Proc. 11th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2010), Desenzano del Garda, Italy, April 2010.

ON THE USE OF AUDIO EVENTS FOR IMPROVING VIDEO SCENE SEGMENTATION

Panagiotis Sidiropoulos[1], *Vasileios Mezaris*[1], *Ioannis Kompatsiaris*[1], *Hugo Meinedo*[2], *Miguel Bugalho*[2,3], *Isabel Trancoso*[2,3]

[1]Informatics and Telematics Institute / Centre for Research and Technology Hellas, 6th Km Charilaou-Thermi Road, Thermi 57001, Greece
[2]INESC-ID Lisboa, Portugal
[3]IST/UTL, Rua Alves Redol 9, 1000-029 Lisboa, Portugal

ABSTRACT

This work deals with the problem of automatic temporal segmentation of a video into elementary semantic units known as scenes. Its novelty lies in the use of high-level audio information in the form of audio events for the improvement of scene segmentation performance. More specifically, the proposed technique is built upon a recently proposed audio-visual scene segmentation approach that involves the construction of multiple scene transition graphs (STGs) that separately exploit information coming from different modalities. In the extension of the latter approach presented in this work, audio event detection results are introduced to the definition of an audio-based scene transition graph, while a visual-based scene transition graph is also defined independently. The results of these two types of STGs are subsequently combined. The application of the proposed technique to broadcast videos demonstrates the usefulness of audio events for scene segmentation.

1. INTRODUCTION

Video temporal decomposition into elementary semantic units is an essential pre-processing task for a wide range of video manipulation applications such as video indexing, non-linear browsing, classification etc. One of the most commonly used elementary semantic units of video is the scene, which is often defined as a Logical Story Unit (LSU) [1], i.e. a series of temporally contiguous shots characterized by overlapping links that connect shots with similar content.

Early approaches to scene segmentation focused on exploiting visual-only similarity among shots [1, 2] to group them into scenes. In [2], the Scene Transition Graph (STG) was originally presented. The Scene Transition Graph method exploits the visual similarity between key-frames of video shots to construct a connected graph, whose cut-edges constitute the set of scene boundaries.

In the last years, several scene segmentation methods that exploit both the visual and auditory channel have been developed, including [3, 4, 5, 6]. In [3] a fuzzy k-means algorithm is used for segmenting the auditory channel of a video into audio segments, each belonging to one of 5 classes (silence, speech, music etc.). Following the assumption that a scene change is associated with simultaneous change of visual and audio characteristics, scene breaks are identified when a visual shot boundary exists within an empirically-set time interval before or after an audio segment boundary. In [4] visual information usage is limited to the stage of video shot segmentation. Subsequently, several low-level audio descriptors (i.e. volume, sub-band energy, spectral and cepstral flux) are extracted for each shot. Finally, neighboring shots whose Euclidean distance in the low-level audio descriptor space exceeds a dynamic threshold are assigned to different scenes. In [5] audio and visual features are extracted for every visual shot and serve as input to a Support Vector Machines (SVM) classifier, which decides on the class membership (scene-change / non-scene-change) of every shot boundary. However, this requires the availability of sufficient training data. Although audio information has been shown in these and other previous works to be beneficial for the task of scene segmentation, higher-level audio features such as speaker clustering or audio event detection results are not frequently exploited. In a recent work [6], the use of audio scene changes and automatic speech recognition (ASR) transcripts together with visual features is proposed; audio scene changes are detected using a multi-scale Kullback-Leibler distance and low-level audio features, while latent semantic analysis (LSA) is used for calculating the similarity between temporal fragments of ASR transcripts. In [7], the combined use of visual features and high-level audio cues (namely, speaker clustering and audio background characterization results) for constructing scene transition graphs was proposed.

In this work, this definition of the scene as a Logical Story Unit is adopted and the method of [7] is extended in order to exploit richer high-level audio information. To this end, a large number of audio event detectors is employed, and their detection scores are used for representing each temporal segment of the audio-visual medium in an audio event space. This representation together with an appropriate distance measure is used, in combination with previously exploited high-level audio (e.g. speaker clustering results) and low-level visual cues, for constructing a combination of different scene transition graphs (Multi-Evidence STG - MESTG) that identifies the scene boundaries. The rest of the paper is organized as follows: an overview of the proposed approach is presented in section 2. Audio event definition and the use of audio events in representing video temporal segments are discussed in Sections 3 and 4, while section 5 presents the proposed MESTG approach. Experimental results are presented in Section 6 and conclusions are drawn in Section 7.

