AUDIO PCA IN A NOVEL MULTIMEDIA SCHEME FOR SCENE CHANGE DETECTION

Marios Kyperountas, Zuzana Cernekova, Constantine Kotropoulos, Marios Gavrielides, Ioannis Pitas

Aristotle University of Thessaloniki, Department of Informatics
Artificial Intelligence and Information Analysis Laboratory
Box 451, GR-54124 Thessaloniki, Greece

ABSTRACT

A novel scene change detection algorithm is proposed in this paper that exploits both audio and video information. Audio frames are projected to the eigenspace and their distance from a reference noise eigenframe is calculated. An analysis is presented that explains why this subspace favors scene cut detection. Video information is used to align audio scene change indications with neighboring shot changes in the visual data, and accordingly to reduce the false alarm rate. Moreover, video fade effects are identified and used independently in order to track scene changes. The detection technique was tested on newscast videos provided by the TRECVID 2003 video test set [1]. The experimental results show that the aforementioned methods used to process the audio and video information complement each other well in tackling the scene change detection problem.

1. INTRODUCTION

The ever-growing amount of digital information has created a critical need for the development of assisting data management algorithms. These algorithms typically aim to group data into meaningful categories, index those categories and provide options for fast browsing and retrieval to the user. Video shot and scene detection is essential to automatic content-based segmentation. A video shot is a collection of frames obtained through a continuous camera recording; similar background and motion typify the set of frames within a shot. While shots can be thought of as the basic units of video grouping, they do not provide effective non-linear access to the information data in terms of their semantic representation. Consequently, higher-level units are used, the video scenes; each scene is defined as a sequence of related shots according to certain semantic rules.

For the task of detecting semantically meaningful scenes, the amount of information from the visual component alone does not provide satisfactory results [2]. In many cases, shot changes are misclassified as scene changes. Typically, an efficient way to separate a video scene change from a shot change is through the use of the audio characteristics. A scene change is most often accompanied by a significant change in the audio characteristics, while a shot change is not. At this point, the definition of a third semantic unit is essential; an audio scene is a semantically consistent audio segment that can be distinguished by basic characteristics of the sound that dominate the signal.

In regards to broadcasting and commercial video, it is well known that program and movie directors use audio not just to convey critical information, such as dialogues, but also to maintain and stimulate the interest of the audience; often it is predetermined for different scenes to be accompanied by dissimilar sounds. In addition to this, highly dissimilar audio characteristics naturally occur in programs such as news shows where the background noise in the news studio most often characteristically varies from the background, or environment, noise of pre-recorded news clips.

Naturally, when both the visual and audio information are incorporated more efficient algorithms can be developed. The actual integration of the two types of information, in order to reach a final decision on scene change detection, can indeed be a challenging problem due to the versatile nature of videos. Unsynchronized visual and audio data, in terms of content, further complicates the data fusion problem.

A scene change detection method that employs both audio and video information is proposed in this paper. The algorithm heavily depends on the audio data in order to detect a scene cut. The expectation for the success of an audio-based detection approach can largely be attributed to the habitual presence of similar background noise, such as car, crowd or room noise throughout a scene. In order to 'visualize' transition periods from one scene to the next, as well as the scene cut points, we use principal component analysis and show why it is an appropriate method for this detection problem. Furthermore, a mathematical analysis illustrates which components of the audio information enhance the scene detection capabilities of our algorithm.

In addition to audio, video information was also utilized. The process initially used visual data to extract shot change information, which is employed in order to synchronize audio scene change indications with the corresponding changes in the video data. Moreover, by considering a time limit in which both an audio scene and a video shot change should occur, the false alarm rate is reduced. Finally, video information was further used as an independent scene change indicator, simply by identifying specific video effects that are commonly employed during scene changes, namely fades: a fade is a transition of gradual diminishing or heightening of visual intensity.

The proposed method was tested on newscast videos provided by the well-established TRECVID 2003 video test set. The experimental results are presented and analyzed in section 6 and conclusions are drawn.

