COMPARING AUDIO AND VIDEO SEGMENTATIONS FOR MUSIC VIDEOS INDEXING

Olivier Gillet and Gaël Richard

GET / Télécom Paris
CNRS LTCI,
37 rue Dareau. 75014 Paris, France
[olivier.gillet, gael.richard]@enst.fr

ABSTRACT

Music videos are good examples of multimedia documents in which the structures of the audio and video streams are highly correlated. This paper presents a system that matches these structures and extracts audio-visual correlation measures. The audio and video streams are independently segmented at two-levels: shots (sections for audio) and events. Audio segmentation is performed at the event level by detecting onsets, and at the section level by a novelty detection algorithm identifying instrumentation changes. Video segmentation is performed at the event level by detecting changes in the motion intensity descriptor, and at the shot level by using a classical histogram-based shot detection algorithm. Audio-visual correlation measures are computed on the extracted structures. Possible applications include audio/video stream resynchronization, video retrieval from audio content, or classification of music videos by genre.

1. INTRODUCTION AND MOTIVATIONS

Automatic indexing of audio-visual documents is the process by which high-level descriptors or semantic representations are extracted from the documents. Such descriptors can include, for example, a temporal structuring in shots of the document, a categorization of each shot or of the entire document, a transcription of all the spoken words and close-captions, etc. Depending on the task to be performed either the audio or the video stream is usually considered. However, joint use of both streams has been successfully used for tasks such as discovering underlying concepts [1], classifying television programs [2] or finding segments of news broadcasts where the interviewed subject is on screen [3]. These generic indexing algorithms are suitable for music-related audio-visual content (television broadcasts of concerts, operas or music videos). However, the specificities of this kind of content could be taken into account not only to develop more robust algorithms, but also to increase the level of details of the extracted information. For this purpose, it is important to understand the different kinds of relationships that can exist between the audio and video streams in music-related content, in order to develop relevant audio-visual approaches. Music videos exemplify the variety of possible semantic relationships between audio and video streams: mainstream music videos show dancers or performers, some others have a narrative content based on higher-level features of the song (structure, mood); while directors like Michel Gondry or Spike Jonze explored new forms of visual metaphors.

In this paper, we define different kinds of correlation measures between the temporal structures of the audio and video streams of a music video, by means of automatic segmentation algorithms. Our main motivation is to infer, from these correlation measures, the way in which the video illustrates the music. Possible applications include the detection of music videos showing the musicians, the retrieval of music pieces from a database to match a given video content, or the temporal resynchronization of mismatched audio and video streams. This paper also aims at bridging the gap between generic video indexing systems, and video analysis systems dedicated to specific music tasks (such as in [4]) by characterizing content which could be processed by the latter.

This paper is organized as follows. All the algorithms involved in the structuring of the audio and video streams are detailed in the next section. Section 3 introduces the audio-visual correlation measures derived from this structuring. Experimental results on a music video database are given in section 4, and, finally, section 5 suggests some conclusions and future directions.

2. AUDIO-VISUAL CONTENT STRUCTURING

An overview of our approach is given in Figure 1. Each stream is temporally structured at two levels. As we are trying to evaluate at which level the video content matches the audio content, it is worth precising that it is not possible to use any multimodal approach for the segmentation. For example, a music video could show performers or dancers, with an editing totally decoupled from the music.

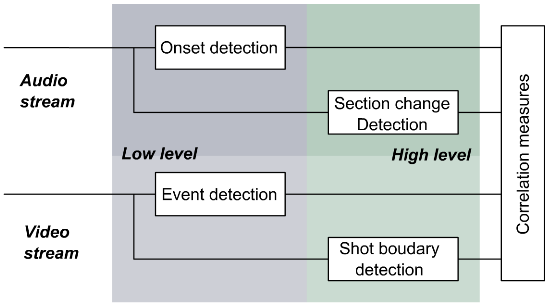

Fig. 1. Overview of the structuring.

2.1. Audio event detection

Some of the most salient events in music signals are note or chord changes, and percussive events. Thus, a good low-level temporal

The authors would like to acknowledge the support of the IST-Kspace network of excellence. This work was partly supported by the MusicDiscover project of the ACI Masse de données.

structuring of a music piece can be achieved by detecting the onsets of such events. As onset detection is an important component of automatic music transcription and beat tracking systems, many approaches have been proposed to solve this problem.