2. OVERVIEW OF THE PROPOSED APPROACH

Scene segmentation is typically performed by clustering contiguous video shots; the proposed MESTG approach is no exception to this rule. Thus, scene segmentation starts with application of the method of [8] for generating a decomposition S of the video to visual shots,

$$S = \{s_i\}_{i=1}^{I} \quad (1)$$

This work was supported by the European Commission under contracts FP6-045547 VIDI-Video and FP7-248984 GLOCAL.

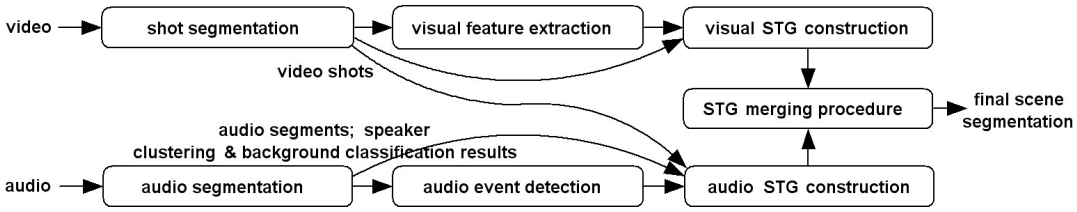

Fig. 1. Overview of the proposed scene segmentation scheme.

Subsequently, as illustrated in Fig. 1, visual feature extraction is performed. Audio segmentation, which includes, among others, speaker clustering and background classification stages [9] [10], is also performed in parallel. This audio segmentation process results in the definition of a partitioning of the audio stream,

$$\mathcal{A} = \{\alpha_x\}_{x=1}^{X}, \quad \alpha_x = [t_x^1, t_x^2] \tag{2}$$

where t_x^1 and t_x^2 are the start and end times of audio segment α_x, $\sigma(\alpha_x)$ denotes the speaker identity of it, if any, and $\beta(\alpha_x)$ denotes its background class that was identified during audio segmentation. Audio event detection, as discussed in the following section, is also performed. Using the resulting features, i.e.

- HSV histograms of shot key-frames,
- Speaker clustering results,
- Audio background classification into one of three categories (noise, silence, music),
- Detection results (confidence values) for 75 audio events,

the proposed MESTG method proceeds with the definition of two types of scene transition graphs (audio STG, visual STG) and of a procedure for subsequently merging their results.

3. AUDIO EVENTS

An audio event is defined, for the purpose of scene segmentation, as a semantically elementary piece of information that can be found in the audio stream of a video. Telephone ringing, dog barking, music, child voice, traffic noise, explosions are only a few of a wide range of possible audio events. As can be deduced from the audio event definition, more than one audio events may coexist in one temporal segment and may even temporally overlap with each other. For example, in a shot where a person stands by a street and talks, several speech- and traffic-related audio events are expected to coexist.

It is intuitionally expected that taking into account audio event detection results may contribute to improved video scene segmentation. This is based on the reasonable assumption that the presence of the same audio event in more than one adjacent or neighboring audio segments may be a good indication of their common scene membership. On the contrary, the presence of completely different audio events in adjacent temporal segments may be a good indication of their different scene membership, which reveals the presence of a scene boundary.

The first step in testing the validity of the above assumptions is the definition of a number of meaningful audio events and of appropriate methods for their detection. In this work, 75 audio events are defined and used; their full list is presented in Table 1. Different methodologies are used for their detection, depending on the event in question. The outcome of event detection is the estimation of a score ("confidence value") in the range $[0,\ 1]$ for each possible temporal segment - audio event pair, expressing our confidence in the specific event being present in the given temporal segment. More specifically,

- Classification using Support Vector Machines as described in [11] is used for the detection of 61 audio events (Dog Barking, Siren, Crowd Applause, Explosion, etc.).
- Classification using Multi-layer Perceptrons or Gaussian Mixture Models as described in [12] is used for audio diarization that leads to the detection of 14 additional audio events (Male Voice, Voice With Background Noise, Music, etc.).

As a result of the event detection process, a vector EV,

$$EV = [ev(1), ev(2), ..., ev(J)], \quad J = 75, \tag{3}$$

of confidence values is extracted and stored for each audio segment.