2. PREVIOUS WORK

Video scene boundary detection, and video structure parsing in general, is a research field that has received much attention from the research community in recent years. Various multimedia approaches are discussed next.

In [3], audio was distributed into four pre-selected classes, and this information was later combined with the probability value for a visual cut detection that segmented the video into shot segments. In order to track scene changes, information from both the video and audio classifiers was used in order to determine if a correlation between adjacent shots exists. The work introduced in [4] used a finite-memory model to independently segment the audio and video data into scenes; then two ambiguity windows were used to merge the audio and video scenes. In [5], scene change detection was based on the dissimilarity index values of audio, color and motion features. Thus, audio-visual breaks were employed to segment the data. In [6], low and mid-level audiovisual features were statistically analyzed according to genre characteristics. These features were directly obtained in the MPEG compressed domain. Then a Linear Machine Decision Tree classifier was used in order to classify each shot into predetermined genre sets. In [7], visual effects such as dissolves or fades were tracked and used in order to nominate possible scene boundaries. Then, the audio data, that were located around a short time frame from where a shot change occurs, were analyzed using the average power of sub-bands. Obviously, this algorithm can suffer from audio and video synchronization problems.

3. MODELLING THE PROBLEM

As mentioned before, scenes can be obtained by grouping semantically correlated shots. However, this definition is quite vague as different people can use different criteria to determine the borders of a particular scene. To make matters worse, different principles are used to define scenes for TV-news programs, talk shows, documentaries or Hollywood movies. Consequently, in order to test an algorithm, it is crucial to define a specific model that avails clear criteria in determining a scene. Several papers try to define models for scene detection, mainly in the field of TV-news, where simple and effective models can be defined [8]. News headings, graphics of the station's logo, anchorperson shots and prerecorded news videos are some of the most common scenes set for news shows.

During news story reports, or commercial breaks, sequential shots are often completely uncorrelated with respect to one another, in terms of visual content; therefore there is a high risk of misclassifying shot as scene changes. Our method remedies this problem by tracking audio scene changes; this is done in a single stage without having to classify the audio frame content into different classes, as it is done in most audio segmentation algorithms. Additional scene changes are tracked by identifying fade in or out effects in the video stream.

4. EIGENFRAMES FOR DETECTION

After studying several news shows it is concluded that scene changes can efficiently be detected by considering the audio background information, or background noise. For example, we can distinguish a scene that represents a report of a soccer game from a scene where a journalist is reporting from a busy street by comparing the characteristic differences between crowd and traffic noise that are present throughout the two scenes. If the variations between the various types of noise are sufficiently large, e.g. due to low signal-to-noise ratios, then they can be considered to be principal modes of variation for the problem of scene segmentation. In order to mathematically discover these modes we decided to use principal component analysis (PCA) where each audio frame is projected to an eigenspace and an eigenframe is created.

PCA uses second order statistics and calculates the principal components of a distribution of audio frames. It aims to extract a subspace in which the variance is maximized and the reconstruction mean square error is minimized by finding the basis vectors of a low-dimensional subspace, as a set of orthonormal eigenvectors [9].

4.1. Significance of background noise

Let $P_1, P_2, ..., P_L$ be the a-priory probabilities of L different noise classes that correspond to L different scenes, or groups. Each class can be modeled by the distribution function $g_i(x)$ with mean μ_i and variance σ_i^2. The grand mean of all groups, is found by

$$\overline{w} = \sum_{i=1}^{L} P_i \mu_i \,. \tag{1}$$

The overall background noise variance is defined as

$$\sigma_w^2 = \sum_{i=1}^{L} P_i \sigma_i^2 + \sum_{i=1}^{L} P_i \left(\mu_i - \overline{w}\right)^2 . \tag{2}$$

The first term in (2) represents the within group variance, while the second corresponds to the between group variance, which is the weighted sum of the squared distances between the means of each group and the grand mean. In regards to PCA, uniform noise, or a single group, is represented in high order principal components; however, when a number of different groups sequentially corrupt the data, the variations in background noise, especially during segments with low signal-to-noise ratios, should be explained in lower order principal components, or in large eigenvalue subspaces. As a result, PCA can help us visualize a separation between the different scenes.

