Scene Change Detection Based on Audio-Visual Analysis and Interaction

Sofia Tsekeridou, Stelios Krinidis, Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki
Box 451, 54006 Thessaloniki, Greece
Tel: +30 31 996361, Fax: +30 31 996304
{sofia,skrinidi,pitas}@zeus.csd.auth.gr

Abstract. A scene change detection method is presented in this paper, which analyzes both auditory and visual information sources and accounts for their inter-relations and coincidence to semantically identify video scenes. Audio analysis focuses on the segmentation of the audio source into three types of semantic primitives, i.e. silence, speech and music. Further processing on speech segments aims at locating speaker change instants. Video analysis attempts to segment the video source into shots, without the segmentation being affected by camera pans, zoom-ins/outs or significantly high object motion. Results from single source segmentation are in some cases suboptimal. Audio-visual interaction achieves to either enhance single source findings or extract high level semantic information. The aim of this paper is to identify semantically meaningful video scenes by exploiting the temporal correlations of both sources based on the observation that semantic changes are characterized by significant changes in both information sources. Experimentation has been carried on a real TV serial sequence composed of many different scenes with plenty of commercials appearing in-between. The results are proven to be rather promising.

1 Introduction

Content-based video parsing, indexing, search, browsing and retrieval have recently grown to active research topics due to the enormous amount of unstructured video data available nowadays, the spread of its use as a data source in many applications and the increasing difficulty in its manipulation and retrieval of the material of interest. The need for content-based indexing and coding has been foreseen by ISO/MPEG that has introduced two new standards: MPEG-4 and MPEG-7 for coding and indexing, respectively [1].

In order to efficiently index video data, one must firstly semantically identify video scenes. The term *scene* refers to one or more successive shots combined together because they exhibit the same semantically meaningful concept, e.g. a scene that addresses the same topic although many shots may be involved. The term *shot* denotes a sequence of successive frames that corresponds to a single camera start and end session. Scene characterization should be content- and

search-dependent. The task of semantic scene identification is rather tedious and no automatic approaches have been reported to date. Usually, low-level processing of the visual data is initially undertaken. Shot boundary detection, i.e., temporal segmentation, is performed and analysis of detected shots follows [2, 3, 4]. Results are enhanced and higher level semantic information can be extracted when other information sources are analyzed, such as aural or textual ones [5, 6, 7, 8]. It is evident that semantic characterization can only be achieved with annotator intervention or by imposing user-defined interaction rules and domain knowledge.

A scene change detection method is presented in this paper which analyzes both auditory and visual sources and accounts for their inter-relations and synergy to semantically identify video scenes. The audio source is analyzed and segmented into three types of semantic primitives: silence, speech and music. Further analysis on speech parts leads to the determination of speaker change instants, without any knowledge on the number or the identity of speakers and without any need for a training process. The video source is processed by a combination of two shot boundary detection methods based on color frame and color vector histogram differences in order to efficiently detect shot boundaries even under various edit effects and camera movement. Combination of the results extracted from single information sources leads to grouping a number of successive shots into a scene according to whether they are in-between two successive speaker change instants or the same music segment accompanies them, or there are long duration silence segments before and after them. If further speaker alternation is attempted, such scenes can also be partially identified as commercials or events or dialogue scenes. In Sect. 2, the tools for low-level audio analysis and segmentation are summarized, while in Sect. 3, video segmentation into shots is reported. In Sect. 4, scene identification by combining both aural and visual information based on interaction rules is presented. Simulation results on a TV serial sequence of around 15 min duration containing many commercials are reported in Sects. 5. Finally, conclusions are drawn in Sect. 6.

2 Audio Analysis

Audio analysis aims at segmenting the audio source into three types of semantic primitives: silence, speech and music. Further processing on speech segments attempts to locate speaker change instants. Segmentation and speaker change identification are achieved by low-level processing methods. In the sequel, the term *audio frame* refers to the shortest in duration audio part used in short-time audio analysis, whereas the term *segment* refers to a group of a variable number of successive frames pre-classified to one of the three predefined audio types.