For this work, we used the onset detector introduced by Alonso et al. in [5]. The input audio signal is analyzed by a short-term Fourier transform, and the temporal variation of the energy in each frequency band is computed by applying an optimal FIR differentiation filter, resulting in the so-called spectral energy flux (SEF). All positive contributions of the SEF are summed to produce a detection function $D_{onset}(k)$ that exhibits sharp peaks at note onsets.

2.2. Audio section change detection

At a higher level, a music piece can be temporally segmented in sections, characterized by distinct dynamic or timbral properties and representing the musical structure of the piece in terms of choruses, verses, fill-ins, etc. Finding such a segmentation from a music signal has particularly useful applications in the automatic generation of music summaries. This problem is traditionnally solved by computing a self-similarity matrix of the signal, and by identifying large blocks within it, or detecting boundaries between blocks [6]. In alternative approaches, an arbitrary segmentation is performed and adjacent segments are grouped if their distance or cross-entropy falls below a given threshold.

Our approach is structured on two main building blocks: a compact but well adapted parametrization and a statistically based novelty detection module. Firstly, each frame of the music signal $x(t)$ is parametrized by a features vector. These features include the Mel Frequency Cepstrum Coefficients (MFCC), the zero-crossing rate, and the first four spectral moments. As we try to model long-term phenomena, and to compensate for the periodic and rhythmic variations of the features, frames are rather long (2 seconds). However, in order to gather a large amount of data and allow accurate decisions, 8 frames are computed per second, in overlapping windows. The resulting representation is noted $X_n(k)$, where k is the frame index and n the feature index.

In the next step, a sliding window $W(k_0)$ of length $2L+1$ centered at frame k_0 is considered. k_0 is a good candidate for being a segment boundary if the content of the "future" data set $S_2(k_0) = X_n(k),\ k \in [k_0, k_0+L]$ is novel relatively to the content of the "past" data set $S_1(k_0) = X_n(k),\ k \in [k_0-L, k_0]$. Let P_1 (resp. P_2, P) be the probability density of the features vector for $S_1(k_0)$ (resp. $S_2(k_0)$, $W(k_0)$). Several statistical methods of the litterature are evaluated on their capacity to measure this novelty.

2.2.1. *Bayesian Information Criterion*

The Bayesian information criterion (BIC) is a classical model or order selection criterion, widely use in speech/music or speakers segmentation problems [7], and is thus used in this work as a reference algorithm. In our case, we want to compare the two following models: If k_0 is a segment boundary, the elements of $S_i(k_0)$ are distributed according to P_i, whereas if k_0 is not a segment boundary, the elements of $W(k_0)$ are distributed according to a single distribution P. In the case of Gaussian distributions, the difference of BIC between the two models can be expressed as:

$$\Delta BIC(k_0) = \frac{1}{2}((2L+1)log|\Sigma| - Llog|\Sigma_1| - Llog|\Sigma_2| - K)$$

where the covariance matrices $|\Sigma_i|$ (resp. $|\Sigma|$) are estimated from $S_i(k_0)$ (resp. $W(k_0)$). The constant K is not explicited here as we are only interested in finding local maxima in $\Delta BIC(k_0)$.

2.2.2. *Probabilistic distance*

We expect segment boundaries to be characterized by a higher probabilistic distance between the estimates of the distributions $\hat{P}_1$ and $\hat{P}_2$. In the case of Gaussian distributions, probabilistic distances such as the Bhattacharyya distance or the Kullback-Leibler divergence can be easily computed from the estimated means and variances of P_1 and P_2. However, using such a Gaussian hypothesis would be wrong, since we are dealing with features computed by non-linear functions, and since P_1 or P_2 can be mixtures of several components corresponding to several sound sources. On the other hand, using non-parametric models would require an expensive numerical integration in a high-dimensional domain, which would be limited by the small number of available samples. A solution to this dilemma using kernel methods has been proposed by Zhou and Chellappa in [8]. By mapping the original data to a higher dimensional Reproducing Kernel Hilbert Space in which the Gaussian hypothesis is valid, the analytical expression of the probabilistic distances in the Gaussian case can be used. Such an approach was successfully used by Essid et al. in [9] to compute distances between musical instruments from a large features set and infer taxonomies.

