes the number of false-detected shot boundaries while detecting all true boundaries.

Our multi-stream video segmentation approach relies on both the visual and audio tracks of the video sequence. Both tracks provide useful segmentation information. In particular, we show in this paper that by combining the results of several simple operations performed simultaneously on the visual and audio tracks, it is possible to obtain reliable segmentation of the video sequence at different levels of abstraction. Such segmentations cannot be achieved by analyzing a single track of information, e.g., the visual track only, unless much more complicated processing, and in particular, understanding, of that track is performed.

Note that the speech/audio stream of a video signal has been previously used for video indexing [5]. Speech information reflects some interesting events in video sequences such as sports games (e.g., crowd cheering, clapping and speak keyword, etc.). Here, we use it for segmentation purposes. The nature of the audio stream accompanying a video sequence typically changes from shot to shot. Furthermore, this information can be used to identify higher level boundaries, e.g., boundaries between TV advertisements. TV advertisements typically consist of multiple shots. Frame image analysis alone cannot detect boundaries between advertisements. On the other hand, each advertisement generally contains its characteristic speech and background music. Further, there is a short silence between adjacent advertisements. Therefore, the audio stream provides us with a useful clue for video segmentation.

2. MULTI-STREAM VIDEO SEGMENTATION TECHNIQUE

Let us now briefly describe the three basic sources of information that we use to segment a video sequence. Recall that the three sources of information are (i) the individual edges detected at a single pixel as a function of time, (ii) the spatio-temporal evolution of intra-frame edges and (iii) the segmentation of the audio stream that accompanies the video sequence.

2.1. Coarse shot boundary identification from temporal edges at individual pixels

The first step in our approach identifies candidate shot boundaries by using the fact that shot boundaries produce temporal edges in the time evolution of the intensity of most individual pixels in a frame or reduced frame. Therefore, the detection of a large number of temporal edges in the individual time evolution of the intensity at each pixel should be indicative of a shot boundary.

To reduce the complexity of this step, we begin by computing coarse frames from a 2-D wavelet transform of a *temporally sub-sampled* video sequence. The sub-sampling factor that we use depends on the frame rate of the underlying video signal. We have used in our experiments temporal sub-sampling factors of 2^3, 2^4 and 2^5 (Figure 1). For example, we typically sub-sample video sequences played at the rate of 12 frame/sec by a factor of 2^2 in the time domain prior to processing.

The spatially reduced and smoothed frames still have enough global information to be used for shot boundary detection [3]. Temporal edges along the time evolution of the intensity at each pixel are detected via a 1-D wavelet transform (as a function of time) of the intensity at each

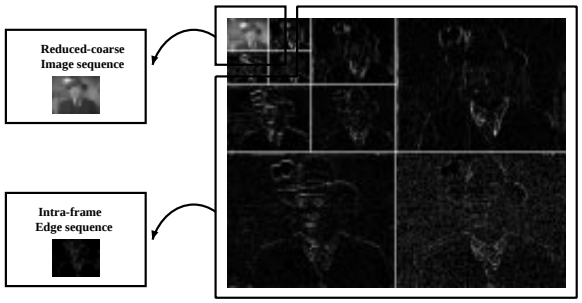


Figure 1. 2-D wavelet decomposition in the spatial domain

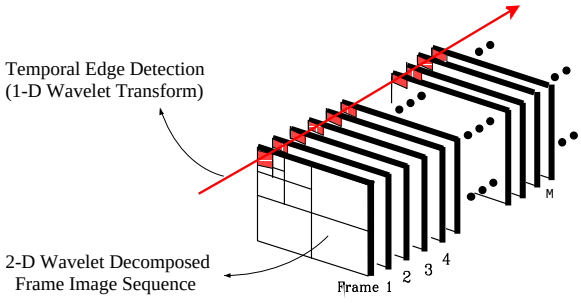


Figure 2. The temporal edge detection for shot boundary

pixel in the reduced image (Figure 2). Sharp edges are located by correlating the wavelet transform coefficients in adjacent scales [6]. By comparing the total number of single pixel temporal intensity edges detected within a reduced frame with a threshold, we obtain a first set of candidate shot boundaries.

