VIDEO SEGMENTATION WITH THE SUPPORT OF AUDIO SEGMENTATION AND CLASSIFICATION

Hao Jiang, Tong Lin and Hongjiang Zhang

Microsoft Research, China
No.5 Zhichun Road, Haidian District
Beijing 100084, China

ABSTRACT

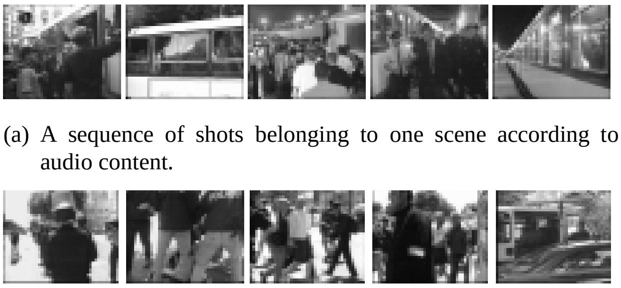
Video structure extraction is essential to automatic and content-based organization, retrieval and browsing of video. However, while many robust shot segmentation algorithms have developed, it is still difficult to extract scene structures or group shots into scenes. In this paper, we present a novel audio assisted video segmentation scheme, in which audio and color information is integrated in video scene extraction. A novel audio segmentation scheme is developed to segment audio tracks into speech, music, environmental sound and silence segments. A robust algorithm for shot grouping based on correlation analysis is also developed to further enhance the scene extraction accuracy.

1. INTRODUCTION

Video structure parsing is the process to extract construction units of video programs and it is essential to automatic and content-based organization and retrieval of video. There are usually two layers of construction units in video: shots and scenes (also often referred as story units). Therefore, a robust video structure parsing method should be able to segment a video program into these two layers. There have been many video parsing algorithms published. However, most of these algorithms utilize only visual information in the segmentation process [1,2]. Color histogram differences and motions between video frames or objects are the most commonly used features in shot segmentation algorithms. While it is very successful when using such features in shot segmentation of video, scene detection using such visual features alone poses many problems.

In general, a scene or story in video program consists of a sequence of related shots according to certain semantic rules. How to group a sequence of related shots into a semantically meaningful scene automatically based on video features is a challenging research topic. In a TV news broadcast, a high-level scene definition can be a news story which is the often separated by anchorperson. Therefore, TV news segmentation can be achieved by anchorperson spotting [3]. However, it is observed that to segment general video programs into semantic scenes, visual information alone cannot achieve satisfactory result; and audio track in a video can provide very useful and complementary semantics cues to aid scene detection.

There have been many works on integrating visual and audio information in video structure and content analysis. In[4], news broadcast is segmented and classified as news, basketball, football, commercial and advertisement segments by combining visual and audio features such that the final segmentation decision is made based on fusion result of both audio and visual boundaries. In[5], audio characteristic changes were described with likelihood ratio of cepstrum coefficient, and visual changes were represented by color difference and motions. These feature vectors were combined with a hidden Markov model to detect shot boundaries. In[6], an audio classification scheme based on heuristic rules was developed and was used to assist video segmentation. Despite many initial successes, integrating of audio and visual information in video structure parsing reminds a challenging research topic for at least following two reasons. First, just like many other data fusion problems, how to determine which features carries more weight in making final decision, especially when the two sources of information indicate to opposite directions. Second, how to measure correlation between consecutive shots is still an open question.



(a) A sequence of shots belonging to one scene according to audio content.

(b) A sequence of shots belonging to one scene according to shot color correlation.

Figure 1: Two examples of video scenes.

In this paper, we present an audio aided video scene segmentation scheme. In our system, audio segmentation and classification and shot correlation analysis based on color correlation are combined to cluster video shots into semantically related groups. We focus more on narrowly defined scene, that is, either a sequence video shots recorded in the same settings, or a sequence of shot anchored by a same anchor person in term of news broadcast. Figure 1 shows two examples, each represent one type of scenes defined in this work. To achieve such scene segmentation, first, audio segment boundaries are detected using a novel audio segmentation and classification algorithm that segments an audio stream into speech, music, environment sound and silence segments. Speech is further segmented into parts of different speakers. The shots within an audio segment are grouped together and market as related. Then, color correlation analysis between shots is performed and a so-called expanding window grouping algorithm is applied such that shots whose objects or background are closely correlated, for instance shots occurring in the same environment, are grouped. In other words, a sequence of shots will be grouped into a scene only when both visual content correlation analysis and audio segmentation detect a common scene boundary.