The Bhattacharyya distance was used for this work. Its expression depends on the eigenvalues and eigenvectors of the Gram matrices K_1 and K_2 computed from $S_1(k_0)$ and $S_2(k_0)$ with a Radial Basis Function kernel. Analytical expressions are given in [8].

2.2.3. *Novelty detection with one-class Support Vector Machines*

A likelihood ratio test can determine whether P_1 and P_2 are the same. This test can be written as:

$$R = \frac{\prod_{x \in S_1(k_0)} P_1(x) \prod_{x \in S_2(k_0)} P_2(x)}{\prod_{x \in W} P_2(x)} = \frac{\prod_{x \in S_1(k_0)} P_1(x)}{\prod_{x \in S_1(k_0)} P_2(x)} > t$$

Thus, performing this test requires the estimation of the densities P_1 and P_2. One-class Support Vector Machines (SVM), which aim at identifying a region of the feature space in which most of the data points lie [10], provide robust and accurate estimates in the form:

$$\hat{P}_i(x) = exp(\sum_k \alpha_k^i K(x, X_k^i) + \Theta)$$

where Θ, α_k^i and the support vectors $(X_k^i)_k \subset S_i(k_0)$ are learnt by the SVM algorithm, and K is a reproducing kernel. The test can be simplified by only estimating $\hat{P}_1$, as the numerator of R, indicating how well the 1-class SVM algorithm will fit its own training set, is expected to be constant. For further details, the reader is invited to refer to [11] in which this method is introduced. Practically, as $S_1(k_0)$ and $S_1(k_0+1)$ share L data points in common, it is not necessary to entirely run the 1-class SVM algorithm as the observation window slides. We can rather remove (if necessary) the outgoing point from the set of support vectors, and perform the optimization starting with the existing set of support vectors.

2.2.4. *Grouping adjacent segments*

In order to make each of these segmentation algorithms more robust, an additional step is added in which adjacent segments are merged if the criterion (BIC, probabilistic distance, SVM output) computed between them falls below the decision threshold.

2.3. Video event onset detection

We define video events as short sequences with a high *intensity of action*, occuring within the same single shot. In the context of music videos, such events could be a dance step, a movement made by musicians during the production of a note, as well as action sequences in narrative music videos. The motion activity descriptor introduced in the MPEG-7 standard [12] captures this notion very well, and we thus based our event onset detection on this feature.

A peculiarity of this descriptor is that it can be directly extracted from a MPEG video stream without decoding any individual frame. In MPEG compressed video streams, the temporal redundancy of image sequences is dealt with by coding groups of frames (known as P-frames) as their difference with a motion-compensated prediction from a previous reference frame. Thus, for each block of a P-frame, a motion vector is available. However, the raw motion vectors are very noisy and unreliable. Two processings are necessary to make them suitable for motion description. Firstly, the motion vectors for non-textured part of the pictures are removed, since the motion estimation algorithms used in MPEG coders is only reliable on textured areas. Non-textured blocks can be easily identified in the compressed domain by their DCT coefficients. Secondly, a median filter is applied to the motion vector field in order to remove outliers. The motion activity descriptor is defined as the standard deviation of the magnitude of the motion vectors, on a 5-grade scale. In order to detect event onsets, we did not quantize the motion activity $MA(k)$, but rather detected its sharp variations by using a differentiation filter h to form a detection function $D_{event}(k) = MA(k) * h(k)$.

2.4. Video shot segmentation

Many algorithms have been proposed to segment a video document in shots - see for example the report of the last annual TRECVid evaluation [13] for a review and comparison of state-of-the-art algorithms. In this work a simple histogram-based approach was used, as it shows good performances for cut detections. For each frame, three 16-bins histograms were computed from the Y, U and V components, resulting in a vector of $N = 48$ features $X_n(k)$. A shot boundary was detected when $D^0_{shot}(k) = \sum_n |X_n(k) - X_n(k-1)| > t$. The threshold t is dynamically adjusted at each frame by applying a median filter to $D^0_{shot}(k)$ across a large window. This is equivalent to using a static threshold with the detrended detection function $D_{shot}(k) = -t_0(k) + \sum_n |X_n(k) - X_n(k-1)|$.

