Multi-Modal Scene Segmentation Using Scene Transition Graphs

Panagiotis Sidiropoulos
Informatics and Telematics Institute / CERTH
Thermi-Thessaloniki, Greece
psid@iti.gr

Vasileios Mezaris
Informatics and Telematics Institute / CERTH
Thermi-Thessaloniki, Greece
bmezaris@iti.gr

Ioannis Kompatsiaris
Informatics and Telematics Institute / CERTH
Thermi-Thessaloniki, Greece
ikom@iti.gr

Hugo Meinedo
Technical University of Lisbon
Lisbon, Portugal
Hugo.Meinedo@l2f.inesc-id.pt

Isabel Trancoso
Technical University of Lisbon
Lisbon, Portugal
isabel.trancoso@inesc-id.pt

ABSTRACT

In this work the problem of automatic decomposition of video into elementary semantic units, known in the literature as scenes, is addressed. Two multi-modal automatic scene segmentation techniques are proposed, both building upon the Scene Transition Graph (STG). In the first of the proposed approaches, speaker diarization results are used for introducing a post-processing step to the STG construction algorithm, with the objective of discarding scene boundaries erroneously identified according to visual-only dissimilarity. In the second approach, speaker diarization and additional audio analysis results are employed and a separate audio-based STG is constructed, in parallel to the original STG based on visual information. The two STGs are subsequently combined. Preliminary results from the application of the proposed techniques to broadcast videos reveal their improved performance over previous approaches.

Categories and Subject Descriptors

I.2.10 [**Vision and Scene Understanding**]: Video analysis

General Terms

Algorithms

Keywords

Scene segmentation

1. INTRODUCTION

Video decomposition into elementary structural units is an extremely significant part of low-level video processing, since it facilitates video indexing, non-linear video browsing, video classification and summarization tasks. Scene segmentation, as part of video temporal decomposition, is crucial for the understanding and efficient organization of video content. In the relevant literature, the scene is described as an aggregation of consecutive semantically relevant shots. There are more than one ways to define the semantic relevance of neighboring shots. In this work, the scene definition as a Logical Story Unit (LSU) that was introduced in [5] has been followed. The LSU is defined as a series of temporally contiguous shots characterized by overlapping links that connect shots with similar content.

Early approaches on scene segmentation focused on exploiting visual-only similarity among shots [5], [11] to group them into scenes. In [11], the construction of a Scene Transition Graph (STG) by examining the visual similarity of shots was proposed. Such uni-modal techniques manage to cluster successfully video shots that exhibit visual similarity, but fail to do so when the scene membership of shots is signified by other means, e.g. by audio content similarity.

In the last years, several scene segmentation methods that exploit both the visual and auditory channel have been developed, including [3], [4], [6], [8]. In [8] 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 empirical time interval before or after an audio segment boundary. In [3] the video is segmented to visual shots, and low-level audio descriptors (i.e. volume, energy, zero crossing rate etc.) are extracted for each shot. Then, neighboring shots whose differences in low-level audio descriptors exceed some empirical thresholds are assigned to different scenes. In [4] audio and visual features are extracted for every visual shot and serve as input to a Support Vector Machine (SVM), which decides on the class membership (scene-change / non-scene-change) of every shot boundary. However, this requires the availability of sufficient training data. In [6] a multi-scale Kullback-Leibler (K-L) distance is exploited to locate audio scene changes, which are aligned with video scene changes

MM'09, October 19–24, 2009, Beijing, China.
Copyright 2009 ACM 978-1-60558-608-3/09/10 ...$5.00.

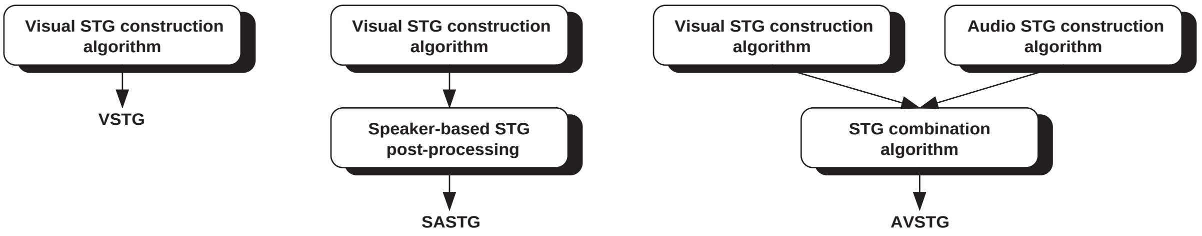
Visual STG construction algorithm
VSTG
Visual STG construction algorithm
Speaker-based STG post-processing
SASTG
Visual STG construction algorithm
Audio STG construction algorithm
STG combination algorithm
AVSTG

Figure 1: Outline of STG-based scene segmentation approaches: (a) Original STG approach of [11] (VSTG), (b) proposed SASTG approach (c) proposed AVSTG approach.

to represent candidate boundaries of scenes. Latent semantic analysis (LSA) is used for calculating the textual content similarity, and finally the multi-modal features are fused to detect scene boundaries. Although audio information has been shown in these and many other previous works to be beneficial for the task of scene segmentation, higher-level audio features such as speaker diarization results are not frequently exploited.