4.2. Calculating eigenframes

The audio stream is segmented into M successive and non-overlapping vectors, the raw audio frames. The mean of these frames is found and subtracted from each frame, thus creating the difference frames. Let $\mathbf{X}$ be a matrix that consists of a set of M difference frames, $\mathbf{X} = \begin{bmatrix} x_1 & x_2 & \dots & x_M \end{bmatrix}$, where each $\boldsymbol{x_i}$ is a $\rho \times 1$ vector. These frame vectors are used in an outer product operation that forms the covariance matrix $\mathbf{C}$ as such:

$$\mathbf{C} = \frac{1}{M} \sum_{i=1}^{\mathrm{M}} x_i x_i^{\mathrm{T}} = \frac{1}{M} \mathbf{X}\mathbf{X}^{\mathrm{T}} . \tag{3}$$

For a typical frame, ρ will be a large number, requiring high computational intensity. Fortunately, a much smaller $M \times M$ matrix problem can be solved in order to determine the eigenvectors of $\mathbf{C}$ by calculating the inner product matrix [9]. The normalized eigenframes can be found by

$$\mathbf{E} = \mathbf{X}\mathbf{Q}\boldsymbol{\Lambda}^{-\frac{1}{2}} , \tag{4}$$

where $diag(\mathbf{\Lambda}) = [\lambda_1, \lambda_2, ..., \lambda_{M'}]$ and $\mathbf{Q}$ is an $M \times M$ matrix that holds M' significant eigenvectors, associated with the largest eigenvalues $\{\lambda_i\}_{i=1}^{M'}$, where $M' < M < \rho$.

4.3. Detecting scene changes

It was observed that in newscast videos usually similar audio background characteristics are present throughout the duration of a scene; moreover, during a scene transition period only background noise is present. As a result, scene transitions are represented in lower levels of signal variance due to the absence of signals with large variations, such as speech, and also to typical drops in signal energy. The varying characteristics from one type of noise to the next enable a relatively precise detection of scene changes; right after the point of where a scene change occurs the signal variance is subject to experience a notable increase. This trend will subsequently be maintained by a typical increase in the signal energy and variance, as foreground signals will begin to appear again in the audio stream.

In order to visualize these trends various background noise frames were collected from several news video sequences and the average noise frame was calculated. Then, the mean of the M raw frames was subtracted and the result was next projected to the eigenspace. Thus, a reference frame was created that enhanced the ability to isolate pure background noise frames. Subsequently, the Euclidian distance between the reference noise frame and the eigenframes was stored in a distance vector.

The downward and upward trends, indicating varying contributions to the overall variance, were clarified by applying a median filter on the distance vector. The median filter gives an additional advantage as it rejects momentary lapses in signal energy, e.g. during speech segments or near other foreground signals with high variance, as scene changes. As a result, scene transitions were represented in the lower dips of the smoothed distance vector and the scene change points could be found by locating the minimum points of those dips. Figure 1 shows the smoothed distance vector, created in one of our experiments, and the location of correctly detected scene changes.

5. VIDEO SHOT BOUNDARY DETECTION

In order to detect the various video shots, the mutual information and the joint entropy between two successive frames was calculated separately on each of the RGB channels, as is proposed in [10].

The mutual information between the frames f_t and f_{t+1} for the red channel with N gray levels is defined as

$$I^R_{t,t+1} = -\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} C^R_{t,t+1}(i,j) \log \frac{C^R_{t,t+1}(i,j)}{C^R_{t,t+1}(i) C^R_{t,t+1}(j)}, \qquad (5)$$

where $C^R_{t,t+1}(i,j)$ corresponds to the probability a pixel having grey level i in frame f_t having grey level j in frame f_{t+1}.