3. AUDIO-VISUAL CORRELATION MEASURES

While it would be possible to compute correlation measures from the list of segments, shots, sections and events boundaries, we decided to follow another approach by directly using the correlations between their respective detection functions - as this approach results in no information loss, and overcomes temporal resolution problems. Moreover, the magnitude of each peak in the detection functions is a good clue of the saliency of the corresponding event, an information worth taking into account.

The general form of these correlation measures is:

$$C_{A/B} = \frac{1}{\sqrt{\sum_k A(k) \sum_k B(k)}} \sum_k A(k)B(k)$$

where the normalization factor $\frac{1}{\sqrt{\sum_k A(k) \sum_k B(k)}}$ favors the matching of detection functions with a similar number of peaks, and penalizes the matching of detection functions with an unrelated number of peaks.

Four correlation measures were defined, corresponding to different relationships between the audio and video content. Firstly, for $A(k) = D_{section}(S_1(k), S_2(k))$ and $B = D_{shot}$ the section/shot change correlation expresses how often a new section of the music is introduced by a shot change.

Secondly, the section change/event correlation defined for $A(k) = D_{section}(S_1(k), S_2(k))$ and $B = D_{event}$ expresses how often section changes in the music are visually translated by changes in the motion activity.

Then, the note onset/shot change correlation defined for $A = D_{onset}$ and $B = D_{shot}$ expresses the synchronicity of shot changes with the rhythmic content of the music.

Finally, the note onset/event correlation expresses the synchronicity of salient motion changes with the rhythmic content of the music. It is defined for $A = D_{onset}$ and $B = D_{event}$.

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1. Evaluation of the audio section segmentation

Since evaluations of the other components are already available [13], only the audio section segmentation algorithm was independently evaluated. For this purpose, 60 full-length pop music signals of various genres were manually segmented according to their structure. The three algorithms presented in 2.2 were evaluated on this database by computing the precision and recall rates for different values of the decision threshold. A sliding window of 15 seconds was used. Section boundaries are considered as correctly detected when they are detected within 2 seconds of the reference boundary. The results are summarized in Figure 2: The algorithm based on the probabilistic distance outperforms the 1-class SVM and the BIC approaches. However, the precision of the 1-class SVM algorithm decreases more slowly as the recall rate increases. In the case of the BIC, the loss of performances can be explained by the lack of data from the past window ; and the invalidity of the Gaussianity hypothesis.

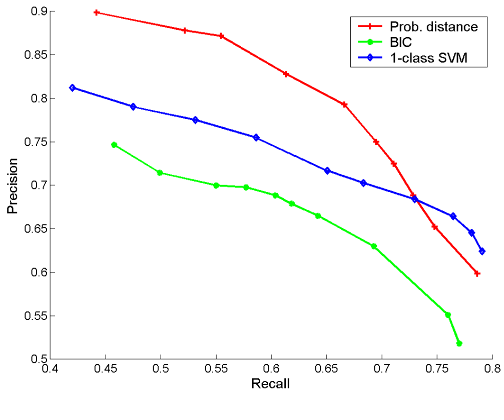

Fig. 2. Recall/Precision curves for the audio segmentation task.

4.2. Experiment 1: Video retrieval from audio

This experiment aims at evaluating whether it is possible, given an audio stream x, to find the corresponding video stream from a music videos database. For this purpose a small database of 30 music videos was considered. All the videos were cut to their minimal common length (1 min 50). Given an audio query x, the videos v_i are ranked according to an audio-visual correlation measure $C(x, v_i)$. The result set consists of the N first ranked videos - different values of N allowing for different precision/recall ratios. The results are given in Figure 3. It can be seen that the best measure for matching the audio and video content is the correlation between section changes in the music and shot boundaries.

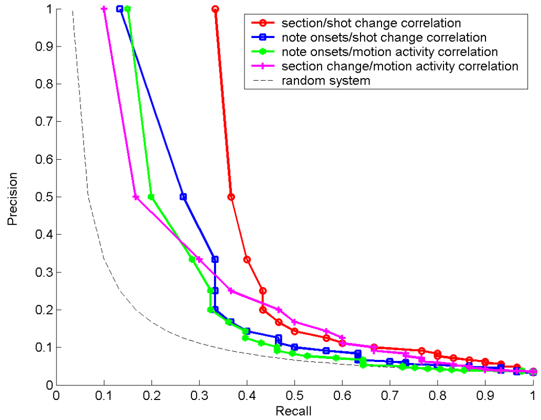

Fig. 3. Recall/Precision curves for the video retrieval experiment.

