Proc. 12th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2011), Delft, The Netherlands, April 2011.

VIDEO SCENE SEGMENTATION SYSTEM USING AUDIO-VISUAL FEATURES

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

[1]Informatics and Telematics Institute / CERTH, Thermi 57001, Greece
[2]Center for Vision, Speech and Signal Processing, University of Surrey, UK.
[3]Technical University of Lisbon, Lisbon, Portugal
{psid, bmezaris, ikom}@iti.gr, {hugo.meinedo, miguel.bugalho,isabel.trancoso}@inesc-id.pt

ABSTRACT

This work demonstrates a new approach to video temporal segmentation into scenes. The utilized technique is based on an audio-visual extension of the well-known method of the Scene Transition Graph (STG). This multi-modal extension exploits both low- and high-level audio-visual descriptors to construct distinct STGs. These STGs are employed into a probabilistic framework that is used for estimating a confidence value on each shot boundary also being a scene boundary. Finally, the thresholding of these confidence values generates the set of experimentally estimated scene boundaries. In this demo both the scene segmentation outcome and some intermediate features that lead to it are demonstrated.

1. INTRODUCTION

Video decomposition into semantic temporal segments, known as scenes, is an essential pre-processing task for a wide range of video manipulation applications, such as video indexing, non-linear browsing, classification etc. Traditionally scene segmentation is performed through uni-modal approaches, i.e. approaches that rely on visual information alone [1]. Although uni-modal techniques are usually sufficient for clustering together shots characterized by pronounced visual similarities, the same does not stand true when the semantic relation between shots is indicated only by other means, e.g. by audio. As a result, multi-modal techniques that combine visual and audio cues has been proposed. Recently the introduction of high-level visual and audio cues has been presented in the relevant literature, e.g. [2], [3].

2. SCENE SEGMENTATION SYSTEM

In this work we demonstrate a multi-modal scene segmentation system, which makes use of both low- and high-level audiovisual features. The simple web-based interface that enables the interaction with the proposed system is shown in Fig. 1. The user can select between a number of pre-processed videos to view the scene segmentation results. After the video is selected, the video player starts, while simultaneously the video's segmentation into scenes is demonstrated and clarified by depicting intermediate results of the scene segmentation process.

More specifically, while the video plays, one key-frame of the current shot as well of the 4 previous and the 4 subsequent shots are also shown. Using these key-frames, the pair of shots, found on either side of the current shot, which are most likely to belong to the same scene according to the low-level or high-level features that are employed in every STG type, are indicated. Furthermore, the estimated confidence value for the hypothesis that the shot boundary between the current shot and the previous one is also a scene boundary is depicted for every type of STGs. Finally, to the right of the video player, two barplots showing 5 visual concepts and 5 audio events, respectively, can be seen. These highlight the detection scores for the visual concepts and audio events which contributed the most to the detection of the shots that are most likely to belong to the same scene. A speaker ID barplot also demonstrates the estimated ID of the most frequent speaker in each of the nine depicted shots.

3. SCENE SEGMENTATION METHOD

The approach that is used as part of the proposed system is a probabilistic multi-modal scene segmentation approach that exploits both high- and low-level audio-visual descriptors. For this purpose four different sets of features are used to produce distinct scene transition graphs. The employed feature sets are i) typically used low-level visual features (HSV histograms), ii) model vectors constructed from the responses of a number of visual concept detectors, iii) typically used audio features (background conditions classification results, speaker ID histogram), and iv) model vectors constructed from the responses of a number of audio event detectors.

More specifically, the HSV histograms of a few key-frames of each shot, or very similar representations, have been extensively used in the relevant literature (e.g. [1]) and are also used in this work. Furthermore, the visual model vectors

are constructed from the responses of trained visual concept detectors and are used in this work as high-level visual features. The typical audio features are estimated by performing audio segmentation, classification according to background conditions, and speaker diarization [4]. Speaker diarization identifies speaker homogeneous segments in the audio stream and further assigns a speaker identity to each, after clustering them. Finally, audio events are the audio equivalent to visual concepts. Audio events are detected with the use of trained audio event detectors that rely on machine learning [5].