Initially, silence detection is performed to identify silence periods and discard them from subsequent analysis. Silence frames are audio frames of only background noise with a relatively low energy level and high zero crossing rate (ZCR) compared to other audio signal types. In order to distinguish silence from other audio signal types, the average magnitude M_t and zero crossing rate Z_t

functions of an M-sample audio frame $x_t(n)$, $n = 0, \ldots, M-1$, are exploited [9]:

$$M_t = \sum_{k=0}^{M-1} |x_t(k)| \tag{1}$$

$$Z_t = \frac{1}{2M} \sum_{k=1}^{M} |\mathrm{sgn}(x_t(k)) - \mathrm{sgn}(x_t(k-1))| \tag{2}$$

$t = 0, .., N-1$, where N is the total number of audio frames. Non-overlapping audio frames of 10msec duration are employed. A convenient approach to robust speech-silence discrimination is end point detection [9], which determines the beginning and end of words, phrases or sentences so that subsequent processing is applied only on these segments. Average magnitude and ZCR thresholds are chosen relative to the background noise characteristics of an apriori known audio interval, its average magnitude and ZCR functions being $M_{t,n}$ and $Z_{t,n}$ respectively. The average magnitude thresholds used by endpoint detection are set equal to:

$$\begin{aligned} M_{thr,up} &= E[M_t] \\ M_{thr,low} &= \max(M_{t,n}) \end{aligned} \tag{3}$$

The ZCR threshold is set equal to: $Z_{thr} = \max(Z_{t,n})$. Such a threshold selection proves to be robust and endpoint detection is satisfactorily performed. Boundaries of words, phrases or entire sentences are well estimated, a useful outcome that is subsequently exploited for audio segmentation and characterization.

Music detection is further performed to discriminate speech from music. Music segments are audio parts having significant high frequency content, high ZCR, different periodicity, compared to speech segments (voiced parts), and usually long duration. The latter is attributed to the fact that music does not usually exhibit silence periods between different successive parts leading to a long audio segment. Thus, in order to distinguish speech from music, four criteria are used: an energy measure, the ZCR, a correlation measure in the frequency domain that attempts to detect periodicity, and, finally, segment duration. Energy, M_t, and ZCR, Z_t, values are evaluated by (1) and (2), respectively, on audio frames of 10 msec duration located inside the current segment S_i, $i = 1, \ldots, N_S$, where N_S is the total number of detected segments other than silence ones. Subsequently, segment-based mean values and variances of M_t and Z_t are estimated, i.e.:

$$\begin{array}{cc} \mu_{M_{S_i}} = E[M_t | t \in S_i] & \mu_{Z_{S_i}} = E[Z_t | t \in S_i] \\ \sigma^2_{M_{S_i}} = E[(M_t - \mu_{M_{S_i}})^2] & \sigma^2_{Z_{S_i}} = E[(Z_t - \mu_{Z_{S_i}})^2] \end{array} \tag{4}$$

Their quotient is considered more discriminative for recognizing music from speech:

$$QM_{S_i} = \frac{\mu_{M_{S_i}}}{\sigma^2_{M_{S_i}}} \tag{5}$$

$$QZ_{S_i} = \frac{\mu_{Z_{S_i}}}{\sigma^2_{Z_{S_i}}} \tag{6}$$

Because both long-term (segment-based) energy and ZCR mean values are higher for music than speech. Besides, due to the existence of voiced and unvoiced parts in speech, long-term variance values of speech segments are expected to be higher than musical ones. In order to take advantage of the long duration periodicity of music, a frequency-based correlation metric C_t is defined between the magnitude spectrums of successive non-overlapping audio frames of 30msec located in segment S_i, $i = 1, \ldots, N_S$:

$$C_t = \frac{1}{M} \sum_{k=0}^{M-1} |\mathcal{F}(x_t(k))| \cdot |\mathcal{F}(x_{t-1}(k))| \tag{7}$$

where $\mathcal{F}(.)$ denotes the Fourier transform operator. If the signal is periodic, x_t and x_{t-1} will have almost identical spectra, thus leading to a high correlation value. Correlation is performed in frequency due to the fact that the Fourier transform remains unaffected by time shifts. In the case of music, C_t is expected to attain constantly large values within S_i. On the other hand, speech, characterized by both periodic (voiced) and aperiodic (unvoiced) parts, will have alternating high and low values of C_t within S_i. Thus, segment-based mean values of C_t, $\mu_{C_{S_i}} = E[C_t | t \in S_i]$ are considered to be adequately discriminative for detecting music. $\mu_{C_{S_i}}$ is expected to be higher for music segments than speech ones. Finally, the segment duration d_{S_i}, $i = 1, \ldots, N_S$, is also employed. Each of the metrics QM_{S_i}, QZ_{S_i}, $\mu_{C_{S_i}}$ and d_{S_i} are individually good discriminators of music. Global thresholding with thresholds:

$$T_M = E[QM_{S_i}] + \frac{\max(QM_{S_i}) - \min(QM_{S_i})}{2} \tag{8}$$

$$T_Z = \frac{7}{8} E[QZ_{S_i}] \tag{9}$$

$$T_C = 2E[\mu_{C_{S_i}}] \tag{10}$$

$$T_d = 5\text{sec} \tag{11}$$

respectively, leads to individual but suboptimal detection of music segments. Combination of these results in order to enhance music detection is based on the validity of the expression:

$$((QM_{S_i} > T_M) \text{ OR } (d_{S_i} > T_d)) \text{ OR } ((QZ_{S_i} > T_Z) \text{ AND } (\mu_{C_{S_i}} > T_C)) \tag{12}$$

If (12) is true for a segment S_i, then this segment is considered to be a music segment. Otherwise, it is declared as a speech segment. It is noted that audio segments, that may contain both speech and music, are expected to be classified according to the most dominant type.

Speech segments are further analyzed in an attempt to locate speaker change instants. In order to do that, low-level feature vectors are firstly extracted from voiced pre-classified frames only [9], located inside a speech segment. Voiced-unvoiced discrimination is based on the fact that unvoiced speech sounds exhibit significant high frequency content in contrast to voiced ones. Thus, the energy

distribution of the frame signal is evaluated in the lower and upper frequency bands (the boundary is set at 2kHz with a sampling rate of 11kHz). High to low energy ratio values greater than 0.25 imply unvoiced sounds, that are not processed further. For audio feature extraction in voiced frames, the speech signal is initially pre-emphasized by an FIR filter with transfer function $H(z) = 1 - 0.95z^{-1}$. Speech frames are used of 30msec duration each with an overlap of 20msec with each other. Each frame is windowed by a Hamming window of size M. Finally, the mel-frequency cepstrum coefficients (MFCC), $\mathbf{c} = \{c_k, k \in [1, p]\}$, are extracted per audio frame [10]. p is the dimension of the audio feature vector. The aim now is to locate speaker change instants used later on for enhancing scene boundary detection. In order to do that, firstly feature vectors of successive K speech segments $S_{K_0}, \ldots, S_{K_0+K}$, are grouped together to form sequences of feature vectors of the form [11]:

$$X = \{\underbrace{\mathbf{c}_1, \ldots, \mathbf{c}_{L_{S_{K_0}}}}_{S_{K_0}}, \underbrace{\mathbf{c}_1, \ldots, \mathbf{c}_{L_{S_{K_1}}}}_{S_{K_0+1}}, \ldots, \underbrace{\mathbf{c}_1, \ldots, \mathbf{c}_{L_{S_{K_0+K}}}}_{S_{K_0+K}}\} \tag{13}$$

Grouping is performed on the basis of the total duration of the grouped speech segments. This is expected to be equal or greater than 2sec, when assuming that only one speaker is talking. Consecutive sequences X and Y of feature vectors of the form (13), with Y composed of K' speech segments and defined by:

$$Y = \{\underbrace{\mathbf{c}_1, \ldots, \mathbf{c}_{L_{S_{K_0+K+1}}}}_{S_{K_0+K+1}}, \ldots, \underbrace{\mathbf{c}_1, \ldots, \mathbf{c}_{L_{S_{K_0+K+K'}}}}_{S_{K_0+K+K'}}\} \tag{14}$$

are considered, having a common boundary at the end of S_{K_0+K} and the beggining of S_{K_0+K+1}. The similarity of these two sequences is investigated by firstly evaluating their mean vectors, μ_X, μ_Y, and their covariance matrices, Σ_X, Σ_Y, and then defining the following distance metric:

$$D_t(X, Y) = (\mu_X - \mu_Y)\Sigma_X^{-1}(\mu_X - \mu_Y)^T + (\mu_Y - \mu_X)\Sigma_Y^{-1}(\mu_Y - \mu_X)^T \tag{15}$$

