ISSN 2280 9481

CINERGIE
il cinema e le altre arti
3
n.s. marzo 2013

Art and Media Files

Soundscape Analysis as a Tool for Movie Segmentation

Automatic music processing is a research area that involves scientists of different disciplines, ranging from computer science to digital audio processing, and from web mining to information retrieval. Research in musicology benefits from automatic tools for music analysis that have been developed in this research area. With this contribution we would like to emphasize how tools developed to process and retrieve music can be applied also to the video domain. The idea is part of the SHOTS project, which goals are described by the contributions of Mario Brenta and Denis Brotto in this issue.

There are a number of application fields to which automatic music processing has been applied. Perhaps the most famous example is automatic song identification[1], a service widely popularized by S*hazam!* app for smartphones, which is also used for near duplicate detection in large music collections. The basic idea behind song identification is that the audio content can be described by a compact representation, called audio fingerprint (as it is done for identifying persons through their actual fingerprints). A similar task, although the underling techniques are usually different, is called automatic cover identification[2]. In this case the task is to identify different interpretations of the same composition, so approaches must be robust to variants in the music interpretation. Another related task is called query-by-humming[3], and in this case the user is asked to provide a personal rendition of the song he is looking for, by singing its main melody or other musically relevant parts.

The basic idea behind music recognition approaches is that it is possible to capture some music dimensions (from the unstructured audio of a performance to the symbolic information of the score) and use them to compute a measure of music similarity[4]. Similar techniques are also exploited in automatic film analysis, mostly for computing video fingerprinting to track unauthorized use of copyright material[5]. While the research on video similarity is focused on commercial applications related to news[6] and sport videos[7], where it is necessary to detect particular events (e.g. the goal of a soccer team) or particular images (e.g. the commercial logo of a brand). Video fingerprint is another popular technique for tracking unauthorized usage of videos protected by copyright.

Another automatic music processing task, which is relevant to the goals of SHOTS project, regards the automatic annotation of audio content, usually called music tagging[8]. Despite the existence of content-based approaches, music access and retrieval is still largely based on metadata descriptors where users have still an active role[9]. Music searches can be based on song title or artist name, but can also be based on more generic descriptors, such as mood, instrumentation, genre, and so on[10]. Although there has been extensive research on the subject, at the state of the art it is still very difficult to extract reliable information from the audio signal, so online services try to collect

this information directly from users tags, which are in the form of keywords or short phrases. Manual tagging is a slow process, suffering from well-known drawbacks such as the *new item* problem, where newly released songs cannot virtually be retrieved because they lack of descriptors, and the classic inconsistency problems of any manual procedure carried out by non-expert users. To this end, automatic music tagging aims at providing methods to associate a set of textual descriptors to corresponding songs using content-features extracted from audio.

Tags are increasingly used to retrieve all the forms of multimedia content. Services for video hosting are extensively exploiting user-generated tags[11] for their internal search engines, and users are continuously invited to tag additional content or, at least, to express an interest towards the content through the "I like" feature or by assigning a number of stars. While music language is not describing real objects, and thus tags are usually related to individual perception of music items, in the case of video tags are usually describing the content of the scene or the plot of the story. Video tagging, especially in the case of essay films, should borrow some of the ideas and tools used in music processing in order to provide new ways to interact with the content.

The goal of this contribution is to show, through simple examples, how video access can be improved by exploiting techniques of automatic music processing to analyze the soundtrack. There is a number of approaches that already exploit audio content for video analysis. In particular, speech processing[12] has been used since decades in the effort to automatically transcribe natural speech and identify the speakers. Automatic categorization of audio sources is used to distinguish between speech, music, and background noise[13] in order to transcribe only parts where speech is actually present. Automatic identification of audio sources is used to detect predefined events in video[14], such as the cheering of the crowd, the engines of F1 cars, and so on. Even if at a minor extent, also music is exploited, in particular to segment a video stream in its components, knowing that changes are usually introduced by musical cues (called jingles or bumpers). The ideas behind the research work presented in this paper are that music can provide a lot of information about video content, and that standard techniques of music processing can be effectively used to improve video access.

This contribution focuses on two tasks, both carried out using only audio information:

- Video segmentation, which is based on the identification of recurring audio events in the form of either background noise or music; this identification can be based on the audio fingerprinting techniques explained above.

- Video labeling, which is based on the recognition of given songs in the soundtrack (either the originals or the covers) using a collection of tagged songs; video tags can be inherited from audio tags in order to provide an initial description of the video content.