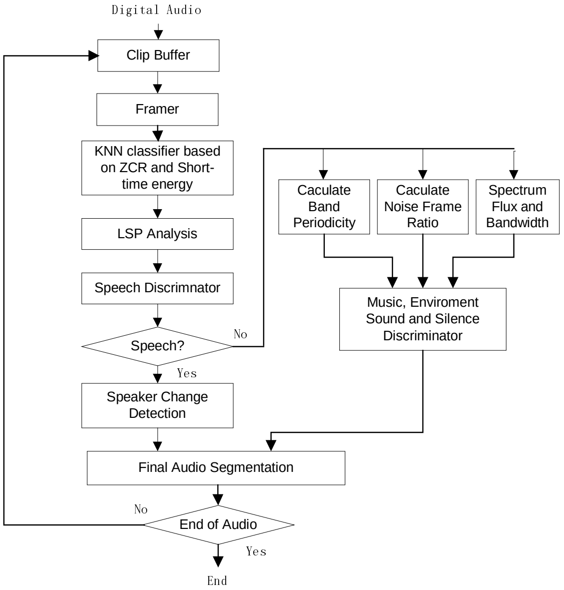
The paper is organized as follows. Audio classification and segmentation algorithm is discussed in Section 2. In Section 3, shot clustering based on the color correlation function is studied. In Section 4, integration of audio and video information is discussed. Experiment evaluation results are given in Section 5.

2. AUDIO ANALYSIS

The system block diagram of the proposed audio segmentation scheme is shown in Figure 2. In this system, the audio segmentation and speaker change detection module is parallel to the audio classification module. The output of audio classification and segmentation are combined together to obtain final audio segmentation result.

2.1 Audio Classification

The proposed audio classification scheme can be divided into two parts. First, discrimination between speech and non-speech segments is performed. That is, a KNN (K-Nearest Neighbor) classifier based on zero crossing rate and short time energy contour is used as a pre-classifier for speech and non-speech discrimination. Then, a GM-VQ (Gaussian Model-Vector Quantization) method based on *line spectrum pair (*LSP) divergence shape analysis [7] is used to refine the classification result and make the final decision. Second, non-speech segments are further classified into music and environment sound based on the audio periodicity and other features.



Digital Audio
Clip Buffer
Framer
KNN classifier based on ZCR and Short-time energy
LSP Analysis
Speech Discrimnator
Speech?
No
Yes
Speaker Change Detection
Caculate Band Periodicity
Caculate Noise Frame Ratio
Spectrum Flux and Bandwidth
Music, Enviroment Sound and Silence Discriminator
Final Audio Segmentation
No
End of Audio
Yes
End

Figure 2. Audio segmentation and classification block diagram

In our method, 10-order LP (Linear Predication) coefficients are analyzed in 25ms non-overlapping frames. Hamming window and band expansion are used. LP coefficients are converted to LSP's before further processing. Then, a GM-VQ method is developed for speech and non-speech discrimination. Gaussian model (GM) is used to describe the distribution of LSP's in a current audio clip. The estimated GM of the audio clips is then compared with a speech class codebook. The distance from input GM to the speech VQ codebook is defined as distance of the GM to the nearest code vector of speech class. That is,

$$D(X,Y) = tr[(C_X - C_Y)(C_Y^{-1} - C_X^{-1})] \qquad (1)$$

where C_X and C_Y are covariance matrixes of random variable *X* and *Y*, respectively. This distance has been proved to be a good measure for speaker identification [7]. If the distance is greater than a threshold, the input audio is classified as non-speech; otherwise, speech signal. Standard K-means VQ clustering is used to form the speech class codebook.

We use the noise free speech as the training data. Noise compensation is added when assigning the threshold. When white noise is added to pure speech, the average distance is nearly anti-proportional to SNR. This shows that the given distance is a good measure to the hearing effect. Two thresholds are used in the refinement procedure. If pre-classification result is speech class, 0db distance is used as the threshold. Otherwise 6db distance is used as the threshold.

