

VIDEO SEGMENTATION WITH THE ASSISTANCE OF AUDIO CONTENT ANALYSIS

Hao Jiang

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

Tong Lin

National Laboratory on Machine
Perception, Peking University
Beijing 100071, China

Hong-Jiang Zhang

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

ABSTRACT

Video structure extraction is essential to automatic content-based organization, retrieval and browsing of video. However, while many robust shot segmentation algorithms have been 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 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. Many video parsing algorithms have been developed. However, most of these algorithms have only utilized visual information in the segmentation process [1,2]. The most commonly used features in these shot segmentation algorithms are color histogram differences and motions between video frames or objects. While the use of such video features in shot segmentation has produced some good results, scene detection based exclusively on visual features poses many problems.

In general, a scene or story in a video program consists of a sequence of related shots according to certain semantic rules. How to automatically group a sequence of related shots into a semantically meaningful scene based on video features is a challenging research topic. In TV news broadcast, a high-level scene definition can be a news story that is often separated by a TV anchor. 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 results; and audio track in a video can provide very useful and complementary semantic cues to aid scene detection.

Many works have been done on integrating visual and audio information in video structure and content analysis. In [4], news broadcast is segmented and classified into news, basketball, football and commercials by combining visual and audio features such that the final segmentation decision is made based on the 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 differences 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 audio and visual information in video structure parsing remains a challenging research topic. There are two reasons for this. First, similar to many other data fusion problems, one need to determine which feature carries more weight in making final decision. This is especially true when the two sources of information do not indicate to the same direction. 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 of video shots recorded in the same setting, or a sequence of shots anchored by the same anchorperson in terms of news broadcast. Figure 1 shows two examples; each represents one type of scene 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 marked 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

Figure 2 shows the system block diagram of the proposed audio segmentation scheme.

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, short time energy contour and spectrum flux[7] 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)* analysis 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 audio periodicity and other features.

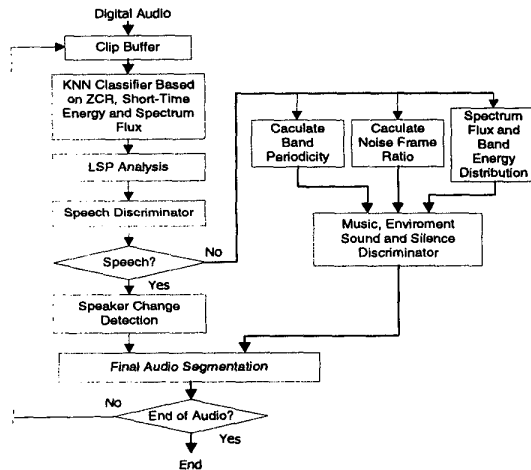


Figure 2. Audio segmentation and classification block diagram.

In our classification algorithm, the basic classification unit is an audio clip of one second in duration. Non-overlapping frames are divided in the audio clip, in which short time parameters such as short time energy (STE) and zero crossing rate (ZCR) are calculated. In speech and non-speech pre-classification we use features - high zero crossing rate ratio (HZCRR) which is the ratio of frames with ZCR greater than 1.5 times the average ZCR in one audio clip, low short time energy ratio (LSTER) which is the ratio of frames with STE less than 0.5 times the average STE in one audio clip and spectrum flux (SF) [7]. Based on this set of features, a KNN classifier is used in our scheme to perform speech and non-speech pre-classification.

To further improve the classification accuracy, we introduce LSP analysis and present a GM-VQ model to verify the pre-

classification result and make the final decision. In our proposed scheme, 10-order LP coefficients (LPC) are analyzed in 25ms non-overlapping frames. Hamming window is applied and a 15Hz bandwidth expansion is used. LP coefficients are converted to LSP before further processing. A Gaussian model is used in our scheme to describe the *pdf* (probability density function) of LSP vector in one audio clip. To increase the precision of estimation, the processing window size should be greater than the audio clip length. Then, the distance between the input audio and speech model is defined by

$$D = \text{tr}[(\hat{C}_{LSP} - C_{SP})(C_{SP}^{-1} - \hat{C}_{LSP}^{-1})] \quad (1)$$

where \hat{C}_{LSP} is the estimated LSP covariance matrix and C_{SP} is the LSP covariance matrix of speech model. In practical applications, a single speech model is found to be insufficient. In our scheme, a speech model codebook is generated using training speech data by the LBG (Linde-Buzo-Gray) algorithm. The size of speech codebook is 4 in our system. The dissimilarity of an audio clip from a speech signal is then defined by the distance between the LSP Gaussian model of the audio clip and the nearest code vector of the speech codebook. The final decision procedure for speech and non-speech classification is based on the distance of audio clip from speech codebook, denoted as D_s . Depending on the pre-classification result, two thresholds are used. If the pre-classification result is speech, a higher threshold, *Threshold1*, is selected; if D_s is greater than *Threshold1*, the audio clip is classified as non-speech, otherwise speech. If the pre-classification result is non-speech, D_s is compared against a lower threshold, *Threshold2*, and final classification is made.

