Multimedia movie segmentation using low-level and semantic features

Bertrand Delezoide
CNRS/Ircam, Paris; CEA/LIST/LIC2M, Fontenay-Aux-Roses; France
bertrand.delezoide@ircam.fr

Abstract

The Movie structure extraction is essential to automatic and content-based organization, retrieval and browsing of movie. However, while many robust shot segmentation algorithms have been developed, it is still difficult to extract structures from a film. In this paper, we present a novel film segmentation scheme, in which low-level and semantic features from audio and image are integrated in movie structure extraction. We show by experiment on 8 films that semantic features bring important information for structure extraction,

1. Introduction

Movie structure parsing is the process of extracting construction units of movies. The problem is important for several reasons: (a) automatic segmentation is the first step towards greater semantic understanding of the film (b) breaking up the film into shorter sequences will help in creating film summaries (c) the determination of the structure will enable a non-linear navigation of the film.

There has been prior work on movie segmentation, usually two layers of construction units are considered: shots and scenes. After the seminal work of [1], work has been done for detecting shot boundaries in a video flow [2]. Other approaches focus on scene detection, like [3] using shot information and multiple cues like audio consistency between shots and the close caption of the speeches of the video. Bolle *et al.* [4] use types of shots and predefined rules to define scenes.

But it is possible to suggest more structural layers. In [5], the authors present a four layer video description and summarization system supported by a semantic and visual similarity strategy. This approach uses a set of predefined high-level ontological categories like action, time, space, etc. to annotate videos. Associated with low-level indexing, such as color and texture, the system achieves reasonable results of video segmentation.

Still, this technique has its drawbacks: (1) Due to the unsatisfactory results of video processing techniques in automatically acquiring video content, manual annotations are used. (2) Audio information is not considered, however, audio analysis can be a successful approach to improve structural segmentation results [3][8].

In this paper, we describe a movie analysis framework for an automatic description and segmentation of commercial films. We propose a four stage hierarchical structure of audio and visual signal of a film. A content description ontology is defined to describe the content of a film. This index is selected so that no more manual annotation is necessary, and that the semantic of the chosen concept improves the segmentation task. We study the efficiency of our model by experimenting the segmentation of 8 commercial films with an unsupervised HMM algorithm. We show that our structure model can be applied to all these films, and that the use of semantic information largely improves the segmentation task.

The paper is organized as follows. In the next section we present the four-layer model of video structure. Bottom layer content index is described in Sect.3. The segmentation algorithm of the four layers is presented in Sect.4. Our hierarchical film structure extraction scheme is tested in Sect.5. Concluding remarks are given in Sect.6.

2. Hierarchical movie structure

It is widely accepted that movie are hierarchically structured into film, scenes and shots. Such structure usually reflects the creation process of the movie. But it is possible to suggest more structural layers. Zhu [5] proposed a **four-layer hierarchical representation** for visual signal of video: shots, group of shots, scenes and group of scenes. We propose to extend it to audio signal and apply it to commercial films. This extension poses two questions: (1) Are these four layers applicable to audio signal? (2) If yes, are the boundaries of audio and visual segments, for a given layer, coincide? The answers are: yes and it depends on the layer considered. That's what we will discuss in

this section.

2.1. First layer: Shots and Clips

A movie **shot** is a video sequence that consists of continuous movie images for one camera action.

Generally, the audio signal is lot less fragmented than the visual signal. Audio is relatively independent from visual montage and more "realistic" to the acoustical environment. Thus, on the first level audio and visual segments boundaries do not correspond. Inspired by several audio scene segmentation algorithms [6][7], we define the "atomic" audio segment: **clip**. A clip is an audio segment homogeneous with basic semantic audio classes, namely: speech, music, speech+music, ambient-sounds and silence. Multiple speakers speech segments are also divided, so that only one speaker is present in a clip. This definition is relevant within the audio montage principle as it corresponds to the way sounds are picked-up and added to the audio track. Moreover, it authorizes the synchronization of every audio scene boundary with a clip boundary; as for image, a visual scene begins and ends with the beginning or the end of a shot.

2.2. Second layer: Group of shots and Group of clips

A **group of shots** is an ensemble of contiguous shots within a scene that are semantically or visually related [8]. In the example presented in Figure 1, a dialogue scene contains 2 groups of shots: the first group presents the characters and the location of the scene; the second group shows a repetitive visual structure of shots on each character face (field, reverse-field). This kind of structure is present in nearly every visual scene: dialogue, chase, flashbacks…

Similarly, we define the second layer audio segment: **group of clips**. A group of clips is an ensemble of clips within a scene that semantically or acoustically related.