D_t is evaluated for the next pairs of sequences X, Y, until all speech segments have been used. The immediate next pair is constructed by shifting the X sequence by one segment, i.e. starting at S_{K_0+1}, and re-evaluating numbers K and K', so that the constraint on total duration is met. This approach is based on the observation that a speaker can be sufficiently modeled by the covariance matrix of feature vectors extracted from his utterances. Furthermore, the covariance matrices evaluated on feature vectors coming from utterances of the same speaker are expected to be identical. Adaptive thresholding follows to locate speaker change instants. Local mean values on a $1d$ temporal window W of size N_W are obtained, without considering the value of D_t at the current location t_0:

$$D_m = E[D_t|_{t \in W, t \neq t_0}]. \tag{16}$$

D_{t_0} is examined to specify whether it is the maximum value of those ones inside the temporal window (possibility of a speaker change instant at t_0). If this is

the case and $D_{t_0}/D_m \geq \epsilon$, where ϵ is a constant controlling the strictness of thresholding, a speaker change instant is detected at t_0. Speaker change instants are a clue for shot or even scene breaks. The method may be further investigated to identify speaker alternation and identify dialogue shots/scenes.

3 Video Analysis

Video analysis involves the temporal segmentation of the video source into shots. Shot boundary detection is performed by combining distance metrics produced by two different shot boundary detection methods. Such a dual mode approach is expected to lead to enhanced shot boundary detection results even under significant camera or object movement or camera effects, thus overcoming the drawbacks of the single modalities in some cases.

The first method estimates color frame differences between successive frames. Color differences, $FD(t)$, are defined by:

$$FD_t = \frac{1}{3N_X \times N_Y} \sum_{\mathbf{x}} ||\mathbf{I}(\mathbf{x};t) - \mathbf{I}(\mathbf{x};t-1)||_1 \tag{17}$$

where $\mathbf{I}(\mathbf{x};t) = [I_r(\mathbf{x};t)I_g(\mathbf{x};t)I_b(\mathbf{x};t)]^T$ represents the vector-valued pixel intensity function composed of the three color components: $I_r(\mathbf{x};t)$, $I_g(\mathbf{x};t)$ and $I_b(\mathbf{x};t)$. By $||.||_1$ the L_1-vector norm metric is denoted. $\mathbf{x} = (x, y)$ spans the spatial dimensions of the sequence (each frame is of size $N_X \times N_Y$) whereas t spans its temporal one. Frame differencing is computationally intensive but seldom any limitations on the processing time are imposed when the task is performed off-line. In order to detect possible shot breaks, the adaptive thresholding approach used for detecting speaker change instants in Sect. 2 is adopted. Such window-based thresholding offers the means of adaptive thresholding according to local content and proves flexible and efficient in gradual camera movements, significantly abrupt object or camera movements, and simple edit effects as zoom-ins and outs (no false positives, over-segmentation is avoided). Abrupt changes are directly recognised.

The second method evaluates color vector histograms of successive frames and computes their bin-wise differences. Summation over all bins leads to the metric that is used for shot break detection. Histogram-based methods are robust to camera as well as to object motion. Furthermore, color histograms are invariant under translation and rotation about the view axis and change only slowly under change of view angle, change in scale, and occlusion. However, histograms are very sensitive to shot illumination changes. To overcome this problem and make the method more robust, our approach operates in the HLS color space and ignores luminance information. Thus, instead of using HLS vector histograms (3-valued vector histograms), the method uses HS vector ones (2-valued vector histograms). Luminance conveys information only about illumination intensity changes, while all color information is found in the hue and saturation domain. Usually, hue contains most of the color information. Saturation is examined and

used to determine which regions of the image are achromatic. In order to evaluate HS vector histograms, the hue range $[0^o, 360^o]$ is divided in 32 equally-spaced bins h_i, $i = 1, \ldots, 32$, and the saturation range $[0, 1]$ in 8 equally-spaced bins s_j, $j = 1, \ldots, 8$. Vector bins are constructed by considering all possible pairs of the scalar hue and saturation bins, leading thus to a total number of 256 vector bins $\mathbf{hs}_k = (h_i, s_j)$, $k = 1, \ldots, 256$. Such an approach translates to a 256 uniform color quantization for each frame. The color vector bin-wise histogram $H(\mathbf{hs}_k; t)$ for frame t is computed by counting all pixels having hue and saturation values lying inside the considered vector bin $\mathbf{hs}_k$ and dividing by the total number of frame pixels. The histogram differences, HD_t, are then computed for every frame pair $(t-1, t)$, by:

$$HD_t = \frac{1}{N_X \times N_Y} \sum_{k=1}^{256} ln(||H(\mathbf{hs}_k; t) - H(\mathbf{hs}_k; t-1)||_1) \tag{18}$$

where k is the vector bin index. By $||\cdot||_1$, the L_1-vector norm metric is denoted. Each frame is of size $N_X \times N_Y$ and t is a temporal spatial dimension of the sequence. Histogram differencing is computationally intensive. In order to detect possible shot breaks, our approach firstly examines the validity of the expression:

$$2 * E[HD_t] \leq \frac{max(HD_t) - min(HD_t)}{2}. \tag{19}$$

If it is true, then the sequence is composed by a unique shot without any shot breaks. In the opposite case, the adaptive thresholding technique introduced for detecting speaker change instants is also employed here, leading to efficient shot break detection. Abrupt changes are directly recognized, but the method is also satisfactorily efficient with smooth changes between different shots.

However, both frame difference and color vector histogram based methods, employed separately, exhibit limited performance, than when combined together. Thus, fusion of single case outcomes is proposed. Specifically, the difference metrics (17) and (18) are multiplied to lead to an overall metric:

$$OD_t = FD_t \cdot HD_t \tag{20}$$

that is adaptively thresholded later on for shot cut detection. Despite its simplicity, such multiplication amplifies peaks of the single case metrics, possibly corresponding to shot cuts, while it lowers significantly the remaining values. The same adaptive thresholding method is employed here as well, leading to enhanced detection compared to the single case approaches. Strong object motion or significant camera movement, edit effects, like zoom ins-outs, and in some cases dissolves (dominant in commercials) are dealt with. Over-segmentation never occurs.

4 Audio-Visual Interaction: Scene Boundary Detection and Partial Scene Identification

Our aim is to group successive shots together into semantically meaningful scenes based on both visual and aural clues and using interaction rules. Multimodal

interaction can serve two purposes: (a) enhance the "content findings" of one source by using similar content knowledge extracted from the other source(s), (b) offer a more detailed content description about the same video instances by combining the content descriptors (semantic primitives) of all data sources based on interaction rules and coincidence concepts. Temporal coincidence due to the temporal nature of video data is a very convenient tool for multimodal interaction.

The combination of the results extracted from the single information sources leads to the grouping of a number of successive shots into a scene according to a number of imposed constraints and interaction rules. It is noted here that, given the results of the presented aural and visual segmentation algorithms, only scene boundaries are determined, while scene charecterization, e.g dialogue scene, can only be partially performed in some cases. Further analysis on those and on additional rules may lead to overall scene characterization. Shot grouping into scenes and scene boundary determination is performed in our case when the same audio type (music or speaker) characterizes successive shots. Partial scene identification is done according to the following concepts:

- commercials are identified by their background music and the many, short in duration, shots that they have.
- dialogue scenes can be identified by the high speaker alternation rate exhibited inside the scene.

5 Simulation Results

Experimentation has been carried on several real TV sequences having many commercials in-between, containing many shots, characterized by significant edit effects like zoom-ins/outs and dissolves, abrupt camera movement and significant motion inside single shots. We shall present here a representative case of a video sequence of approximately 12 min duration that has been digitized with a frame rate of 25fps at QCIF resolution. The audio track is a mixture of silence, speech, music and, in some cases, miscellaneous sounds. The audio signal has been sampled at 11kHz and each sample is a 16bit signed integer. In the sequel, firstly the performance of the various aural and visual analysis tools presented in Sects. 2 and 3 will be investigated. Then, scene change detection will be examined and partial scene characterization will be attempted.

In order to evaluate the performance of the audio segmentation techniques, the following performance measures have been defined:

- Detection ratio: the % ratio of the total duration of correctly detected instances versus that of the actual ones,
- False alarm ratio: the % ratio of the total duration of falsely detected instances versus that of the actual ones,
- False rejection ratio: the % ratio of the total duration of missed detections versus that of the actual ones,