These four sets of features are used to produce different STGs. We have used a probabilistic technique that involves the independent creation of multiple STGs of each type, where a "type" means here an STG that uses a specific set of features, to reduce the dependence of the proposed approach on arbitrarily chosen parameters. Then, for every pair of adjacent shots the number of STGs that have identified the boundary between these shots as a scene boundary, divided by the total number of generated STGs of this type, is calculated and used as a measure of our confidence on this shot boundary also being a scene boundary, based on the features that this STG type employs. Subsequently, these confidence values are linearly combined to result in a cumulative confidence value, which is then thresholded to produce the final video scene boundaries.

More algorithmic details on preliminary versions of the scene segmentation method used in this work and also on specific elements on it can be found in [2], [3].

4. CONCLUSIONS

In this work, we presented a demo of a multi-modal scene segmentation method that makes use of both low and high level audiovisual features as well as a probabilistic STG combination framework. This demo enables the understanding of the contribution that each different type of features makes to the correct identification of scene boundaries, and also the showcasing of the good performance of the overall system.

5. ACKNOWLEDGMENTS

This work was supported by the European Commission under contract FP7-248984 GLOCAL.

6. REFERENCES

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

[2] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso, "On the use of audio events for improving video scene segmentation," in *11th Int. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2010, pp. 1–4.

[3] V. Mezaris, P. Sidiropoulos, A. Dimou, and I. Kompatsiaris, "On the use of visual soft semantics for video temporal decomposition to scenes," in *Fourth IEEE Int. Conf. on Semantic Computing (ICSC)*, 2010, pp. 141–148.

[4] 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, 2007.

[5] M. Bugalho, J. Portelo, I. Trancoso, T. Pellegrini, and A. Abad, "Detecting audio events for semantic video search," in *Interspeech*, 2009, pp. 1151–1154.

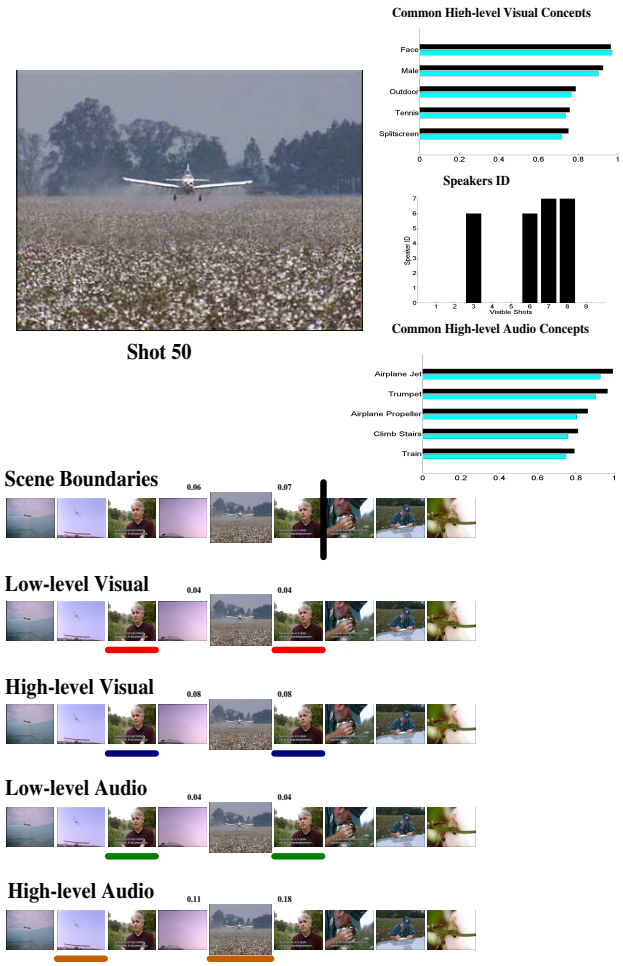

Fig. 1. User interface of the scene segmentation system. The video player is shown in the top left corner of the interface. Below it, intermediate and final results of the scene segmentation are shown, while to the right of the player some of the employed features are plotted.