Enabling a rapid on-the-fly view of the content of a movie requires segmenting the movie and describing the segments in a user-compatible manner. The difficulty resides in extracting relevant semantic information from the audiovisual signal, both for the segmentation and the description. We introduce in this paper audio scenes and chapters in movies and present an algorithm for automatically segmenting a video based on the audio stream only. Audio scenes and chapters are defined as the equivalent of shots and scenes in the visual domain. A tree-like audio-based structure of a video is proposed. A chapter is then classified into different chapter categories. The automatic solution to audio scene and chapter segmentation and classification is evaluated on manually segmented and classified videos.

Keywords Video segmentation · Audio classification · Audio scenes · Piecewise Gaussian Model

1 Introduction

Structuring a video based on its semantic content is of increased relevance to allow intelligent video navigators, skimmers and search engines with the increase in the use of digital video archives. Video structuring is the problem of video segmentation into semantic units, also called story units, and the creation of a storyboard similar to the table of contents in a book. Automatically generating a storyboard of a video will probably enhance the experience of video retrieval both for consumer and professional applications. The objective of video structuring and description is to permit a user to have an on-the-fly idea of the content of a video or a scene.

In this paper, the term semantics is defined as a mid or high level interpretation of low-level signal features. Examples of mid-level interpretation include for instance, speech, music, cars, trees... Higher level understanding of these mid-level features, such as the classification of video content into “calm dialog”, “battle”, “car pursuit”, etc. is a real challenge for the content-based multimedia indexing community. This paper tries to use low- and mid-level audio features for content-based video structuring and description.

Starting from basic units such as shots, video structuring consists of grouping these units into scenes that are semantically homogenous and grouping scenes into chapters and iteratively continuing the grouping till covering the entire document. For example, several shots of one telephone conversation are considered as one scene. The underlying action is the telephone conversation. Several scenes telling one underlying story may be grouped together in one chapter. This kind of grouping is a semantic grouping based on a high level of understanding of the content. Although video segmentation into shots or the grouping of similar shots into scenes can be considered as a mature domain, semantic scene and chapter segmentation seems a harder problem. The main reason for the difficulty of scene, or story unit, determination is the high level of semantics engaged in the process, making even a manual segmentation of a video into scenes and chapters subjective and difficult to some extent.

In this paper, we consider structuring a video based on the audio stream only. We restrict our analysis to movies since they offer a major component in the entertainment market especially with future applications such as interactive TV or Video On Demand services.

The main motivation behind our analysis of the audio stream for video structuring is a preliminary experiment that we have conducted on the ability of human subjects to detect scene changes when listening to a movie with no information on the visual stream. We have observed that human subjects were able to structure a movie into semantic units even when they do not understand the language of the movie. Human subjects generally base their judgments on the combination of the acoustical environment and the changes in the mood, described by the mood of speech and music. This ability of understanding the semantic structure...
of a movie based on the audio stream only is probably due to the extensive use of the music and audio effects by the movie makers in order to convey semantic information to the spectators.

Moreover, the use of audio information for video segmentation offers the advantage of being computationally less demanding and having fewer dimensions than the visual information.

As a movie is classically structured into frames, shots, scenes, and chapters, we introduce in this paper an audio-only structure into audio scenes, and audio chapters at different levels of abstraction. An audio scene is defined as the time instants of a video containing a homogenous acoustical environment and a homogenous semantic environment, such as calm speech or loud music. The same criterion is used to group several consecutive audio scenes into audio chapters.

A table of contents of a movie is therefore created. The algorithm presented in this paper for the segmentation of the audio stream is based on a combination of spectral and semantic dissimilarities. We also present a grouping algorithm for merging consecutive audio scenes into scenes, or chapters, having a higher level of abstraction. Each chapter is moreover classified into different chapter categories.

2 Related work

Video scene segmentation has gained an important effort from the research community. In the majority of cases, the segmentation is based on an analysis of the visual stream of a video. Generally the video is first segmented into shots, a domain that can be considered as mature at present with high accuracy algorithms and standard methodologies for evaluation [1, 2]. Several shots are then grouped together on the basis of similarities in their color histograms, objects, motion, rhythm, and so on [3–6].

The audio stream has been considered in some cases as a complement to the visual stream. Simple algorithms for audio analysis and segmentation were generally used.

Another family of algorithms for video segmentation is the one considering an audio-only segmentation. Few papers in the literature fall in this category, to which the work presented here belongs.

The basic approach generally used for audio-based video segmentation can be considered as a blind segmentation. It consists of defining a similarity measure between neighboring time windows based on signal features such as spectral or cepstral ones. Peaks in the similarity measure correspond to potential scene boundaries.

Sundaram presents in [7] an audio scene segmentation technique based on low-level audio signal features, such as cepstral and energy features. The technique is based on a correlation measure between the envelope of the audio features in an attention time window (16 s) and the envelope of the features in a memory time window (17 s). This technique can actually be considered as a technique based on a measure of similarity between consecutive time windows, attention and memory windows. Cao et al. [8] demonstrate experimentally that a distance between spectral distributions in two consecutive time windows is more convenient than the correlation measure presented by Sundaram.

Minami et al. in [9] try to use a simple speech/music classification in videos in order to enable novel video browsers. The authors use information about the peaks in the spectrogram for speech and music classification. A comprehensive description of the significance of the audio stream in a video is presented in the paper; the authors base their analysis on video production rules. A video browser application is developed based on the speech/music classification.

Pfeiffer in [10] presents an algorithm for video segmentation based on audio analysis. The author uses spectral vectors as audio signal features. The Euclidian distance between vectors from present and a mean of an exponential prevision of the vectors from the past is used as a measure of similarity to detect scene boundaries.

In Ref. [11], shots are merged together using an audio-based similarity measure. The audio similarity measure is based on a Euclidian distance of the means and variances of low-level audio features such as Energy, Spectral centroid, and Spectral flux.

Atalan in [12] combines basic speech/music/silence detection with face detection in order to detect dialog scenes using a Hidden Markov Model. The speech/music/silence detection uses energy thresholding and frequency analysis.

A simple speech/music/silence detection algorithm is used in [13] in order to improve shot and scene detection in videos.

From this brief review of the state-of-the art in audio-assisted video scene segmentation, we can conclude that the proposed algorithms are more likely to be compared to shot segmentation in the visual domain. The segmentation process is generally based on low level acoustic features and sliding time windows. Whereas the process of generating scenes from shots requires a higher level of understanding of the video content, such as objects, rhythm and other semantic information. The audio segmentation presented in the literature, since it is based on low-level acoustic features and generally with little or no knowledge about the semantic content, can be considered as a blind segmentation. We argue that including semantic information to the segmentation process of the audio stream will permit a segmentation level more similar to that of scenes than to shots. Moreover, this semantic information may be used to describe the content of a scene.

Consider for instance a dialog scene between two persons, one from indoor and the other from outdoor. The acoustical environment of the two persons will be probably different. The mismatch in the acoustical environment will be translated by different characteristics if analysed in the low-level feature space, such as the cepstral or spectral feature space. Figure 1 illustrates an example from an English movie “Bridget Jones Diary”; it consists of a dialog scene