After speech discrimination, non-speech class is further classified into music, environment sound and silence segments. In our scheme, silence detection is performed first based on short time energy in a one-second period. If the short time energy is lower than a threshold, the segment is classified as silence. To discriminate music and environment sound, noise/periodic signal discrimination based on correlation analysis is performed for each no-overlapping frame (25ms) in each audio clip (1 second). The ratio of noise frames in a given audio clip is defined as *noise frame ratio (NFR).* If the NFR of a frame is large, the audio clip tends to be environment sounds. However, it is usually not sufficient to only characterize the full-band periodicity of audio signals. Therefore, we introduce an additional feature, *band periodicity(BP),* based on sub-band correlation analysis. We choose bands 500~1000Hz, 1000~2000Hz, 2000~3000Hz, and 3000~4000Hz in computing *BP's.* DC-removed full-wave regularity signal is also used in the correlation analysis [8].

Furthermore, *spectrum flux*[9] and *band energy distribution* are also used in music/environment sound discrimination in the proposed scheme. A rule-based model is then used to discriminate music and environment sounds based on these two features. Periodicity is used as a first measure. If either of the NFR or four-band periodicity of an audio clip is lower than a predefined threshold, the clip is classified as noise-like environment sounds. Otherwise, band energy distribution and spectrum flux of the clip are checked. If the spectrum flux is greater than a threshold, or the energy in high band exceeds another threshold, the clip is classified as environment sounds. Strong periodicity environment sounds such as tone signal are discriminated by checking *band periodicity* and *spectrum flux.* Music is finally segmented out by excluding all above conditions. The thresholds used here are determined by experiments.

2.2 Audio Segmentation and Speaker Change Detection

A sliding window method is used in our scheme for audio segmentation and speaker change detection. That is, LSP's of two successive sliding windows of an audio clip are extracted and compared to determine if they belong to the same class. The distance measure used in the comparison is the divergence shape as (1) and is denoted as D_i, where i indicates the time stamp. To reliably detect audio segment boundaries, the following conditions are examined:

$$D_i > D_{i-1},\ D_i > D_{i+1} \text{ and } D_i > TH \qquad (2)$$

where TH is a threshold. The first two conditions guarantee a local peak exists. The last condition can prevent too low peaks from being detected. It is important to choose an appropriate window size in the segmentation process. Experiments show that a three-second window provides the good performance in terms of temporal resolution and appropriate feature smoothness.

3. SHOT CLUSTERING BASED ON COLOR CORELATION ANALYSIS

In video structure parsing, we first use a standard histogram comparison algorithm to detect shot boundaries [1]; then consecutive shots are grouped according to their correlations. A new method named *expanding window* is designed to group correlated shots into scenes. We assume that each scene should contain at least 3 shots and therefore, initially, the size of expanding window is 3. Every time a new shot is detected, the color correlation scores of this shot with the last three shots in are calculated. The maximum score in the window is denoted *v*. If *v* is greater than a threshold, the shot is absorbed into the current window and the window size is increased by 1. The threshold is dynamically defined as *mean-var*, where *mean* and *var* are the mean and variation of maximum score, respectively, in the expanding window. Otherwise, as illustrated by Figure 3, we consider one more shot, and start a new scene if and only if

$$\min\{left(0),left(1)\}<\min\{right(0),right(1)\} \qquad (3.1)$$

$$\max\{left(0),left(1)\}<\max\{right(0),right(1)\} \qquad (3.2)$$

where

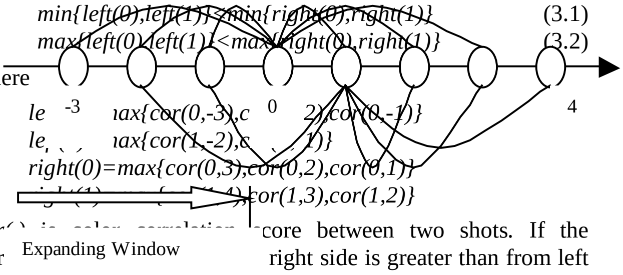
$$left(0)=\max\{cor(0,-3),cor(0,-2),cor(0,-1)\}$$

$$left(1)=\max\{cor(1,-2),cor(1,-1)\}$$

$$right(0)=\max\{cor(0,3),cor(0,2),cor(0,1)\}$$

$$right(1)=\max\{cor(1,4),cor(1,3),cor(1,2)\}$$

cor(.) is color correlation score between two shots. If the attr[illegible] right side is greater than from left side, a new scene starts. Otherwise, the current scene absorbs this shot.