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 average energy in the audio clip. If it is lower than a threshold, the segment is classified as silence. In discriminating music and environment sound, two new features have been introduced, *noise frame ratio (NFR)* and *band periodicity (BP)*. Noise/periodic signal discrimination based on correlation analysis is performed for each non-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)*. In addition to *NFR*, *BP* is derived based on sub-band correlation analysis. We choose bands 500-1000Hz, 1000-2000Hz, 2000-3000Hz, and 3000-4000Hz in computing *BPs*.

Furthermore, *spectrum flux* 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. Periodicity is used as a first measure. If either of the four-band periodicity of an audio clip is lower than a predefined threshold or the *NFR* is larger than a given threshold, the clip is classified as noise-like environment sound. 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 the above conditions. The thresholds used here are experimentally determined.

2.2 Speaker Change Detection

A sliding window method is used in our scheme for speaker change detection. LSPs extracted in the audio classification stage are buffered for two successive sliding windows. Then, separate Gaussian models are constructed for the LSP data in the two windows. The Gaussian models of early and later windows are compared to determine if they belong to the same speaker. Let C_i and C_{i-1} denote the LSP covariance matrix of audio clip i and audio clip $i-1$ respectively, we define,

$$D_i = \text{tr}[(C_i - C_{i-1})(C_{i-1}^{-1} - C_i^{-1})] \quad (2)$$

To reliably detect speaker change boundaries, the following conditions are examined:

$$D_i > D_{i-1}, D_i > D_{i+1} \text{ and } D_i > TH \quad (3)$$

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

3. SHOT CLUSTERING BASED ON COLOR CORRELATION 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 are calculated. The maximum score in the window is denoted as 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-std*, where *mean* and *std* are the mean and standard deviation of maximum score, respectively, in the expanding window. Otherwise, as illustrated in Figure 3, we consider one more shot, and start a new scene if and only if

$$\min\{\text{left}(0), \text{left}(1)\} < \min\{\text{right}(0), \text{right}(1)\} \quad (4.1)$$

$$\max\{\text{left}(0), \text{left}(1)\} < \max\{\text{right}(0), \text{right}(1)\} \quad (4.2)$$

where

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

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

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

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

$\text{cor}(\cdot)$ is color correlation score between two shots. If the attraction to shot 0 and 1 from right side is greater than from left side, a new scene starts. Otherwise, the current scene absorbs this 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.

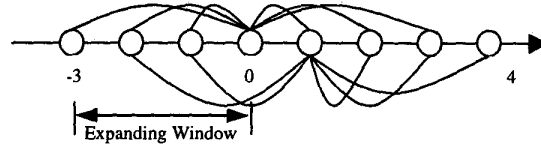


Figure 3. Expanding window shot grouping method.

Color correlation scores between two shots, $\text{cor}(\cdot)$ in (4), is calculated by dominant color object comparison and tracking between the two shots as follows. 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, thus forming a 3D color histogram 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 a spatial domain. Rather, we consider pixels falling into a dominant region in the color space an object, which often does not represent a spatial object in a frame.

Color objects in different 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 extracts the temporal change of content in a shot, which is usually difficult to obtain with key-frame based representations of shot content [1]. Only those color objects having a longer duration are retained, 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 intervals. When a shot break and an audio break are detected simultaneously within a one-second interval, the boundary of the sequence of shots is marked as a potential scene boundary. Generally audio breaks can be classified into two categories. One is the audio class change, such as changes from a speech segment to music or environment sound. The other kind of audio break takes place

when the speaker changes. Both kinds of breaks are used in 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 the 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

A set of TV news broadcasts has been chosen as our test data. News was chosen primarily because 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, each about half an hour long, and a CCTV(China Central TV) sport news program. The test data consists of approximately 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 first converted to mono-channel audio before further processing.

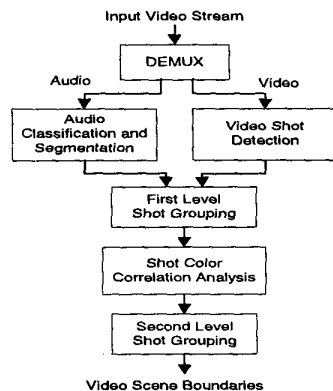


Figure 4. Audio aided video segmentation system diagram.

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

$$R = \frac{\text{Correct Detection}}{\text{Total Boundaries}} \quad (5.1)$$

$$P = \frac{\text{Correct Detection}}{\text{Correct Detection} + \text{False Detection}} \quad (5.2)$$

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 (3) 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 tested with TV news broadcast data, the scheme can be easily expanded to general video segmentation.

Future work to extend proposed video scene segmentation scheme includes a more sophisticated audio-video fusion model in making 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] H.J. 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] 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