The joint entropy for this channel is defined as

$$H^R_{t,t+1} = -\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} C^R_{t,t+1}(i,j) \log C^R_{t,t+1}(i,j). \qquad (6)$$

The total mutual information and entropy is the corresponding sum from all three channels.

A small value of the mutual information $I_{t,t+1}$ leads to a high probability of having a cut between frames f_t and f_{t+1}. In order to detect possible shot cuts, an adaptive thresholding approach was employed, as in [10].

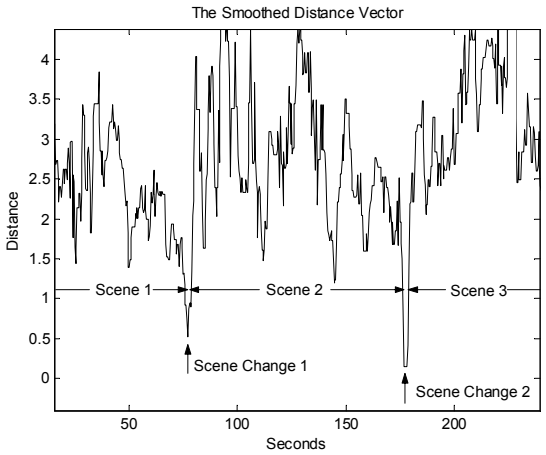

Figure 1: Scene changes in the smoothed distance vector

In order to detect video fades, the joint entropy criterion was employed, which measures the amount of the information that is carried between frames; its value decreases during fades. For our purposes, only the points at which a fade in effect started or a fade out effect ended were identified, by simply applying a threshold. In the cases where both effects were found to exist within a small time frame t_d then the shot cut was set at the average temporal point. Shot cuts due to fade effects were classified as scene change indicators.

6. INTEGRATING AUDIO AND VIDEO INFORMATION

After processing the audio track in order to locate potential scene change points and the video track to find possible shot changes, including fade effects, we describe a process that integrates all information in order to enhance the scene change detection capabilities. This process consists of two steps:

1) Define a temporal window W_I, with length that corresponds to time duration t_I selected such as to represent the maximum allowed time between the audio and the video data to convey information about semantically corresponding events. Whenever a scene change indication from the audio information does not match with any shot change indication from the video information, within a time frame t_I, then the scene change is rejected as a false alarm; otherwise, the audio scene change point is aligned with the closest video shot change point and a scene cut is designated at that location.

2) Shot cut points that correspond to video fade effects are classified as scene cuts.

7. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed technique's accuracy was tested on two half-hour-long newscast videos from the well established reference video test set TRECVID 2003. The video had a frame rate of 29.97 fps and each frame was resized to half of the original resolution, at 176-by-132 in order to speed up calculations. The audio track

was converted to an 8-bit mono channel with a sampling rate of 11.25 KHz. Audio frames were extracted, with each one corresponding to roughly one third of a second, or to 10 video frames. The ground truth for video scene change points was provided by human observations.

We define the algorithm that was used to extract scene change information from the audio data as process '*P1*'. The eigenvectors associated with the largest 35 eigenvalues were used to project the data. Let process '*P2*' be the algorithm that was used in order to detect scenes from the video data, by tracking fade effects; for this procedure t_d was set at 5 seconds. Process '*P3*' is defined as the method that we used to integrate audio scene and video shot detection information, as is described in section 6. The value of the parameter t_1 was set to 3 seconds. Finally, '*P4*' is a process that collects the detection indications from processes *P2* and *P3*. *P4* represents our proposed solution to the scene change detection problem.

Table 1 presents an analysis of the capabilities of these algorithms to detect different categories of scenes. The table verifies the high efficiency of the integration process of the audio and visual data that we followed. As expected, the audio-based algorithm performs poorly in finding the separation points between commercials, which are typically more 'refined' TV segments that are created on a very similar formula; usually music or various sound effects are inserted that override environment or background noise. Fortunately, fade effects separate commercials thus the detection process is well complemented. The false alarm incidents for processes *P1*, *P2*, *P3* and *P4* are correspondingly 32, 8, 14 and 22; more than half of the false alarm incidents in the audio-based process were corrected by the audiovisual integration process.