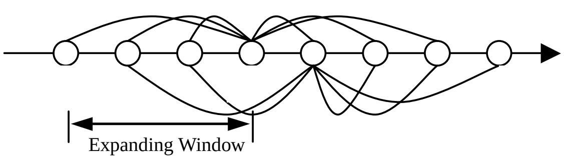


Figure 3. Expanding window shot grouping method. Shot 0 is the current shot.

It is essential to define correlation between two shots since shot grouping relies on such correlation scores. In contrast to many published methods, we use shot correlation rather than similarity in grouping shots into scenes since two shots belonging to a scene are usually highly correlated according to some rules. Therefore, we have introduced a color correlation score to measure shot correlation quantitatively.

Color correlation scores between two shots, *cor (.)* in (3), is calculated by dominant color object comparison and tracking between the two shots as following. First, pixels of each frame in one shot, or DC blocks in I frames when MPEG1/2 video are used, are projected into the HSV color space. Then, HSV color space is quantized with 10 values for H and U, 5 values for V (the lightness) based on the pixel distribution of the frame, thus forming a 3D color histograms of the frame. All dominant local maximum points are identified within a small neighborhood in the color histogram. A sphere surrounding each local maximum point in the color space is defined as a color object. Only dominant colors are counted in the 3D color histograms, because they capture the most significant color information of a frame and are more resilient to noise. It is worth noticing that we do not perform object segmentation in spatial domain, rather, we consider pixels falling into a dominant regions in the color space an object, which often not represent a spatial object in a frame.

Color objects in difference frames are tracked in the HSV color space, which allows lighting conditions to change gradually. If the centers of two color objects in two consecutive frames are close, these two color objects are recognized as the same color object. Such a color tracking process extract the temporal change of content in a shot, which is usually difficult to obtain with key-frame based representations of shot content [1]. Only the color objects having longer duration are retained as dominant object in a shot, which correspond to dominant objects or background in one shot. Therefore, our dominant color objects represent both structural content in a frame and temporal content in a shot.

To measure shot correlation, the mean size of each color object in a frame is weighted with duration in a shot and normalized within a shot. Dominant color objects that have a longer duration are more important and thus have higher weights. Finally, histogram intersection is made to get a correlation score.

4. INTEGRATION OF AUDIO ANALYSIS AND SHOT CORRELATION ANALYSIS

In this section, we discuss how to combine these two parts together for a more robust video segmentation scheme. In our system, the scene determination procedure can be divided into two stages, as illustrated in Figure 4. At the first stage, shots of a video sequence are clustered based on audio analysis. Audio breaks are first detected in one-second interval. When a shot break and an audio break are detected simultaneously within the one-second interval, the boundary of the sequence of shots is marked as a potential scene boundary. Generally audio break can be classified into two categories. One is the audio class change, such as changes from a speech segment to music or speech to environment sound. The other kind of audio breaks takes place when the speaker changes. Both of the two kinds of break are used in the video scene detection in our system. In news broadcast, such scene segmentation will result in many fragmented scenes since the process depends heavily upon audio segmentation. For instance, an interview scene between two persons may be broken into two or more scenes by this first step, since one shot could have only one person's speech, while another shot may only have other person's voice. Therefore, at the second stage of scene extraction, shot grouping using the expanding window method based on shot correlation analysis as presented in Section 3 is performed. In this step, a sequence of shots whose objects or background are closely correlated, for instance shots occurring in the same environment, are grouped by the color correlation algorithm. In other words, the potential scene boundaries detected in the first step by audio analysis pass will be marked as final scene boundaries when they coincide with that determined by the color correlation analysis.

5. EXPERIMENT RESULTS

We have chosen a set of TV news broadcasts as our test data, since news has explicit structures (ground truth) which will make the evaluation more objective. The test data set includes news broadcasting material in the MPEG7 content set CD17, each about half an hour long, and a CCTV(China Central TV) sport news program . The testing data consists of about 800 shots and 100 scenes. The audio track in the test set is sampled at 44.1kHz in two channels. In the experiment, stereo audio is firstly converted to mono-channel audio before further processing.