In this work, two multi-modal automatic scene segmentation techniques are proposed, both building upon the Scene Transition Graph (STG) and exploiting higher-level audio features such as speaker diarization and audio classification results (Fig. 1). The rest of the paper is organized as follows: Background information on the STG technique is given in Section 2. Sections 3 and 4 present the Speaker Assisted STG (SASTG) and the Audio-Visual STG (AVSTG) proposed in this work. Experimental results are presented in Section 5, and conclusions are drawn in Section 6.

2. SCENE TRANSITION GRAPH

The STG is a well-known idea that was firstly proposed in [11] for visual-only scene segmentation based on shot grouping. This STG (denoted VSTG in the sequel) comprises nodes, which contain a number of visually similar and temporally neighboring shots, and edges, which represent the time evolution of the story [11].

The VSTG construction starts with the generation of a segmentation S of the video to visual shots,

$$S = \{s_i\}_{i=1}^{I} \text{ where } s_i = \{f_k\}_{k=b_i}^{e_i} \qquad (1)$$

Shot s_i is defined as the set of frames between frame f_{b_i} and f_{e_i}. Two video shots are considered similar if they contain at least one pair of similar frames:

$$D(s_i, s_j) = \min_{m,n} \left(d(f_m, f_n) \right) \qquad (2)$$

where $b_i \leq m \leq e_i$ and $b_j \leq n \leq e_j$. Although the similarity of all frames of both shots needs to be evaluated according to this criterion, a set of selected key-frames is often used instead, for reducing computational complexity.

The visual similarity values $D(s_i, s_j)$ between each pair of shots s_i, s_j in the video, providing that s_i, s_j are less than an empirical time threshold T apart, are calculated and used for grouping shots that are similar into the same cluster. The clustering criterion requires each shot to be similar to every other shot in the same cluster. The termination of the clustering procedure is controlled by an empirical similarity threshold. From the clusters and the temporal ordering of the video shots, a scene transition graph is constructed, where nodes represent the shot clusters and a directed edge is drawn from a node to another if there is a shot represented by the first node that immediately precedes any shot represented by the second node. Finally, the set of ”cut-edges” of the scene transition graph is evaluated. A cut-edge is defined as an edge, which if removed, results in two disconnected graphs [2]. The collection of all cut edges of the VSTG constitutes the set of scene boundaries.

Various criteria have been developed for the measurement of the visual similarity of two frames. It was shown in [11] that the selection of visual descriptors and visual similarity metric does not play a critical role in the performance of a scene segmentation system. In the present work, HSV histograms of a few key-frames of each shot (selected by k-means clustering of all frames of the shot) were used as shot descriptors; the $L1$ metric was used as the similarity metric.

3. SPEAKER ASSISTED STG

Although visual similarity is strong evidence that two temporally neighboring shots belong to the same scene, visual dissimilarity is not by itself strong evidence for the opposite; in this case, the possibility of the shots being similar in terms of audio content should be examined. The latter however is not taken into account by the VSTG or any other visual-only scene segmentation approach, which explains why such techniques fail to correctly group shots to a scene when their scene membership is signified by audio content similarity.

In order to alleviate this drawback, the construction of a Speaker Assisted Scene Transition Graph (SASTG) is proposed. Instead of using low-level audio features such as energy etc. to define an audio similarity metric for the shots, we exploit the reasonable assumption that the presence of the same speaker in adjacent shots indicates that these shots belong to the same scene, while the opposite does not provide any information. This assumption is generally valid and can be used to discard erroneous scene boundaries that have been identified by the VSTG.

To realize the SASTG, speaker diarization results extracted using the method of [1], [7] are employed in this work. Speaker diarization is the process of identifying in the audio stream a set A of temporal segments that are homogeneous according to the speaker identity,

$$A = \{a_p\}_{p=1}^{P} \text{ where } a_p = [t_p^1, t_p^2] \qquad (3)$$

and further assigning a speaker identity $\sigma(.)$ to each speaker segment, by means of clustering the latter. t_p^1 and t_p^2 are the start and end times of segment a_p.

Then, the construction of the SASTG is done as follows:

- Step 1. A VSTG is constructed, according to section 2 and setting its construction parameters to values that favor over-segmentation. The reason for the latter choice is that, as opposed to over-segmentation errors, under-segmentation ones cannot be corrected in the subsequent steps of the SASTG algorithm.