Both approaches are described in the following, as part of the initial research work to be used within SHOTS project.

Video Segmentation

Broadcast TV contains a number of recurrent events, which are related to the segmentation of a continuous video flow in different elements. In most cases, these elements introduce, open or close items in the palimpsests; hence the identification of these events can be used for video segmentation. In the particular case of commercial TVs, recurring elements are related to ads and promos (the former are commercial advertisements for external brands and products, the latter are teasers for forthcoming programs of the broadcasters). For our goals, promos are of particular interest because, for a given movie or TV show that is promoted, there are a number of different promos. While video content may change depending on the emphasis on different aspects of the TV programs, promos normally share the same music content, which is used as a signature.

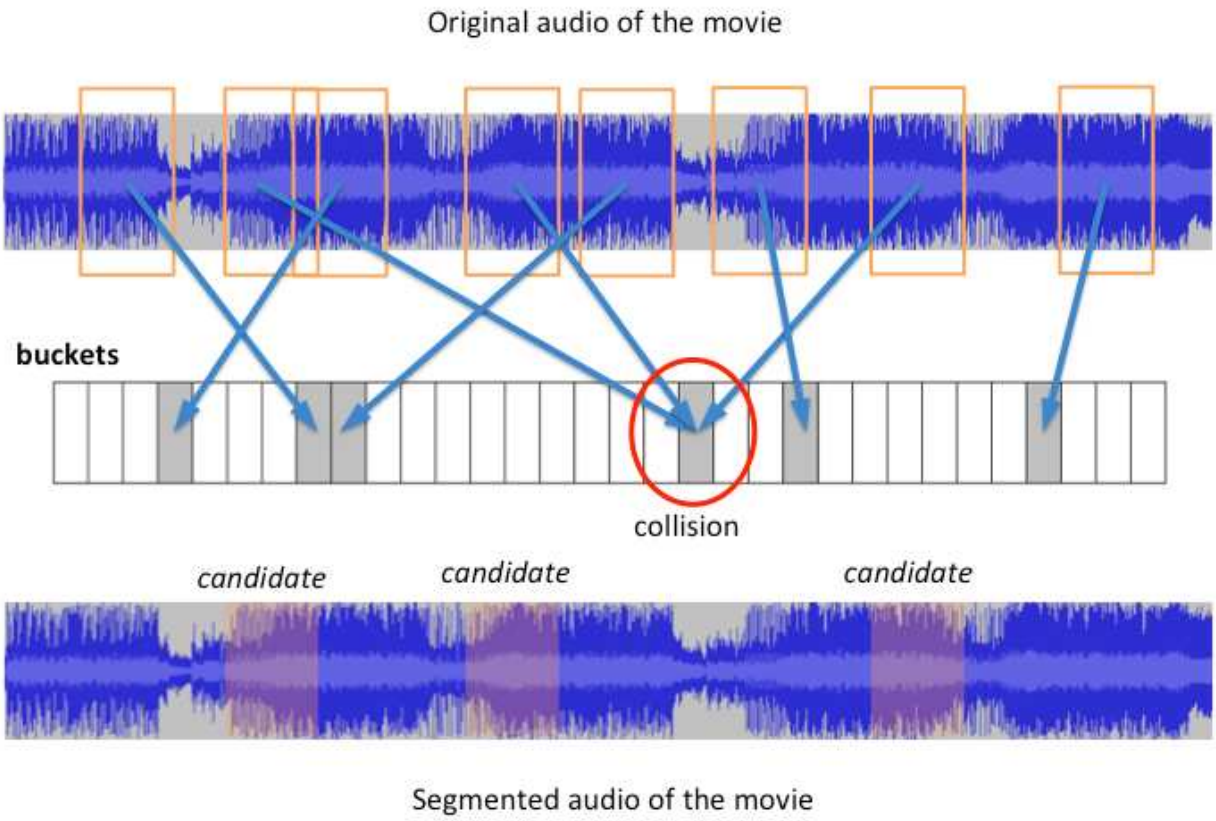
The workflow for audio-based video segmentation was the following:

- Extract audio fingerprints from the video signal, that is compute a set of coarse descriptors that are related to the spectral content of the audio signal. Audio fingerprints are robust to small changes in the audio content, due to additional noise, lossy compression and other artifacts. This means that two excerpts with similar audio content should have a similar set of fingerprints.

- Apply hashing to the fingerprints, that is transform them in a sequence of integer values providing that similar fingerprint are transformed in the same value (this is called technically a collision between hashes).

