Scene Change Detection using Audiovisual Clues

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

Artificial Intelligence and Information Analysis Laboratory,
Department of Informatics, Aristotle University of Thessaloniki,
Box 451, GR-54124 Thessaloniki, Greece
Email: {mkyper, zuzana, costas, marios, pitas}@zeus.csd.auth.gr

Abstract

A novel scene change detection algorithm is proposed in this paper that exploits both audio and video information. Audio frames are projected to an eigenspace that aims to 'discover' the changes in the audio track caused by the variations of background audio. An analysis explains why the selected subspace is suitable for detecting scene changes, even more so than the large-eigenvalue subspace used in [1]. Video information is used to align audio scene change indications with neighboring shot changes in the visual data by considering certain timing restrictions, 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 [2]. The experimental results show that the aforementioned methods that are used to process the audio and video information complement each other well when tackling with 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. Such algorithms typically aim to group data into meaningful categories, index these 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 and is considered to be the basic unit of video grouping. Video scenes provide effective non-linear access to the information data; each scene is defined as a sequence of related shots according to certain semantic rules. Unlike shot changes, a scene change is most often accompanied by a significant change in the audio characteristics.

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 usually characteristically varies from the background/environment, noise of pre-recorded news clips.

A scene change detection method that employs both audio and video information is proposed in this paper. The expectation for the success of the audio-based part of the proposed 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 project the audio frames to a particular eigenspace with the expectation to distinguish, as much as possible, the variations in background audio that occur during scene transitions.

In addition to audio, visual information was also utilized 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 8 and conclusions are drawn, in section 9.

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 have developed, some of which are discussed next.

The work introduced in [3] 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 [4] the audio feature extraction relates to the detected video shots. Unlike most audio feature-based segmentation algorithms, the content of the audio is no longer relevant; a scene change is indicated simply based on the differences of the audio features from the corresponding adjacent shots. In [5], audio was distributed into four pre-selected classes, and this informa-

tion was later combined with the probability value for a visual cut detection that segmented the video into shot segments. To find scene changes, information from both the video and audio classifiers was used in order to determine if a correlation between adjacent shots exists. 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.

3. The TV-News Model

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 [7]. News videos have a rather definite structure that makes them well suitable for content analysis. News headings, graphics of the station's logo, anchorperson shots and prerecorded news videos are some of the most common scenes that are set for news shows.

4. Eigenframes for Detection

Scene changes in news shows 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 the 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 extracts a subspace in which the variance is maximized and the reconstruction mean square error is minimized by finding the orthonormal basis vectors, or eigenvectors, of a low-dimensional subspace [8].

4.1. Significance of background noise

Let $P_1, P_2, ..., P_L$ be the a-priori probabilities of L different noise classes that correspond to L different scenes, or groups. Each class can be modelled 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 larger eigenvalue subspaces. As a result, PCA can help us visualize a separation between the different scenes.

4.2. Eigen analysis and scene change detection

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, $\mathbf{X} = \left[\boldsymbol{x}_1 | \boldsymbol{x}_2 | ... | \boldsymbol{x}_M\right]$, 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 of $\mathbf{X}$, $\mathbf{C}$. In order to find the eigen-components of $\mathbf{C}$ a $\rho \times \rho$ problem needs to be solved. However, for a typical frame, ρ will be a large number, requiring high computational intensity.

The significant, or non-zero, eigenvalues of $\mathbf{C}$ are equal to the eigenvalues of the inner product matrix [8], $\mathbf{R}$, created as such:

$$\mathbf{R} = \frac{1}{M} \mathbf{X}^{\mathrm{T}} \mathbf{X} \,. \tag{3}$$

As a result, the much smaller $M \times M$ matrix problem is solved. The normalized eigenframes can be found by

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

where $diag(\boldsymbol{\Lambda}) = \left[\lambda_1, \lambda_2, ..., \lambda_M\right]$ consists of the eigenvalues of $\mathbf{R}$ that are associated with the eigenvectors matrix $\mathbf{Q}$.

Usually, only a subset of the possible $min(\rho, M-1)$ non-zero eigenframes contained in $\mathbf{E}$ is retained for projecting the data to eigenspace. Any audio frame, i.e. $\boldsymbol{\upsilon}$, can be projected to this eigenspace by

$$\mathbf{v} = \mathbf{E}^{\mathrm{T}} \boldsymbol{\upsilon} \,. \tag{5}$$

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 reappear in the audio stream.

In order to visualize these trends various background

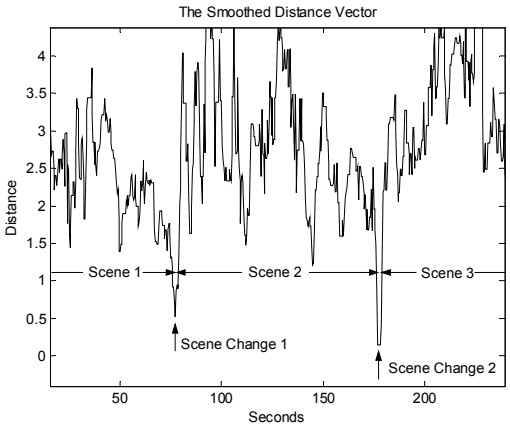

Figure 1: *Scene changes indications from the audio.*

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. Refining the Audio Information

In order to optimize our algorithm a 'filtering' procedure was applied to the original audio signal in order to remove signal components that either do not contribute to the detection process' success or downgrade its performance. When dealing with PCA, a convenient method for 'filtering' or denoising comes by refining the selected set of projection components. Let us consider a decomposition of the audio signal to background audio, transmission noise and the remaining information signal. The proposed scene change detection method uses the expectation that the largest variations in background audio occur in temporal neighborhoods that 'bridge' two scenes where the information signal's energy and variance reside at low levels.

