SEGMENTATION OF MUSIC VIDEO STREAMS IN MUSIC PIECES THROUGH AUDIO-VISUAL ANALYSIS

Gabriel Sargent, Pierre Hanna, Henri Nicolas

Université de Bordeaux, LaBRI - UMR 5800, F-33400 Talence, France

ABSTRACT

Today, technologies for information storage and transmission allow the creation and development of huge databases of multimedia content. Tools are needed to facilitate their access and browsing. In this context, this article focuses on the segmentation of a particular category of multimedia content, audio-visual musical streams, into music pieces. This category includes concert audio-video recordings, and sequences of music videos such as the ones found in musical TV channels. Current approaches consist in supervised clustering in a few audio classes (music, speech, noise), and, to our knowledge, no consistent evaluation has been performed yet in the case of audio-visual musical streams. In this paper, we aim at estimating the temporal boundaries of music pieces relying on the assumed homogeneity of their musical and visual properties. We consider an unsupervised approach based on the generalized likelihood ratio to evaluate the presence of statistical breakdowns of MFCCs, Chroma vectors, dominant Hue and Lightness over time. An evaluation of this approach on 15 manually annotated concert streams shows the advantage of combining tonal content features to timbral ones, and a modest impact from the joint use of visual features in boundary estimation.

***Index Terms*—** Multimedia signal processing, segmentation, audio-visual stream, music video

1. INTRODUCTION

The development of communication and information technologies allows the storage and broadcasting of large collections of audio-visual contents. A large part of these collections consists in audio-visual musical streams, *i.e.* concert recordings and music video playlists broadcasted through internet services or TV channels. We focus here on the estimation of the temporal boundaries (start time, end time) of western popular music pieces occurring in such streams. Such an estimation can be useful to navigate within the stream (automatic chaptering) and extract statistical informations from it (e.g. providing the number of music pieces and their occurrences). Moreover, it can help the cross-referencing of music pieces from different multimedia documents for copyright protection.

This work was supported by the Mex-Culture project (ANR-11-IS02-0001).

Estimating the boundaries of music pieces within an audio stream is a difficult problem : instrumental breakdowns can be introduced on purpose by the band during concerts, or by the video producer for scripting issues. On the opposite, music pieces can be played successively without any pause between them, keeping locally similar properties such as a stable timbre or a constant tonality. One can wonder if the combination of both timbre and tonal features improves the estimation, as well as information provided by other modalities. For example, the video track associated to a music piece generally consists in a sequence of shots taken from a limited number of ambiances and environments. This article explores this issue through the evaluation of an unsupervised approach combining descriptions of the audio and visual modalities in terms of Mel-Frequency Cepstral Coefficients, Chroma vectors, Hue and Lightness.

In section 2 is presented an overview of the existing and related approaches for music piece boundary estimation. Section 3 describes the problem statement and working assumptions. The proposed approach is described in section 4 and evaluated according to different configurations in section 5.

2. RELATED WORKS

Few approaches have been proposed to locate temporal boundaries of music pieces in audio-visual musical streams. [1] and [2] present browsing and summary generation systems which include such a segmentation based on audio and visual features, whose full description and evaluation are beyond the scope of these papers. For both of them, the segmentation is driven by the segmentation of the audio which is then corrected thanks to visual features (color, lightness).

Audio-based approaches can be found in [3] and [4]. Both cases require to divide the audio stream into short frames classified according to a Support Vector Machine. The audio classes are defined using temporal or spectral features such as zero crossing, MFCC and LPC. No tonal feature is used. In [3], the segmentation is refined according to several heuristics : duration and metadata guides the fusion or division the obtained segments. In [4], the segmentation is obtained from the adaptive thresholding of an homogeneity criterion built from the frame classification along with the RMS energy curve calculated over time. The final segmentation is obtained by searching the most probable segmentation through Bayesian inference.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

The localization of music pieces within musical streams can be related to music/speech/noise discrimination. Indeed, two successive music pieces are often marked with applauses, speech or periods of silence. However, such audio events can be found within music pieces : applauses at the end of instrumental solos, instrumental interruptions during live songs, speech and silent parts resulting from video production (see for example Lady Gaga's *Paparazzi* music video)... On the opposite, songs may be played without timeout. Few of these approaches are based on musical properties.