- Analyze the time axis to highlight the instants in which there is a maximum number of collisions. Since collisions are expected during recurring audio elements, they are likely to be related to the presence of ads and promos.

The figure gives a visual representation of the process. From unstructured video, the presence of recurring elements is highlighted though the analysis of collisions. A threshold on the average number of collisions is set in order to select potential boundaries in the video flow.



Description of the experiment. We used a collection of about six days (144 hours) of continuous video taken from a normal palimpsest of a commercial TV broadcaster. The video thus contains complete movies, news, and other programs interleaved by ads, promos, bumpers that appeared on average every 15 minutes. Video information has been discarded from the analysis, and used only to evaluate the effectiveness of the approach in correctly segmenting the palimpsest. Continuous video was collected through direct acquisition from the antenna signal, thus we expected problems in terms of variable audio quality and double conversion from digital (the

original song) to analogue (the aired signal) and back to digital (the recording).

The songs used for creating ads and promos were available, so we decided to run two parallel experiments: identifying collisions with the prototype audio descriptors, that is with the fingerprints of the songs used as soundtracks for ads and promos, and identifying collisions of the video content with itself, that is without any prior knowledge about the audio content.

A first measure of the effectiveness of the approach can be obtained from the number of correct segment labeling. When prototypes were used, the system achieved 82% of correct identifications, with problems due to possible audio post-processing to create the soundtrack. It is common practice to use different material of a well-known song and remix it in order to fit with the time of the video. When prototypes were not used, the percentage of correct identification reduced to 78%, which is still a satisfactory recognition rate. It is likely that the high redundancy of the palimpsests as regards ads and promos provides enough information also for blind identification. In both cases, the main problems were given by spoken only ads and promos.

Yet, the most important measure is the effectiveness of video segmentation. It has to be noted that, in the particular case of commercial TV, ads and promos are consistently presented in groups of at least four-five items. This makes segmentation a much simpler task because several ads and promos contributes in providing evidence of a boundary in the video stream. In fact, the approach achieved a very high rate, about 97%, of correct segmentations .

Discussion. The good results with this dataset suggests that audio-based video segmentation is a feasible task. Of course, the soundtrack of a movie provides much less redundancy in the audio content. Yet, it is likely that for a relevant number of movies the presence of recurring music themes and the consistent use of audio effects (which are usually taken from audio digital libraries) will make this task feasible. A similar approach can also be used to identify indoor and outdoor scenes, in case changes are emphasized by a change in the soundscape. Moreover, the identification of particular elements as part of the audio content can provide important cues for describe the movie content, as addressed in the next experiment.

Video Labeling

Music is often used to emphasize the message provided by the video. A part from the case of music purposely composed for a movie, or for a TV show, manual tagging and data mining can be exploited to associate textual descriptors to the original songs. As it has been explained earlier, automatic music tagging is one application of automatic music processing and has been extensively studied in the last years. The basic idea of the proposed approach is that music tags can represent an initial approximation of the description of a movie content.