Since our approach initially analyzes a temporally subsampled video sequence, it needs to refine its boundary estimates by focusing on windows consisting of the full-rate frames around the location of the detected candidate shot boundaries. In particular, we perform the procedure that we described above on each of the full-rate frames in these windows. This refinement step also eliminates false detections due to certain types of camera/object motions and to strobe lighting.

Note also that in addition to reducing the complexity of the boundary detection procedure (especially for long video sequences), our progressive (coarse-to-fine) approach for producing candidate shot boundaries can effectively deal with gradual shot transitions (e.g., dissolving, fading in and fading out). As can be seen from Figure 4, our method works well with both color or luminance information. With color images, we use the sum of the detection results of the Y, Cb, and Cr sequences respectively as a metric for identifying candidate shot boundaries. However, we have found that the luminance information is usually sufficient for shot boundary detection.

2.2. Spatio-temporal tracking of intra-frame edges

The main limitation of the first step in our procedure is that it is very sensitive to the camera/object motion and, in particular, to fast moving objects close to the camera (rapid panning [8]). Figure 5 shows part of a shot that contain such fast moving objects. To eliminate this sensitivity, we use a coarse spatio-temporal intra-frame (spatial) edge tracking procedure. The rationale behind this step is that intra-frame edges are not likely to move by much from frame to frame within a single shot.

Since this second step is meant to be a refinement step, our procedure need not use very accurate and complex edge detection procedures within a single frame. We use a simplified version of the edge detection approach of [7] operating on the spatially reduced and smoothed frames computed in step 1. Further, we only operate on the full-rate frames that are in a window around the candidate shot boundaries identified in step 1.

Edge tracking between successive frames is performed by evaluating differences in the numbers of edge pixels within blocks of pixels in the spatial edge maps that correspond to the reduced image frames. The size of the blocks depends on the amount of motion that we expect within a single shot. In particular, we simply compute the difference between the number of edge points in each fixed size-block in the edge maps corresponding to two successive reduced image frames. This coarse approach to edge tracking can effectively deal with fast moving objects as well as object rotations.

Only frames that display a large number of temporal edges within the individual pixel intensity level time histories as well as large differences between the number of intra-frame edge pixels within blocks in successive frames are retained as candidate shot boundaries. All other candidate shot boundaries computed in step 1 are eliminated.

Figure 6 shows the result of integrating steps 1 and 2. False boundaries detected by step 1 are completely eliminated. It is also easier to identify the real shot boundaries from the combined detection.

2.3. Audio Signal in Video

As mentioned in the introduction, the audio stream accompanying a video sequence can provide valuable information pertaining to shot segmentation. We analyze the audio stream using a procedure that is reminiscent of the 1-D edge detection approach of [6].

Specifically, we use a filter bank (Figure 3) to divide the audio signal into five unequal-widths subbands. We then detect sharp temporal variations in the power of the output of each subband filter between successive 80 msec segments of the signal. Temporal variations are detected using simple differences. Finally, we correlate the strength of the variations found in the five bands using simple multiplication. Figure 7(b) shows that sharp peaks in the resulting curve correspond to the boundaries of 3 advertisements sequence. Integrating with the result of image-based segmentation allows us to understand the structure of these advertisements sequences in detail.

3. RESULTS

We have tested our approach on video sequences digitized at 320 x 240 spatial resolution and about 12 frame/sec. The effectiveness of the first step of our procedure is revealed by an analysis of the 500 frames of the last climax scene of the