4. AUDIO EVENT-BASED SEGMENT REPRESENTATION AND SIMILARITY EVALUATION

For enabling the effective representation of temporal segments in the audio event space, and the evaluation of segment dissimilarity on the basis of audio events, two tasks are necessary: the normalization of the extracted audio event vectors, and the definition of an appropriate event vector distance measure.

Audio event vector normalization is motivated by the diversity of the distribution of confidence values among different event detectors for a given video. This is in part due to the differences in the actual frequency of appearance of different events within the video. For example, in a video with a female narrator speaking throughout the entire video and a thunder-like sound being heard in just a couple of shots, it is expected that the "female voice" audio event will receive very high confidence values in many shots, while the "thunder" audio event is likely to receive high or moderate confidence values in just the shots where the thunder-like sound is heard and even lower values in all others. However, the high or moderate confidence values that the latter audio event receives should be considered as a strong indication in favor of those shots' common scene membership. In order for them to receive the due attention during scene segmentation, the normalization of confidence values depending on their distribution for each audio event is proposed, and a very simple (most likely non-optimal) normalization approach is adopted in this work. Specifically, if $ev(j)$ is the initial confidence value of the j-th audio event in a temporal segment, and $maxev_j$ is the maximum value of the j-th audio event in all the temporal segments of the video, then the normalized confidence value $\tilde{ev}(j)$ is:

$$\tilde{ev}(j) = \frac{ev(j)}{maxev_j} \tag{4}$$

Table 1. List of audio events

Airplane Engine Jet	Car	Animal Hiss
Baby Whining or Crying	Bear	Bell Electric
Bell Mechanic	Big Cat	Crowd Applause
Bite Chew Eat	Bus	Buzzer
Airplane Engine Propeller	Cat Meowing	Donkey
Child Voice	Cow	Child Laughing
Clean Background	Birds	Wind
Digital Beep	Dog Barking	Dolphin
Chicken Clucking	Female Voice	Drink
Elephant or Trumpet	Electricity	Explosion
Door Open or Close	Fire	Fireworks
Music Background	Glass	Gun Shot Heavy
Gun Shot Light	Hammering	Helicopter
Horn Vehicle	Pig	Insect Buzz
Moose or Elk or Deer	Saw Manual	Male Voice
Wolf or Coyote or Dog Howling	Insect Chirp	Morse Code
Telephone Ringing Digital	Frog	Music
Non Vocal Music	Speech	Vocal Music
Noise Background	Paper	People Talking
Voice With Background Noise	Rattlesnake	Saw Electric
Telephone Ringing Bell	Sheep	Siren
Telephone Band	Whistle	Motorcycle
Voice With Background Music	Traffic	Train
Walk or Run or Climb Stairs (Soft)	Thunder	Horse Walking
Walk or Run or Climb Stairs (Hard)	Typing	Water

Following event vector normalization, the definition of a shot dissimilarity measure is based on the assumption that not only the difference of audio event confidence values between two segments, but also the absolute confidence values themselves, are important. Indeed, if for a given audio event two segments present similarly low confidence values, the only deduction that can be made is that this audio event is most probably not present in both segments; no conclusion can be drawn on the semantic similarity of these two segments. On the contrary, if two segments present similarly high confidence values, then it can be inferred that the same audio event is present in both segments, and this concurrence reveals a significant semantic similarity. The commonly used Minkowski distance would not satisfy the above requirements since it depends only on the difference of the confidence values. Instead of it, a variation of the Chi-test distance is employed in this work. If $\tilde{EV}_1$, $\tilde{EV}_2$ are two normalized audio event vectors, then their distance D is defined as:

$$D(\tilde{EV}_1, \tilde{EV}_2) = \sqrt{\sum_{j=1}^{J} \frac{(\tilde{ev}_1(j) - \tilde{ev}_2(j))^2}{\tilde{ev}_1(j) + \tilde{ev}_2(j)}} \quad (5)$$

It can be seen that this dissimilarity measure does not depend only on the difference of the audio event vectors, satisfying the previously discussed dissimilarity measure requirements.

5. MULTI-EVIDENCE SCENE TRANSITION GRAPH METHOD

5.1. Audio STG definition

The definition of the ASTG is based on the following assumptions:

- Scene boundaries are a subset of the visual shot boundaries of the video (i.e. a visual shot cannot belong to more than one scenes).
- Each audio segment cannot belong to more than one scenes. The same holds for a set of temporally consecutive audio segments that share the same $\sigma(.)$, $\beta(.)$ values and exhibit similar audio events. Two audio segments are said to exhibit similar audio events if the distance between their audio event vectors, as defined in Section 4, is lower than an empirical threshold.
- Audio event similarity and the distribution of speaker identities across two shots (or two larger temporally contiguous video segments) can serve as measures of audio similarity.