- Step 2. For a scene boundary defined in the VSTG, construct the set A^- of speaker segments a_p within a time window of length T_s immediately before it, and the the set A^+ of speaker segments a_p within a time window of same length immediately following it.

- Step 3. Check if a pair of speaker segments a_p, a_q exists, for which $a_p \in A^-$, $a_q \in A^+$, $\sigma(a_p) = \sigma(a_q)$ and $t_q^1 - t_p^2 \leq T_s$. If at least one such pair of segments exists, merge the two nodes of the VSTG that represent the shot clusters to which a_p, a_q belong.

Steps 2 and 3 are repeated for all scene boundaries of the VSTG. The (remaining) cut edges of the resulting SASTG signify the detected scene boundaries.

4. AUDIO-VISUAL STG

In the algorithm of section 3, speaker diarization was used for post-processing the VSTG. In this section, a more generic audio-visual STG (AVSTG) approach is proposed. More specifically, audio information (not restricted to speaker diarization) is initially exploited for estimating the audio similarity of visual shots independently of their visual similarity; based on this audio similarity evaluation, an Audio STG (ASTG) is constructed. Although only audio classification and speaker diarization results are employed in this work for audio similarity evaluation, additional audio features could be introduced to the process to further improve its outcome. Then, the results of the ASTG and the VSTG of section 2 are appropriately combined.

4.1 Audio STG construction

For the construction of the ASTG, the audio stream of the video undergoes segmentation, classification according to background conditions, and speaker diarization [1], [7]. These processes result in the definition of a partitioning of the audio stream,

$$\mathcal{A} = \{\alpha_x\}_{x=1}^{X} \text{ where } \alpha_x = [t_x^1, t_x^2] \tag{4}$$

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. Possible background classes are noise, silence and music.

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 that one scenes).

- Each audio segment or set of temporally consecutive audio segments that share the same $\sigma(.)$ and $\beta(.)$ values cannot belong to more that one scenes.

- The distribution of speaker identities across two shots (or two larger temporally contiguous video segments) can serve as a measure of audio similarity.

Direct consequence of the first assumption is that visual shots serve, similarly to the case of the VSTG, as the basic units for constructing the ASTG. Based on these assumptions, the ASTG is constructed as follows:

- Step 1. The audio similarity of temporally adjacent audio segments α_x, α_{x+1} is examined, starting from α_1; if $\sigma(\alpha_x) = \sigma(\alpha_{x+1})$ and $\beta(\alpha_x) = \beta(\alpha_{x+1})$, then the two audio segments are merged. For simplicity, the audio segments resulting from this merging stage and used in the next one 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 similarity $\Delta(.)$ of their speaker identity distributions; the employed clustering criterion is similar to that used in VSTG construction.

- Step 4. The ASTG is constructed with nodes representing the unit clusters; directed edges are drawn from a node to another if there is a unit represented by the first node that immediately precedes any unit represented by the second node.

The speaker identity distribution of a video unit is:

$$H_x = [h_1 \ \ h_2 \ ... \ h_V] \tag{5}$$

where V is the total number of speakers in the video according to the speaker diarization results and h_v is defined as the fraction of time that speaker v is active in a video unit over the total duration of the same unit. The $L1$ metric is used as similarity function $\Delta(H_x, H_y)$.

4.2 Visual and Audio Scene Transition Graph merging

As expected, it was experimentally found that the scene boundaries estimated using either the VSTG or the ASTG alone depend significantly on the values of their construction parameters. In order to combine the visual and audio analysis results and simultaneously reduce the dependency of the proposed approach on parameters, we propose a technique for combining the VSTG and ASTG results that involves the creation of multiple ASTGs and VSTGs using different parameter values each time. In particular, following the creation of multiple VSTGs, the fraction p_i^v of VSTGs where the boundary between shots s_i and s_{i+1} was identified as a scene boundary over the total number of generated VSTGs is calculated and used as a measure of our confidence on this shot boundary also being a scene boundary, based on visual information. The same procedure is followed for audio information, resulting in confidence values p_i^a. Both p_i^v and p_i^a receive values in the range $[0, 1]$. Subsequently, these confidence values are linearly combined to result in an audio-visual confidence value p_i:

$$p_i = V \cdot \sum_{j=-y}^{y} w_j^v \cdot p_{i+j}^v + U \cdot \sum_{j=-y}^{y} w_j^a \cdot p_{i+j}^a. \tag{6}$$

In the above formula, U and V are global parameters that control the relative weight of the ASTG and VSTG in the audio-visual scene boundary estimation. Weights w_j^v and w_j^a control the contribution of the boundary between shots

Table 1: Performance evaluation of VSTG, SASTG, AVSTG and [8].