Obviously, transmission noise should be removed from the audio track, as it can only decay the detection performance. Subsequently, it is noted that the principal modes of variation of the audio track are primarily related to variations in the information signal. Naturally, the variations of the information signal are also represented in lower modes of signal variance, but to a lower extend. Lastly, we note that lower eigenvalue subspaces, but not low enough to represent transmission noise, are more indicative to the presence of background noise and in addition represent components of the information signal with smaller overall contributions to the signal variance.

Based on these preliminary observations, it is expected that the optimum selection for the projection components should include neither the highest nor the lowest modes of variance. In order to ensure that the largest part of the foreground signals' and transmission noise representation is largely suppressed in the 'frame space', only the eigenvectors that are associated with the eigenvalues relating to 50-90% of the total variance are selected to comprise the eigenspace. For comparison purposes it is noted that in [1] the eigenvectors associated with the largest 35 eigenvalues, corresponding to roughly 25% of the total variance, were used to project the data.

6. 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 [9].

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

$$I^R_{t,t+1} = -\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} C^R_{t,t+1}(i,j)\log\frac{C^R_{t,t+1}(i,j)}{C^R_{t,t+1}(i)C^R_{t,t+1}(j)}, \quad (6)$$

where $C^R_{t,t+1}(i,j)$ corresponds to the probability a pixel having grey level i in frame f_t and grey level j in frame f_{t+1}, in the red channel. The same figure can be measured in the other channels as well.

The joint entropy for channel R is defined as

$$H^R_{t,t+1} = -\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} C^R_{t,t+1}(i,j)\log C^R_{t,t+1}(i,j) \quad (7)$$

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 [9].

In order to detect video fades, the joint entropy criterion was employed, which measures the amount of 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. If more than one effect was 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.

7. 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_1, 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.

8. Experimental Results and Discussion

The accuracy of the proposed technique 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 '*AudioPCA*'. The eigenvectors associated with the ordered eigenvalues that related to 50-90% of the total variance were used to project the data. Let process '*VideoFades*' 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 to 5 seconds. Process '*AudiovisualAlign*' is defined as the method that we used to integrate audio scene and video shot detection information, as is described in section 7. The value of the parameter t_1 was set to 3 seconds. Finally, '*AuViFuse*' is a process that collects the detection indications from processes *VideoFades* and *AudiovisualAlign*. *AuViFuse* represents our proposed solution to the scene change detection problem. There is a huge advantage in treating audio data false alarm rates by using the actual video data for the problem of video scene change detection. The false alarm incidents for processes *AudioPCA*, *VideoFades*, *AudiovisualAlign* and *AuViFuse* are correspondingly 27, 8, 10 and 18. More than half of the false alarm incidents in the audio-based process were corrected by the *AudiovisualAlign* integration scheme.

In order to evaluate the performance of the segmentation method the '*Recall*' and '*Precision*' measures were used. The *Recall* measure 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 1 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.

Finally, the *AudioPCA* process was significantly improved by projecting the audio frames to the subspace that was proposed in section 5 than to the one that was used in [1]. The improvement in the *Recall* rate reaches 5.3%, while the *Precision* rate improved by 6.4%.

	Recall (%)	*Precision* (%)
AudioPCA	57.9	62.0
VideoFades	55.2	84.0
AudiovisualAlign	57.9	81.5
AuViFuse	**84.2**	**78.1**

Table 1: *Evaluation results for each process.*

9. Conclusion

A novel multimedia method that was implemented in order to detect scene changes in videos was presented. The detection scheme integrates indications from the audio and video data in order to produce higher detection and lower false alarm rates. An eigenspace that excluded the lowest and highest order components was used, in order to better discover the variations caused by background audio. 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.

Acknowledgements

This work has been partially supported by the VISNET Network of Excellence, funded by the Commission of the European Communities (Contract No. 506946).

References

[1] M. Kyperountas, Z. Cernekoca, C. Kotropoulos, M. Gavrielides and I. Pitas, "Audio PCA in a novel multimedia scheme for scene change detection," *in Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP2004)*, Montreal, May, 2004 – accepted for publication.

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

[3] 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.

[4] S-C. Chen et. al., **"**Scene change detection by audio and video clues," *in Proc. IEEE Int. Conf. Multimedia and Expo (ICME2002)*, vol. 2, pp.365 -368, 2002.

[5] 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.

[6] M. Sugano et. al., "Shot genre classification using compressed audio-visual features," *in Proc. IEEE Int. Conf. Image Processing (ICIP2003)*, vol. 2, pp. 17-20, Barcelona, Spain, 14-17 Sep, 2003.

[7] R 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.

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

[9] 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.