focusing initially on the performance of the aural analysis tools. Thus, silence detection exhibits a remarkable performance of 100% detection ratio and 0% false rejection ratio, achieving to locate entire words, phrases or sentences. Rare occasions of unvoiced speech frames being classified as silence frames have only been observed leading to a false alarm ratio of 3.57%. There was no case of silence being classified as any other kind of audio types searched for. Music detection exhibits 96.9% detection ratio, 3.1% false rejection ratio, because some music segments of short duration are being confused as speech. It has 7.11% false alarm ratio, because it confuses some speech segments as music ones. On the other hand, speech detection is characterized by 86.2% detection ratio, 13.8% false rejection ratio and 2.4% false alarm ratio by mistaking music segments as speech. Finally, speaker change instant detection attains a suboptimal performance mainly attributed to the fact that covariance matrices and their inverse ones are insufficiently evaluated given a limited number of feature vectors extracted from 2sec duration segments. However, the use of bigger audio segments would imply that the same speaker is speaking during a longer duration, which would be long in many cases. Speaker change instants are evaluated with a detection accuracy of 62.8%. We have 30.23% false detections, while missed detections are of a percentage of 34.89%. Enhancement of this method may be achieved by simultaneously considering other similarity measures as well, as shown in [11]. Despite, however, of the suboptimal performance of speaker change instants detection, their use during audio-visual interaction for scene boundary detection leads to a satisfactory outcome, in combination with the other segmentation results.

In order to evaluate the performance of the visual segmentation methods, that is, the shot boundary detection methods presented in Sect. 3, the following performance criteria are used [2]:

$$\text{Recall} = \frac{\text{relevant correctly retrieved shots}}{\text{all relevant shots}} = \frac{N_c}{N_c + N_m} \tag{21}$$

$$\text{Precision} = \frac{\text{relevant correctly retrieved shots}}{\text{all retrieved shots}} = \frac{N_c}{N_c + N_f} \tag{22}$$

where N_c denotes the number of correctly detected shots, N_m is the number of missed ones and N_f is the number of falsely detected ones. For comparison purposes and to illustrate the strength of combining different methods and fusing results, the above criteria are also measured for the single shot detection methods presented in Sect. 3. Results for the single cases as well as the combined one are presented in Table 1. Adaptive thresholding that leads to the decision about shot boundaries is performed using two different lengths for the local windows: $W = 3$ and $W = 5$. It can be observed that the combined method attains the best results for $W = 5$. No false detections are made and the missed ones are rather few even under dissolve camera effects. The color vector histogram difference method is inferior in performance compared to the color frame difference method because histograms do not account for spatial color localization. However, the histogram approach is better under illumination changes. To illustrate the discriminative power of all temporal difference metrics considered

Table 1. Recall and Precision values achieved by the Shot Boundary Detection methods.

Method	$W = 3$		$W = 5$	
	Recall	Precision	Recall	Precision
Color Frame Difference	0.7047	0.5866	0.8456	0.7975
Color Vector Histogram Difference	0.3356	0.2155	0.5705	0.4271
Combined Method	0.9329	0.9858	0.9396	1.0

in the shot cut detection methods, i.e., the color frame difference metric FD_t, the color vector histogram difference metric HD_t and the combined difference metric OD_t, Fig. 1 is given, where parts of these temporal difference metrics are shown. One can easily observe how more easily distinguishable are peaks in the third plot, even in parts of the video sequence where a lot of action and movement is dominant, and how less varying are the rest values.

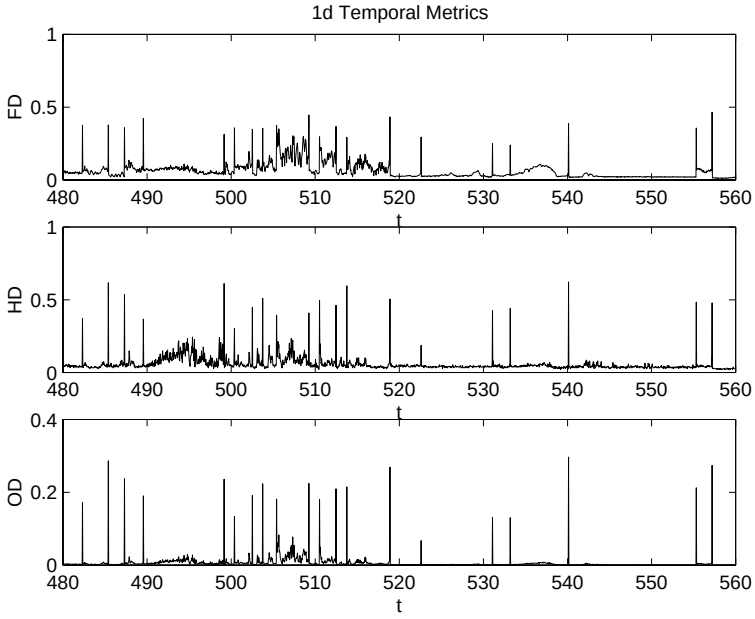

Fig. 1. Evaluated $1d$ temporal difference metrics: FD_t (top plot), HD_t middle plot, OD_t bottom plot, for a certain temporal part of the input sequence.