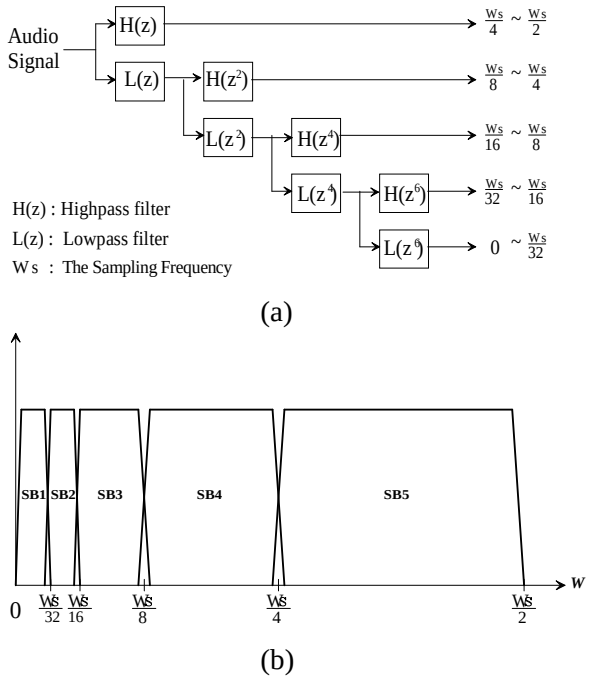


Figure 3. (a) Filter Bank (b) Five unequal-widths subbands in the frequency domain

movie "THE STING". Figure 4(a) and (b) show the results that we obtained using luminance and color-information respectively. All detected peaks correspond to actual shot boundaries. Further, no boundary was missed. The need to integrate the first two steps of our procedure appears when we analyze 360 frames from the movie, "BROADCASTING NEWS" are used. Figure 6(a) and (b) show the results of the pixel-based step and the coarse intra-frame edge tracking step respectively. The result of the combined two steps is shown in Figure 6(c). False detected boundaries due to camera panning are effectively eliminated.

The effectiveness of the audio segmentation step is illustrated in Figure 7. This figure corresponds to the segmentation of the audio stream extracted from the video sequence of 3 TV advertisements ("Burger King", "Cheese", "Folgers"). The audio signal is sampled at 22 KHz. Figure 7(a) and (b) show the result of steps 1 and 2 of our procedure only and those of step 3 (audio-based method) respectively.

4. CONCLUSION

We introduced a new video sequence segmentation technique which is combined three source of information. The first stream in particular identifies candidate shot boundaries. The latter two streams are used to sift these candidate shot boundaries. The novel combined audio and visual video sequence segmentation technique provides higher-level boundaries beyond the simple shot boundary information that can be obtained with other approaches that rely on the visual data stream only. Our approach is particularly suited to hierarchical video structuring, progressive refinement of temporal video segmentation and content-based video indexing and retrieval using joint audio-video signal processing.

REFERENCES

[1] S.W. Smoliar and H. Zhang, "Content-Based Video Indexing and Retrieval,", *IEEE Multimedia*, 1(2), pp. 62-72, 1994.

[2] H. Zhang, Y. Gong, S.W. Smoliar and S.Y. Tan, "Automatic Parsing of News Video," *International Conference on Multimedia Computing and Systems*, pp. 45-54, 1994.

[3] B.L. Yeo and B. Liu, "Rapid Scene Analysis on Compressed Videos," *IEEE Trans. on Circuits and Systems For Video Technology*, 5(6), Dec. 1995.

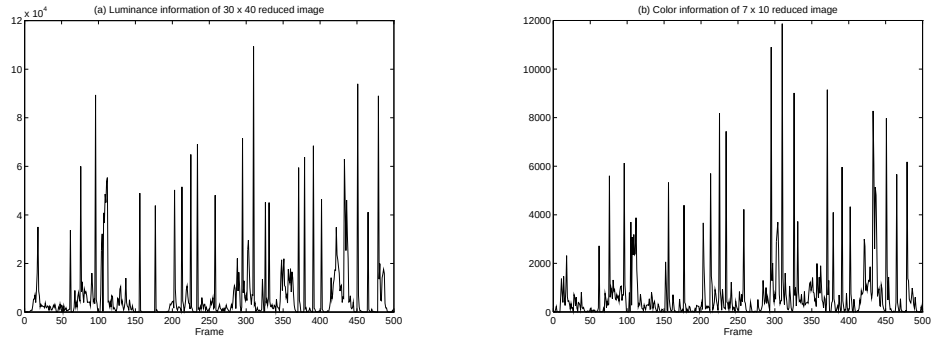
[4] P.R. Hsu and H. Harashima, "Detecting Scene Changes and Activities in Video Databases." *ICASSP 94*, vol. 5, pp. 33-36, Apr. 1994.