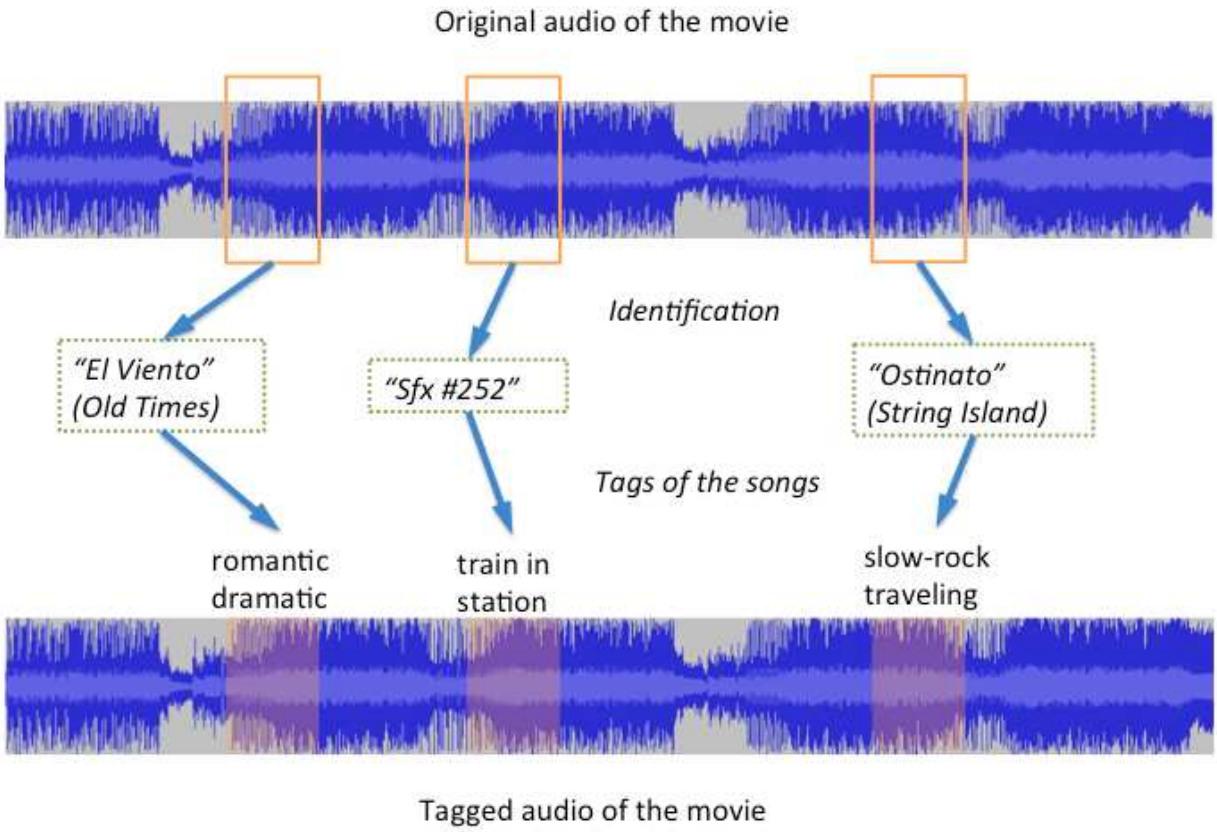
The workflow for audio-based video labeling was the following:

- Index a music library of music and audio effects, in order to carry out efficient identification of the video content. To this end, initiatives such as *FreeSound*, which makes available thousands of audio effects and is increasingly used in video productions, represent an optimal source of information.

- Apply cover identification techniques to the audio content. A typical approach is to carry out an approximate search using an hashing function (eventually the same function used for video segmentation can be exploited) and then refine the results using more sophisticated techniques such as statistical models or approximate string matching.

- Label video with the corresponding audio, retrieve the available tags for the identified songs (usually genre, mood and usage for video, and the name of the source for audio effects), and inherit the audio tags to video.

The figure gives a visual representation of the process. Given the soundtrack of a movie, or a TV program, automatic cover identification is exploited to associate known songs and audio effects to its elements. Music tags, titles, and names of the audio effects are then used as video descriptors. The example reported in the figure is taken from the TV serial “The lady of the lake” and the particular excerpt is related to a scene where the protagonist is leaving her home town. Although it is not possible to infer this situation from audio only (without using at least speech), the retrieved tags can provide a rough description of the video content.



Description of the experiment: We used four video productions aired by a commercial TV, for a total length of about five hours of video. The broadcaster is also the producer of the videos and owns the music library used to create the soundtrack. Thus it was possible to perform automatic cover identification with the ground truth directly available. As regards the music library, it contains about 9,000 songs of production music (i.e. music composed and performed to be used only as soundtrack). Each song has a title, chosen by the composer who is often also the performer, and belongs to a category (represented between brackets in the figure). Afterwards, a group of music experts tagged each song of the library using a controlled vocabulary of both genre and mood tags. The library also contains about 800 audio effects, which have only an Id and are tagged only describing the audio source.

Also in this case we carried out two parallel experiments. On the one hand, we measured the effectiveness of music identification techniques in order to analyze the correctness of the association between songs and video. On the other hand we carried out a qualitative evaluation of how music tags can be an effective descriptor of video content.

As regards the identification rate, there is a substantial difference between music, which achieved 95% of correct identifications, and audio effects, which had on average a much lower value of 71%. In particular, there were some effects that were never recognized, due to their non-stationary nature (bursts, slams) or to a short length. It can be noticed that, in other experiments of cover identification, the rate of correct song identification drops to 53% in case of live music. Yet, this is not normally the case in movie productions especially for TV movies and serials.