Finally, the performance of the method according to scene boundary determination is investigated. The sequence under study contains 18 different scenes being either dialogue ones, or action ones, or commercials or the serial logo

displays. During boundary detection, those shots that exhibit the same speaker speaking or the same music part are combined together into a scene. The boundaries of the scenes are further extended according to shot boundaries. For example, if the same speaker is found to be speaking during frames 100 and 200, while shot boundaries have been detected to exist to frames 85 and 223, then scene boundaries are further extended to those, based on the enhanced performance of our shot boundary detection. Cases have been observed that extent scene boundaries to even a different speaker or music segment. Thus, dialogues may be identified if the speaker changing points in a scene are rather high. Results show that 13 out of 18 scenes are correctly detected, 12 are false detections (an actual scene is recognized as more than one due to the non-overlapping of speaker boundaries, music boundaries and shot boundaries), while 5 scene boundaries are missed. The performance is good considering that simple rules are imposed for scene boundary detection. Further investigation for scene characterization as well as incorporation of other analysis tools to define more semantic primitives and enhancement of methods attaining a suboptimal performance will be undertaken.

6 Conclusions

Content analysis and indexing systems offer a flexible and efficient tool for further video retrieval and browsing, especially now that distributed digital multimedia libraries have become essential. When such tasks combine semantic information from different data sources (auditory, visual, textual) through multimodal interaction concepts, enhanced scene cut detection and identification is possible. In this paper, a scene boundary detection method has been presented that attains promising performance. Both aural and visual sources are analyzed and segmented. The audio types used are speech, silence and music. Video segmentation into shots is performed by a remarkably efficient method that combines metrics used by the two distinct approaches. Interaction of the single source segmentation results leads to the determination of scene boundaries and the partial scene characterization.

References

[1] P. Correia and F. Pereira, "The role of analysis in content-based video coding and indexing", *Signal Processing, Elsevier*, vol. 66, no. 2, pp. 125–142, 1998.

[2] A. Del Bimbo, *Visual Information Retrieval*, Morgan Kaufmann Publishers, Inc., San Francisco, California, 1999.

[3] M.R. Naphade, R. Mehrotra, A.M. Ferman, J. Warnick, T.S. Huang, and A.M. Tekalp, "A high-performance shot boundary detection algorithm using multiple cues", in *Proc. of 1998 IEEE Int. Conf. on Image Processing*, Chicago, Illinois, USA, 4-7 Oct. 1998, vol. 1, pp. 884–887.

[4] N. Dimitrova, T. McGee, H. Elenbaas, and J. Martino, "Video content management in consumer devices", *IEEE Trans. on Knowledge and Data Engineering*, vol. 10, no. 6, pp. 988–995, 1998.

[5] R. Lienhart, S. Pfeiffer, and W. Effelsberg, "Scene determination based on video and audio features", in *Proc. of 1999 IEEE Int. Conf. on Multimedia Computing and Systems*, Florence, Italy, 1999, pp. 685–690.

[6] C. Saraceno and R. Leonardi, "Identification of story units in audio-visual sequences by joint audio and video processing", in *Proc. of 1998 IEEE Int. Conf. on Image Processing*, Chicago, Illinois, USA, 4-7 Oct. 1998, vol. 1, pp. 363–367.

[7] C. Saraceno, "Video content extraction and representation using a joint audio and video processing", in *Proc. of 1999 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 15-19 Mar. 1999, vol. 6, pp. 3033–3036.

[8] S. Tsekeridou and I. Pitas, "Audio-visual content analysis for content-based video indexing", in *Proc. of 1999 IEEE Int. Conf. on Multimedia Computing and Systems*, Florence, Italy, 1999, vol. I, pp. 667–672.

[9] L. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, N.J.: Prentice Hall, 1978.

[10] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[11] P. Delacourt and C. Wellekens, "Audio data indexing: Use of second-order statistics for speaker-based segmentation", in *Proc. of 1999 IEEE Int. Conf. on Multimedia Computing and Systems*, Florence, Italy, 1999, vol. II, pp. 959–963.