In order to evaluate the performance of the segmentation method the '*Recall*' and '*Precision*' measures, inspired by receiver operating characteristics in statistical detection theory, were used. The *Recall* measure, also known as the true positive function or sensitivity, corresponds to the ratio of correct experimental detections over the number of all true detections. The *Precision* measure corresponds to the accuracy of the method considering false detections and it is defined as the number of correct experimental detections over the number of all experimental detections. Table 2 illustrates how our method evaluates based on these two criteria. The multimedia approach proposed in this paper for the detection of scene changes presents promising results on both accounts.

Scene Category	Scene Ch. per Category	Detected Scene Changes		
		P1 or P3	*P2*	*P4*
News Headings	2	2	0	2
Studio Transition	2	0	0	0
Anchorperson	20	16	8	16
Prerecorded Vid.	16	12	0	12
Commercials	36	10	34	36
Total	**76**	**40**	**42**	**66**

Table 1: Analysis of detection by scene category.

Evaluation of Scene Change Detection Capabilities	*Recall* (%)	*Precision* (%)
P1	52.6	55.6
P2	55.2	84.0
P3	52.6	74.1
P4	**86.8**	**75.0**

Table 2: Evaluation results for each process.

8. CONCLUSION

A novel method was proposed in order to detect scene changes in videos. The detection scheme integrates indications from the audio and video data in order to produce higher detection and lower false alarm rates. The method was tested on newscast videos and results are very promising. For this specific test additional work can be done, using news-specific knowledge, in order to secure higher detection results.

9. ACKNOWLEDGEMENTS

This work has been supported by the Commission of the European Communities in the framework of the Methods for Unified Multimedia Information Retrieval (MOUMIR) project, (HPTN-CT-2000-00111).

10. REFERENCES

[1] NIST, *TREC Video Retrieval Evaluation*, http://www-nlpir.nist.gov/projects/trecvid/.

[2] H. Jiang, T. Lin, and H.J. Zhang, "Video segmentation with the assistance of audio content analysis," *in Proc. IEEE Int. Conf. Multimedia and Expo (ICME2000),* pp. 1507-1510, 2000.

[3] C. Saraceno and R. Leonardi, "Audio as support to scene change detection and characterization of video sequences," *in Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP 1997),* vol. 4, pp. 2597-2600, 1997.

[4] H. Sundaram and S.-F. Chang, "Video scene segmentation using video and audio features," *in Proc. IEEE Int. Conf. Multimedia and Expo (ICME2000),* pp. 1145-1148, 2000.

[5] J. Huang, Z. Liu, and Y. Wang, "Integration of audio and visual information for content-based video segmentation," *in Proc. IEEE Int. Conf. Image Processing (ICIP1998)*, vol. 3, pp. 526-530, 1998.

[6] M. Sugano et. al., "Shot genre classification using compressed audio-visual features," *in Proc. IEEE Int. Conf. Image Processing (ICIP2003)*, Barcelona, Spain, 14-17 Sep, 2003.

[7] A. Yoshitaka, and M. Miyake, "Scene detection by audio-visual features," *in Proc. IEEE Int. Conf. Multimedia and Expo (ICME2001)*, pp. 49-52*,* 2001.

[8] M. De Santo et. al., "Dialogue scenes detection in MPEG movies: a multi-expert approach," in *Lecture Notes in Computer Science*, vol. 2184, pp. 192-201, September 2001.

[9] M. Turk, "A random walk through eigenspace," *IEICE Trans. Information and Systems*, Vol.E84-D, No.12, pp. 1586-1595, Dec. 2001.

[10] Z.Cernekova, C.Nikou, and I.Pitas, "Shot detection in video sequences using entropy-based metrics," *in Proc. 2002 IEEE Int. Conf. Image Processing (ICIP2002)*, Vol. 3, pp. 421-424, 2002.