Based on these assumptions, an ASTG is constructed as follows:

- Step 1. The similarity of temporally adjacent audio segments α_x, α_{x+1} is examined, starting from α_1. Denoting $\tilde{EV}_x$, $\tilde{EV}_{x+1}$ the audio event vectors of α_x, α_{x+1} respectively, the two audio segments are merged if $\sigma(\alpha_x) = \sigma(\alpha_{x+1})$, $\beta(\alpha_x) = \beta(\alpha_{x+1})$, and $D(\tilde{EV}_x, \tilde{EV}_{x+1}) < T_{ev}$, where T_{ev} is an empirically defined threshold. For simplicity, the audio segments resulting from this merging step and used in the next step continue to be denoted α_x.
- Step 2. Merging of visual shots is performed: for every α_x, the visual shots that temporally overlap with it by at least T_a msec are merged to a video unit.
- Step 3. The video units formed in step 2 are clustered according to the dissimilarity of their speaker identity distributions $\Delta(.)$ and the distance of their audio event vectors $D(.)$. The two dissimilarity measures are linearly combined to produce a one-dimensional distance measure. Assignment of two video units to the same cluster requires both this distance measure and the temporal distance between them to be lower than certain thresholds.
- Step 4. A connected graph with nodes representing the clusters of video units is formed, and a directed edge is drawn from a node to another if there is a shot included the first node that immediately precedes any shot included in the second node [2], [7]. The collection of cut-edges, i.e. the edges, which if removed, result in two disconnected graphs, constitutes the set of estimated video scene boundaries.

It should be noted that the speaker identity distribution of a video unit is:

$$H_x = [h_1 \;\; h_2 \; ... \; h_G] \quad (6)$$

where G is the total number of speakers in the video, according to the speaker clustering results, and h_g, $g = 1, \ldots, G$, is defined as the time that speaker g is active in the video unit divided by the total duration of the same video unit. The $L1$ metric is used as similarity function $\Delta(H_x, H_y)$.

5.2. Visual STG definition

Similarly to ASTG, a scene transition graph based on visual information (VSTG) is defined. The VSTG comprises nodes, which contain a number of visually similar and temporally neighboring shots, and edges which represent the time evolution of the story. Visual similarity of shots is evaluated by calculating the Euclidean distance of HSV-histogram vectors of shot key-frames. More details on the visual scene transition graph can be found in [2].

5.3. Visual and Audio Scene Transition Graph merging

In [7] we introduced a probabilistic scene transition graph merging approach that combines the visual and audio STGs and simultaneously reduces the dependency of the proposed approach on STG

Table 2. Performance evaluation of MESTG

Method	VSTG [2]	[3]	AVSTG [7]	MESTG
Coverage (%)	79.18	77.93	83.86	**85.75**
Overflow (%)	17.81	13.88	11.05	**10.71**

construction parameters. Similarly to this approach, in this work multiple VSTGs are created, each using a different randomly selected set of parameter values. Then, the fraction p_i^v of VSTGs that identify the boundary between shots s_i and s_{i+1} as a scene boundary (i.e. the number of such VSTGs, divided by the total number of generated VSTGs) is calculated and used as a measure of our confidence on this being a scene boundary, based on visual information. The same procedure is followed for audio information using multiple ASTGs, resulting in confidence values p_i^a. Subsequently, these confidence values are linearly combined to result in an audio-visual confidence value p_i:

$$p_i = V \cdot p_i^v + U \cdot p_i^a \tag{7}$$

Finally, all shot boundaries for which p_i exceeds a threshold form the set of scene boundaries estimated by the proposed MESTG approach. In the above formula, U and V are global parameters that control the relative weight of the ASTGs and VSTGs in the audio-visual scene boundary estimation.

6. EXPERIMENTAL RESULTS

For experimentation, a test-set of 7 documentary films (229 minutes in total) from the collection of the Netherlands Institute for Sound & Vision[1] was used. Application of the shot segmentation algorithm of [8] to this test-set and manual grouping of the shots to scenes resulted in 237 ground truth scenes. For evaluating the results of the proposed and other scene segmentation techniques, the Coverage and Overflow measures, proposed in [13] for scene segmentation evaluation, were employed. Coverage measures to what extent frames belonging to the same scene are correctly grouped together, while Overflow evaluates the quantity of frames that, although not belonging to the same scene, are erroneously grouped together. The optimal values for Coverage and Overflow are 100% and 0% respectively.