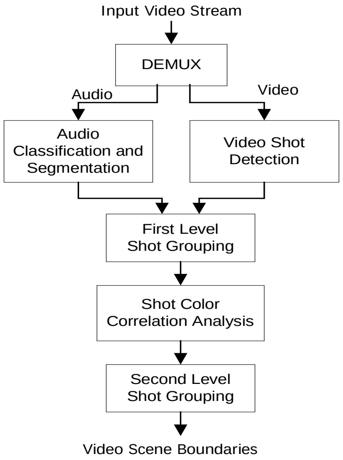

Input Video Stream
DEMUX
Audio
Video
Audio Classification and Segmentation
Video Shot Detection
First Level Shot Grouping
Shot Color Correlation Analysis
Second Level Shot Grouping
Video Scene Boundaries

Figure 4. Audio aided video segmentation system diagram

The performance is discribed with recall R and precision P, as

$$R = \frac{Correct\ Detection}{Total\ Boundaries} \quad (4.1)$$

$$P = \frac{Correct\ Detection}{CorrectDetection + False\ Detection} \quad (4.2)$$

In the evaluation, we first tested scene detection by combining only shot boundary detection and audio boundary detection. That is, we only perform the first step of our scene extraction process. In the experiment, the threshold in (2) is set low to achieve low missing detection rate. Experiments show that the first step scene extraction alone can not achieve satisfactory results. This is because the speech of the interviewee is often not very fluent and many audio breaks can be detected in one video scene: Video scenes are often fragmented, due mainly to speaker changes within a scene. This problem is overcome by the shot correlation analysis in the second step of scene extraction process.

Recall	Precision
91.9%	86.8%

Table 1. Scene grouping performance with audio break and color correlation analysis.

The overall video scene segmentation result based on the test set is listed in Table 1. As it is shown, the recall rate is very high. However, there are still many false detections (about 15%), often induced by the dramatic color changes during one news event.

6. SUMMARY

In this paper, we have presented an audio aided scheme for video scene structure parsing. A novel algorithm for the audio segmentation and classification, including an efficient audio segmentation and speaker change detection algorithm, is presented. Shot grouping method based on an expanding window algorithm is also discussed. Experiments show that audio break information can improve significantly the performance of video segmentation. Though the proposed scheme has been mainly test with TV news broadcast data, the scheme can be easily expanded to general video segmentation.

Future work to extend proposed video scene segmentation scheme include a more sophisticated audio-video fusion model in make final scene segmentation decision by integrating segmentation results from both audio and video content analysis.

7. REFERENCES

[1] H.J.Zhang, A.Kankanhalli, and S.W.Smoliar, *Automatic Partitioning of Full–Motion Video*, Multimedia Systems, Vol.1, No.1, pp.10-28, 1993

[2] P. Aigrain, H.J. Zhang, and D. Petkovic, *Content-based Representation and Retrieval of Visual Media: A State-of-the-art Review. Multimedia Tools and Applications,* 3(3):179-202, November 1996

[3] HongJiang Zhang, Yihong Gong, Smoliar S.W., Shuang Yeo Tan. *Automatic parsing of news video.* IEEE Proceedings of the International Conference on Multimedia Computing and Systems, 1994. pp. 45-54.

[4] Z. Liu, Y. Wang and T. Chen, *Audio Feature Extraction and Analysis for Scene Segmentation and Classification* Journal of VLSI Signal Processing Systems, June 1998

[5] J.S.Boreczky and L.D. Wilcox. *A Hidden Markov Model Frame Work for Video Segmentation Using Audio and Image Features.* Proceedings of ICASSP'98, pp.3741-3744, Seattle, May 1998.

[6] T. Zhang and C.-C. J. Kuo. *Video Content Parsing Based on Combined Audio and Visual Information.* SPIE 1999, Vol.IV, pp. 78-89.

[7] J. P. Campbell, JR. *Speaker Recognition: A Tutorial.* Proceedings of the IEEE, vol.85, no. 9, September 1997.

[8] A. V. McCree and T. P. Barnwell, *Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding,* IEEE Trans. on Speech and Audio Processing, vol. 3, No. 4, pp 242-250. IEEE, Jul 1995.

[9] E.Scheirer and M. Slaney, *Construction and Evaluation of a Robust Multifeature Music/Speech Discriminator*. Proc. ICASSP 97, vol II, pp 1331-1334. IEEE, April 1997