Method	VSTG	SASTG	AVSTG	[8]
Coverage (%)	80.98	82.44	88.78	83.90
Overflow (%)	12.09	8.47	9.56	12.14

s_i and s_{i+1} (for $j = 0$) and also of the $2 \cdot y$ neighboring shot boundaries to the calculation of the confidence value p_i. Finally, all shot boundaries for which p_i exceeds a threshold form the set of scene boundaries estimated by the proposed AVSTG approach.

It should be noted here that, due to the specifics of the AVSTG construction, the AVSTG can efficiently handle complex scenes. For example, a scene involving multiple persons participating in a conversation will be identified correctly on the basis of a number of audio and visual similarities among its constituent shots that the AVSTG takes into account. The same holds (though probably to a lesser degree, as indicated by the experimental results) for the SASTG method.

5. EXPERIMENTAL EVALUATION

In order to conduct a preliminary experimental evaluation of the proposed approaches, a test-set of 3 documentary films from the collection of the Netherlands Institute for Sound & Vision[1] was used. The videos had a total duration of approximately 115 minutes. The shot segmentation algorithm of [9] was applied to the test-set, followed by the proposed scene segmentation algorithms. Ground truth grouping of shots to scenes was generated manually by annotators, following the Logical Story Unit definition [5]. The resulting database included 123 ground truth scenes.

For evaluating the results, the coverage and overflow metrics that were proposed in [10] specifically for scene segmentation evaluation were employed. Coverage measures to what extent frames that belong 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.

In order to generate the VSTG and SASTG, the required parameter values were chosen by experimentation, while for constructing the AVSTG, 1000 ASTGs and VSTGs were constructed with different pre-set parameter values. Weights U and V were set equal to 0.28 and 0.72, while both w_j^v and w_j^a were set equal to 0 for all $j \neq 0$ and equal to 1 for $j = 0$.

The results of evaluation are shown in Table 1, where it can be seen that the SASTG performs better that the VSTG approach of the literature, while the AVSTG is shown to result in further overall improvement in performance (significantly better coverage and slightly inferior overflow) compared to the SASTG. Comparison with the method of [8], which detects scene boundaries by examining the presence of an audio boundary within an empirical time window before or after a visual shot boundary, also reveals the efficiency of the proposed approaches. In particular, the AVSTG outperforms the method of [8] both in terms of coverage and overflow, while the SASTG demonstrates significantly lower (thus, better) overflow compared to [8], at the expense of a slight concurrent decrease in coverage.

6. CONCLUSIONS

In this work segmentation of videos into scenes was examined and two novel techniques that combine visual and audio information to this end were presented. A preliminary experimental evaluation of them and comparison with literature approaches revealed promising results. Future work includes the optimization of the weights controlling the combination of VSTGs and ASTGs and the evaluation of their impact on scene segmentation performance, as well as thorough comparative evaluation of the proposed techniques on a more extensive data-set.

7. ACKNOWLEDGMENTS

This work was supported by the European Commission under contract FP6-045547 VIDI-Video.

8. REFERENCES

[1] 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*, 2007, May 2007.

[2] J. A. Bondy and U. Murty. *Graph Theory with Applications*. Macmillan Publishing Group, London, 1976.

[3] S.-C. Chen, M.-L. Shyu, W. Liao, and C. Zhang. Scene change detection by audio and video clues. In *Proc. IEEE ICME*, pages 365–368, August 2002.

[4] N. Goela, K. Wilson, F. Niu, and A. Divakaran. An svm framework for genre-independent scene change detection. In *Proc. IEEE ICME*, pages 532–535, July 2007.

[5] 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*, 9(4):580–588, June 1999.

[6] W. Jinqiao, D. Lingyu, L. Qingshan, L. Hanqing, and J. Jin. A multimodal scheme for program segmentation and representation in broadcast video streams. *IEEE Trans. on Multimedia*, 10(3):393–408, April 2008.

[7] H. Meinedo. *PhD Thesis*. IST, Technical University of Lisbon, Portugal, March 2008.

[8] N. Nitanda, M. Haseyama, and H. Kitajima. Audio signal segmentation and classification for scene-cut detection. In *Proc. IEEE Int. Symp. on Circuits and Systems*, volume 4, pages 4030–4033, May 2005.

[9] E. Tsamoura, V. Mezaris, and I. Kompatsiaris. Gradual transition detection using color coherence and other criteria in a video shot meta-segmentation framework. In *Proc. IEEE ICIP-MIR 2008*, pages 45–48, October 2008.

[10] J. Vendrig and M. Worring. Systematic evaluation of logical story unit segmentation. *IEEE Trans. on Multimedia*, 4(4):492–499, December 2002.

[11] M. Yeung and B.-L. Yeo. Segmentation of video by clustering and graph analysis. *Computer Vision and Image Understanding*, 71(1):94–109, July 1998.

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