Using the above test-set and measures, the proposed approach (MESTG) was compared with the audio-visual scene segmentation technique (AVSTG) of [7], the method of [3], and the visual scene transition graph (VSTG). For constructing the latter, the required parameter values were chosen by experimentation as in [2]. For the MESTG and AVSTG approaches, the probabilistic merging procedure discussed in Section 5.3 was followed, involving the creation of 1000 ASTGs and 1000 VSTGs with different parameters for estimating the required probability values. Weights V, U of Eq. 7 were tuned with the use of least squares estimation and one video manually segmented into scenes; the resulting values were 0.482 and 0.518 respectively. The results of experimentation are shown in Table 2, where it can be seen that the use of audio events in MESTG leads to an increase of Coverage by 1.89% and a decrease of Overflow by 0.34% compared to the AVSTG. The MESTG approach also significantly outperforms the methods of [2] and [3].

[1] http://instituut.beeldengeluid.nl/

7. CONCLUSIONS

In this work the use of high-level audio events for the improvement of scene segmentation performance was examined and a multimodal scene segmentation technique exploiting audio events and other audio-visual information was proposed. The proposed technique was shown to outperform previous approaches that did not exploit high-level audio events. Future extensions of this work include the use of additional audio events and experimentation with additional measures for evaluating similarity in the audio event space.

8. REFERENCES

[1] A. Hanjalic and R. L. Lagendijk, "Automated high-level movie segmentation for advanced video-retrieval systems," *IEEE Trans. On Circuits and Systems for Video Technology*, vol. 9, pp. 580–588, June 1999.

[2] M. Yeung and B.-L. Yeo, "Segmentation of video by clustering and graph analysis," *Computer Vision and Image Understanding*, vol. 71, pp. 94–109, July 1998.

[3] E.F.N. Nitanda, M. Haseyama, and H. Kitajima, "Audio signal segmentation and classification for scene-cut detection," in *IEEE Int. Symp. on Circuits and Systems*, 2005, vol. 4, pp. 4030–4033.

[4] A. Chianese, V. Moscato, A. Penta, and A. Picariello, "Scene detection using visual and audio attention," in *ACM Int. Conf. on Ambient Media and Systems*, 2008, vol. 4, pp. 4030–4033.

[5] K.W. Wilson and A. Divakaran, "Discriminative genre-independent audio-visual scene change detection," in *SPIE Conf. on Multimedia Content Access: Algorithms and Systems III*, 2009, vol. 7255.

[6] W. Jinqiao, D. Lingyu, L. Qingshan, L. Hanqing, and J.S. Jin, "A multimodal scheme for program segmentation and representation in broadcast video streams," *IEEE Transactions on Multimedia*, vol. 10, pp. 393–408, April 2008.

[7] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, and I. Trancoso, "Multi-modal scene segmentation using scene transition graphs," in *ACM Multimedia*, 2009, pp. 665–668.

[8] E. Tsamoura, V. Mezaris, and I. Kompatsiaris, "Gradual transition detection using color coherence and other criteria in a video shot meta-segmentation framework," in *IEEE ICIP Workshop on MIR*, 2008, pp. 45–48.

[9] R. Amaral, H. Meinedo, D. Caseiro, I. Trancoso, and J. Neto, "A prototype system for selective dissemination of broadcast news in European Portuguese," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, May 2007.

[10] H. Meinedo, *Audio pre-processing and speech recognition for Broadcast News, PhD Thesis*, IST, Technical University of Lisbon, March 2008.

[11] M. Bugalho, J. Portelo, I. Trancoso, T. Pellegrini, and A. Abad, "Detecting audio events for semantic video search," in *Interspeech 2009*. ISCA, 2009.

[12] I. Trancoso, T. Pellegrini, J. Portelo, H. Meinedo, M. Bugalho, A. Abad, and J. Neto, "Audio contributions to semantic video search," in *IEEE Int. Conf. on Multimedia and Expo*, 2009, pp. 630–633.

[13] J. Vendrig and M. Worring, "Systematic evaluation of logical story unit segmentation," *IEEE Transactions on Multimedia*, vol. 4, pp. 492–499, December 2002.