Group of shots and group of clips boundaries usually do not correspond. In the example the audio scene contains one group of clips that correspond to a repetitive auditory structure: A speaking, B speaking, A speaking…

2.3. Third layer: Scenes

Scenes are a concept that is much older than motion pictures, ultimately originating in the theater. Traditionally, a scene is a continuous sequence that is

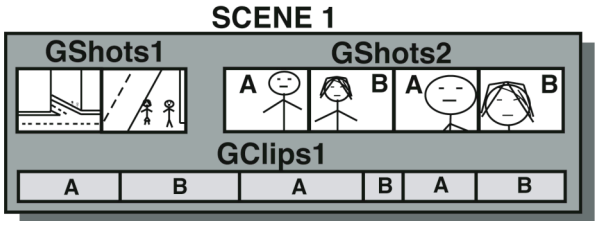

Figure 1. Audio and visual structure within a scene

temporally and spatially cohesive *in* the "real world". In films, audio and visual scene boundaries coincide so that the spectator may identify the unity of the sequence. Frequently, audio boundary may occur slightly before the visual boundary, this characteristic will be considered in the segmentation algorithm.

2.4. Fourth layer: Acts or group of scenes

Within a film it is possible to extract groups of scene or **acts** that have a global underlying semantic. In general films can be separated into three acts: the presentation of the characters and the subject, the confrontation of the characters with the subject, the ending.

It is interesting for indexing or retrieval to put the stress on these structures. But it is complicated to envision an automatic segmentation of such structures as they lay on complex semantics. It would require the ability to extract meaning from the film, a task well known to be extremely difficult for computers. Nevertheless, we believe that audio analysis may help in some cases. Indeed, the musical soundtrack relies on this semantic structure. As audio content vary slower than visual content, we believe that some audio features, as tempo or cepstre, may show a similar structure.

3. First layer content

The temporal organization of shots and clips, realized by montage, constitute the most specific foundation of the cinematographic language. To keep the "perceptive dynamism" during the film, filmmakers are due to respect some rules of continuity and homogeneity in the content of consecutive shots and clips. A rupture of this homogeneity induces a rupture in the psychological tension that indicates a change of structural sequence.

The choice of the features used to segment films is therefore dictated by the physical characteristics that respect this homogeneity (e.g. chromaticity, set, ambient sounds).

3.1. Low-level features

Image: Chromaticity and lighting conditions are described by a color-histogram [9].
Audio: Cepstral features are known to be a good signature of ambient sounds and music [8], as they express the energy distribution of sounds in the frequency space.

For both shots and clips, temporal length is extracted, as its brutal change may indicate the end of a sequence.

3.2. Semantic features

Image: Identification of objects [10] (e.g.: faces, cars) and textures [11] (e.g.: sky, grass, building) in shots, gives a description of a scene set. The hierarchical localization of a shot in indoor/outdoor, city/landscape, store/office/dwelling, furnishes information on the spatial coherence of a scene.
Audio: Classification of audio in speech, music, ambient-sounds, silence and speakers segmentation, present an interest for detecting breaks in the soundtrack.

These semantic features are represented in a semantic space by the relevance rate of their classification estimated by the classifier. These rates express the confidence in the automatic classification.

3.3. Feature fusion

Low-level and semantic features are then concatenated, so that the segmentation algorithm may analyze the set of features. We obtain one feature vector per shot and one per clip.

Features used here come from different extraction models. They describe multiple physical characteristics and media. Therefore, it is necessary to apply standardization: for each dimension of the vector the mean is set to 0 and the deviation is set to 1.

4. Segmentation algorithm

The global segmentation algorithm is presented in figure. It consists in (1) segmenting the first layer structure (shots and clips), (2) determining the corresponding sequences of features (one for audio, one for image), (3) segmenting hierarchically the three higher layers from top to bottom.

3.1. Shots segmentation

A standard algorithm developed in CEA [13] realizes shots segmentation by local maxima detection of a color and luminosity based function (recall=98%, precision=86%). For each shot a "mean" image is extracted as the local minima of the observation function. Image features are extracted from this image, which, we suppose, contains enough information about the shot it summarizes.

3.2. Clips segmentation

A supervised HMM segmentation model learned on annotated films of reference segments audio in speech/music/music+speech/ambient-sounds/silence (recall=96%, precision=94%). Segments containing speech are then segmented in one-speaker segments with an unsupervised HMM technique [12] (recall=87%, precision=76%). Those approaches are now widely spread and give good results of segmentation. Audio features are extracted from the integrality of each clip.

3.3. Higher layers segmentation

A common technique of scene segmentation consists in merging contiguous shots based on their similarity with local minima detection [8] or an unsupervised classification algorithm [5]. Several refinements of clustering algorithms have been proposed in order to take temporal contiguity constraints into account. But we found more appropriate to formulate this constraint using a unsupervised HMM approach developed by [14]. This general approach makes minimal assumptions regarding the content or structure of the source. Given an appropriate parameterization and distance measures, the approach is applicable to all media and data types.