4.3. Experiment 2: Audio-visual association characterization

The 30 music videos used in the previous experiment were manually classified into 4 categories: narratives, dance/performance, video sampling (in which video excerpts are triggered according to the music score) and abstract compositions - videos without narrative content or obvious relationship with the performance of the music. Each music video is represented in a 2D plane using two correlation measures as axes (see an example on Figure 4). It can be seen that the correlation measures defined above can help in the characterization of music videos according to the association relationship between their video and audio content. Especially, the correlation between note onsets and motion activity is typical of music videos produced by video sampling techniques.

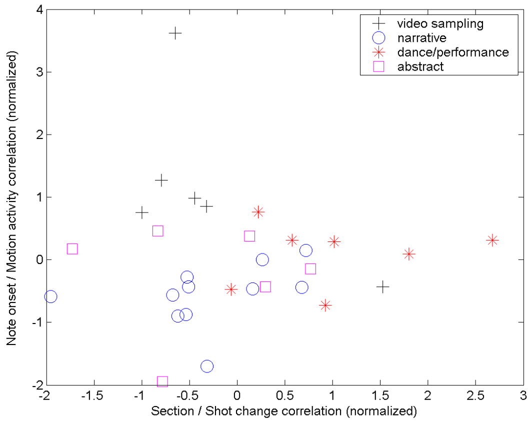

Fig. 4. Characterization of audio-visual associations.

5. CONCLUSION AND FUTURE WORK

This paper presented different audio and video segmentation algorithms operating at two levels - individual events and shots or sections. We defined, from their output, four correlation measures between the audio and video streams. As a preliminary experiment showed, they can be used as ranking functions for cross-modal queries. These correlation measures can also help in the characterization of music videos, as they are strongly related to the semantic relationships between the music and the way in which it is illustrated. This work is a first step toward music-aware video indexing systems, capable of identifying in music or concert videos which sequences are suitable for multimodal music analysis.

6. REFERENCES

[1] L. Xie, L. Kennedy, S.-F. Chang, A. Divakaran, H. Sun, and C.-Y. Lin, "Discovering meaningful multimedia patterns with audio-visual concepts and associated text," in *Proc. Int. Conf. Image Processing*, 2004.

[2] J. Huang, Z. Liu, , Y. Wang, and T, "Classification of tv programs based on audio information using hidden markov model," in *IEEE Workshop on Multimedia Signal Processing*, 1998.

[3] A. Albiol, L. Torres, and E. Delp, "Combining audio and video for video sequence indexing applications," in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2002.

[4] O. Gillet and G. Richard, "Automatic transcription of drum sequences using audiovisual features," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, March 2005.

[5] M. Alonso, G. Richard, and B. David, "Extracting note onsets from musical recordings," in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2005.

[6] G. Peeters, A. La Burthe, and X. Rodet, "Toward automatic music audio summary generation from signal analysis," in *Proc. 3rd Int. Conf. Music Information Retrieval*, 2002.

[7] B. Zhou and J.H.L. Hansen, "Unsupervised audio stream segmentation and clustering via the bayesian information criterion," in *Proc. Int. Conf. Spoken Language Processing*, 2000.

[8] S. Zhou and R. Chellappa, "From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel hilbert space," in *IEEE Trans. Pattern Analysis and Machine Intelligence*, to be published.

[9] S. Essid, G. Richard, and B. David, "Instrument recognition in polyphonic music based on automatic taxonomies," in *IEEE Trans. Speech and Audio Processing*, to be published.

[10] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," in *Adv. in Neural Information Processing Systems 12*, 1999.

[11] G. Loosli, S. G. Lee, and S. Canu, "Context changes detection by one-class svms," in *Workshop on Machine Learning for User Modeling*, 2005.

[12] S. Jeannin and A. Divakaran, "Mpeg-7 visual motion descriptors," in *IEEE Trans. Circuits and Systems for Video Technology*, 2001, vol. 11, pp. 720–724.

[13] W. Kraaij, A. F. Smeaton, P. Over, and J. Arlandis, "Trecvid 2004 - an overview," National Institute of Standards and Technology (NIST), 2005.