[5] Y.L. Chang, W. Zeng, I. Kamel and R. Alonso, "Integrated Image and Speech Analysis for Content-Based Video Indexing," *Proceedings of Multimedia '96*, pp. 306-313, Sep. 1996.

[6] Y. Xu, J.B. Weaver D.M. Healy and J. Lu, "Wavelet Transform Domain Filters: A Spatially Selective Noise Filtration Technique," *IEEE Trans. on Image Processing*, vol. 3, No. 6, Nov. 1994.

[7] S. Mallat and S. Zhong, "Characterization of Signals from Multiscale Edges," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 14, No. 7, Jul. 1992.

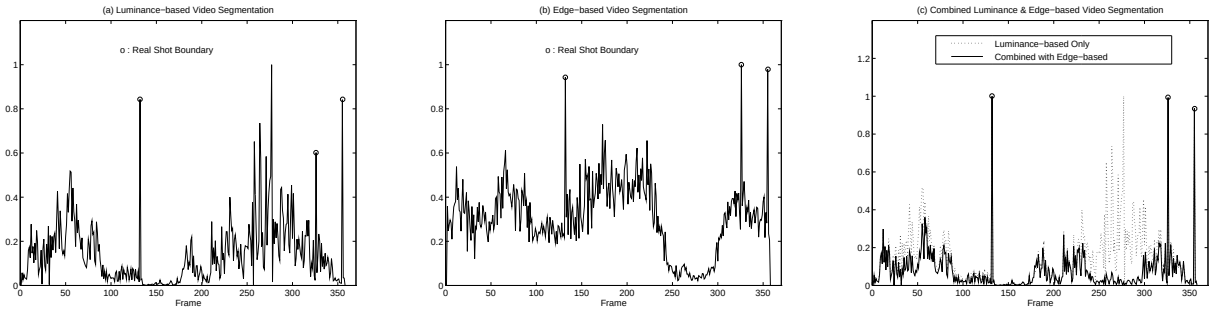
[8] Y. Tonomura, A. Akutsu, Y. Taniguchi, and G. Suzuki, "Structured Video Computing," *IEEE Multimedia*, 1(3), pp. 34-43, 1994.



(a) (b)

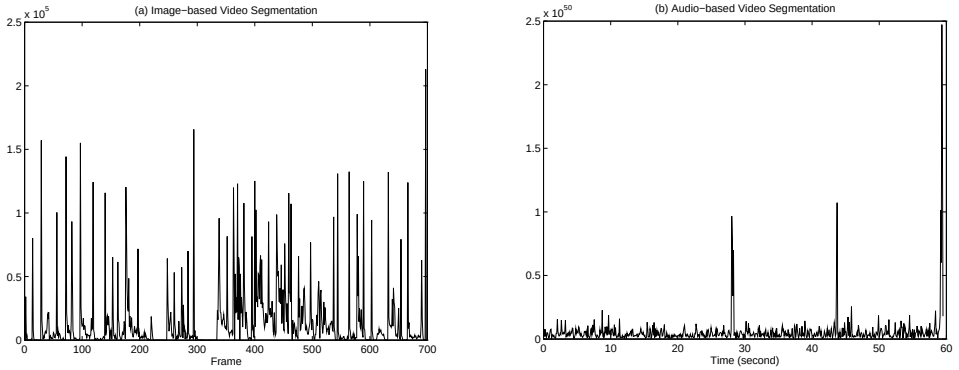
Figure 4. (a) 30 x 40 reduced image in Luminance (b) 7 x 10 reduced image in Color (Y, Cb, and Cr)

Figure 5. Shot of camera/object movement causing the false detections (Frame 263-278)



(a) (b) (c)

Figure 6. (a) Pixel intensity-based method (b) Intra-frame Edge-based method (c) Integrated method



(a) (b)

Figure 7. (a) frame image-based segmentation (b) audio-based segmentation