Evaluation of the effectiveness of music genre and mood labels as descriptors of the video content has been carried out informally matching the inherited tags with the plot of the story. Video were annotated by about twenty genre and mood labels. As expected, genre gives little information about the video content, since it is linked to features such as music style, historical period, and instrumentation. Song mood was found to be correlated to the mood expressed by the movie excerpt, although the statistical significance of this correlation has not been computed. Audio effects were good indicators for inferring the video content. A subsequent categorization of audio effects could be useful to infer the context of the video.

Discussion. Results with this dataset are encouraging, although the effectiveness of music tags as descriptors of video content has been carried out informally. The availability of music tags for large music collections, often in the form of social tags provided by the end users, allows us to extend the approach also to essay film and, most of all, to commercial productions. Clearly the semantic of music tags might not have the ability to describe in detail a movie, besides the fact that audio effects may not be identified at all. Yet, it is likely that the use of the soundtrack could improve video tagging

Final Remarks and Future Work

Audio content is a rich source of information for video processing techniques, which need to be studied in more detail. In this paper we show some directions in which audio content can be used to segment or to characterize a video. Video processing can be inspired by results in audio processing, or it can be used as an additional tool to refine the results obtained from the audio.

As explained by Mario Brenta and Denis Brotto in this issue, SHOTS is an ongoing project with very large objectives. The purpose of this contribution was to provide an initial demonstration of the possibility of video processing using audio. In the future we will explore other techniques, in particular related to segmentation and characterization of audio/video source. An important aspect will be the automatic analysis of the soundscape, in order to identify the location were a scene has been shot. Speech analysis deserves particular attention, because of its importance in films. To this end, much attention will be paid to characterize the kind of speech (monologues, dialogues, verbal fights, and so on) in order to provide alternative way of accessing the video material.

Nicola Orio

---

[1] Wei Li, Yaduo Liu, and Xiangyang Xue, “Robust audio identification for MP3 popular music”, *Proceedings of ACM SIGIR Conference*, 2010, pp. 627-634.

[2] Emanuele Di Buccio, Nicola Montecchio, and Nicola Orio, “FALCON: FAst Lucene-based Cover sOng identification”, *Proceedings of the International Conference on Multimedia*, 2010, pp. 1477-1480.

[3] Erdem Unal, Shrikanth S. Narayanan, and Elaine Chew, “A statistical approach to retrieval under user-dependent uncertainty in query-by-humming systems”, *Proceedings of the ACM SIGMM*

International Workshop on Multimedia Information Retrieval, 2004, pp. 113-118.

[4] Nicola Orio. "Music Retrieval: A Tutorial and Review", *Foundations and Trends in Information Retrieval*, n. 1(1), 2006, pp 1-90.

[5] Xavier Anguera, Pere Obrador, and Nuria Oliver, "Multimodal video copy detection applied to social media", *Proceedings of the SIGMM Workshop on Social media*, 2009, pp. 57-64.

[6] Mohammad Ghoniem, Dongning Luo, Jing Yang, and William Ribarsky, "NewsLab: Exploratory Broadcast News Video Analysis", *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, 2007, pp. 123-130.

[7] Ning Zhang et al., "A Generic Approach for Systematic Analysis of Sports Videos", *ACM Transactions on Intelligent Systems Technology*, n. 3(3), 2012, pp. 46:1- 46:29.

[8] Riccardo Miotto, and Gert R. G. Lanckriet, "A Generative Context Model for Semantic Music Annotation and Retrieval", *IEEE Transactions on Audio, Speech & Language Processing*, n. 20 (4), 2012, pp. 1096-1108.

[9] Cynthia C.S. Liem et al., "The need for music information retrieval with user-centered and multimodal strategies", *Proceedings of the ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies*, 2011, pp. 1-6.

[10] Charles Inskip, Andy MacFarlane, Pauline Rafferty, "Towards the disintermediation of creative music search: analysing queries to determine important facets", *International Journal on Digital Libraries*, n. 12(2-3), 2012, pp. 137-147.

[11] Savitha H. Srinivasan and Mayank Kukreja, "Tagboards for video tagging", *Proceedings of the ACM international conference on Multimedia*, 2008, pp. 905-908.

[12] Yuh-Lin Chang et al., "Integrated image and speech analysis for content-based video indexing", *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, 1996, pp. 306-313.

[13] Rui Cai, Lie Lu, and Alan Hanjalic, "Unsupervised content discovery in composite audio", *Proceedings of the ACM international conference on Multimedia*, 2005, pp. 628-637.

[14] Robert Mertens et al., "Acoustic super models for large scale video event detection", *Proceedings of the ACM Workshop on Modeling and Representing Events*, 2011, pp. 19-24.