First, for each media, given the sequence of features extracted from clips and shots, the acts structure is extracted by HMM algorithm. Boundaries from audio and image are then merged with a simple time-constrained nearest-neighbor algorithm [8]. Second, scenes structure is segmented within each detected act: (1) grouping shots and clips (2) merging boundaries from media. Third, the audio signal within each scene is segmented in groups of clips and the visual signal in groups of shots.

5. Experiment

In this section we shall discuss the experimental results of our models. The data used to test our models come from 8 commercial movies chosen to be representative of movie genres: action, romantic, social… The ground truth structure is annotated manually by an experimented spectator, based on experience and DVD marking.

We present results of segmentation for the 8 films and 4 layers of structure in the Table 1. Segmentation using only low-level features and both low-level and semantic features are compared. These results show that the exploitation of semantic information largely improves the segmentation of films.

	Group of Shots		Group of Clips		Scenes		Acts	
Recall	79	**87**	78	**82**	83	**90**	54	**56**
Precision	74	**75**	71	**76**	68	**79**	68	**71**

Table 1. Hierarchical segmentation results

This is obvious with scene segmentation for which the localization information gives the system the ability to perceive the spatial unity of shots within a scene. At group of shots and clips layer, the improvement is significant too, indeed the information given by faces detection and audio type classification (speech, music…) is important to segment this structure layer. At acts layer, the improvement is less significant; we believe that the semantic of the features used here is still to low to describe the information needed for this kind of segmentation.

6. Conclusion

In this paper, we have addressed the problem of film hierarchical segmentation for commercial movies. We have proposed a four layers film structure to describe the audio and visual signal of the film. We have selected a fully automatic content description index to describe the audio and visual content at different levels of semantic: low-level and high-level. We have developed a segmentation algorithm based on the visual and auditory content homogeneity within film sequences.

The film segmentation algorithm was tested on a set of 8 commercial movies. It works well, for all the structure layers, giving, notably, a scene detection result of 90% recall and 79% precision. We believe that these results are very encouraging when we keep the following consideration in mind: (a) the content index was entirely extracted automatically (b) the segmentation algorithm is rudimentary.

They are some clear improvement for this work: (a) The segmentation algorithm and nearest-neighbor fusion could integrate probability of detecting boundaries, that would limit some missed points; (b) some high-level semantic description could be added to the index, as action (e.g. car chasing, kiss), or characters and objects location and movements; (c) As the subtitles may convey useful concepts, it could be interesting to add this textual information to the system.

7. References

[1] A. Nagasaka and Y. Tanaka, "Automatic Scene-Change Detection Method for Video Works", *2nd Working Conference on Visual Database Systems*, 1991, pp. 119-133.

[2] A. F. Smeaton, P. Over and R. Taban, "The TREC-2001 Video Track Report*", The Tenth Text Retrieval Conference (TREC 2001)*, NIST, 2001.

[3] Y. Li, W. Ming and C.-C. Jay Kuo, "Semantic video content abstraction based on multiple cues", *IEEE International Conference on Multimedia and Expo (ICME)*, 2001.

[4] R. M. Bolle, B.-L. Yeo and M. M. Leung, "Video Query: Behond the Keywords*", IBM Research Report RC 20586 (91224)*, 1996.

[5] X.Zhu, J.Fan, A.Elmagarmid and X.Wu, "Hierarchical video content description and summarization using unified semantic and visual similarity", 2003.

[6] C. Saraceno and R. Leonardi, "Audio as support to scene change detection and characterization of video sequences", In *ICASSP*, 1997, pp 2597–2600.

[7] H. Jiang, T. Lin, and H.J. Zhang, "Video segmentation with the assistance of audio content analysis", In *ICME*, 2000, pp 1507–1510.

[8] H.Sundaram and S.F.Chang, "Determining computable scenes in films and their structures using audio-visual memory models", 2000.

[9] Y.-C. Cheng and S.-Y. Chen, "Image classification using color, texture and regions". *Image and Vision Computing*, 2003, 21 :759–776.

[10] P. Viola and M. Jones, "Robust real-time object detection", *International Journal of Computer Vision*, 2002.

[11] C. Millet, "Génération de sémantiques spatiales de différents niveaux". *Master's thesis, CEA/LIC2M*, 2004.

[12] L. Rabiner, "A tutorial on hidden markov model and selected applications in speech". *Proccedings of the IEEE*, 1989, 77(2):257–285.

[13] P. Josserand, "Detection de transitions à l'intérieur d'une séquence video en vue de son indexation", Master's thesis, Université du Littoral de Calais, 2000.

[14] J.Foote and M.Cooper. "Media Segmentation using Self-Similarity Decomposition," *Proc. SPIE Storage and Retrieval for Multimedia Databases*, 2003, Vol. 5021, pp. 167-75.