The approach in [5] segments concert videos according to a categorization of key frames according to visual objects, e.g. musical instruments, band members... This is achieved by a SVM with visual features and video production features. This article concludes on the efficiency of visual features compared to production ones.

Generic approaches have been proposed for video segmentation using audio and video in the scope of scene detection [6]. They mainly consist in a shot segmentation step from visual analysis, and a scene segmentation obtained by the grouping of contiguous shots with similar audio-visual properties. The ranking of these approaches according to their performance remains a difficult issue as existing evaluation databases, which don't contain musical streams, vary from one work to another.

3. PROBLEM SPECIFICATION

The audio-visual musical streams considered consist in a sequence of music pieces associated to a visual stream such as live or scripted scenes. A pop music piece can be described as a temporal object built on related smaller objects in relationship [7]. The associated visual stream generally exhibits a limited number of ambiances or environments per music piece, which can be described with global properties such as colors or lightness. Such properties may change significantly from one piece to another.

Assuming that the audio and video streams show statistically stable global properties, we model the whole audio-visual music stream as a sequence of homogeneous segments over time, in terms of timbre, tonal content, color and lightness. As a music piece never appears twice in a row, we assume that the features of two consecutive pieces have different statistics. This leads us to characterize the temporal boundary between two segments by a statistical breakdown of the audio-visual properties of the stream.

Two music pieces can be separated by non-musical segments such as silence, crowd noises and speech, which we also consider as globally homogeneous segments in the scope of our problematic.

4. SEGMENTATION APPROACH

The segmentation approach is composed of two main steps : the extraction of audio and visual features, and the estimation of the temporal boundaries of music pieces through the calculation and combination of homogeneity breakdown criteria.

4.1. Audio and visual features

We describe the music video stream as a sequence of audio and visual features. As they are extracted from different modalities with different time resolutions, we choose to express them at a common time-scale, empirically set to a sampling period of 0.5 seconds.

We consider musical properties of the audio through the use of Mel-Frequency Cepstral Coefficients (MFCC) and Chroma vectors. A vector of MFCCs is obtained by filtering the log-power spectrum of a signal with bandpass filters whose frequency responses are regularly spaced at the Mel frequency scale. This filtered spectrum is then decomposed with a discrete cosine transform. The resulting set of coefficients roughly describes the spectral envelope of the input signal [8] and it is often considered as a way to describe its overall musical timbre [9]. A Chroma vector is a set of coefficients which quantizes the energy associated to the twelve semi-tones of the chromatic scale over the signal's whole spectrum in western music theory [10]. They constitute a description of the tonal content of the input signal. An homogeneous sequence of chroma vectors over time can be interpreted as the use of local key.

The visual part of the musical stream is described as a sequence of dominant color and lightness values over time. We consider an image through its Hue Lightness Saturation (HLS) model. The dominant color of an image corresponds to the most represented value of Hue in the image (in practice the index of the maximal value in the Hue histogram). The dominant Lightness is obtained using the same process for the Lightness component. The Saturation component, associated to more subtle color changes, is currently left apart.

4.2. Statistical breakdown criterion

As assumed in section 3, the boundary between two music pieces is reflected by a statistical breakdown of the stream's properties over time. We therefore evaluate for each time instant t if it coincides with a statistical breakdown of its neighboring features.

Let $y = \{y_n\}_{1 \leq n \leq 2N}$, $N \in \mathbb{N}$ be the sequence of feature vectors contained within an analysis window centered on t, composed of two parts $y^1 = \{y_1, ..., y_N\}$ (neighboring features before t) and $y^2 = \{y_{N+1}, ..., y_{2N}\}$ (neighboring features after t). y is assumed to be a sequence of observations generated by a sequence of independent random variables $Y = \{Y_n\}_{1 \leq n \leq 2N}$ under antagonistic assumptions H_0 and H_1. As in [11] for music vs. speech discrimination, the presence of a statistical breakdown at t is evaluated through the logarithm of the Generalized Likelihood Ratio (GLR), defined as :

$$\log(\text{GLR}) = \log \frac{P(y|H_1)}{P(y|H_0)} = \log \frac{P(y^1|G_1)P(y^2|G_2)}{P(y|G_0)} \quad (1)$$

where H_0 assumes that y can be modeled with a single Gaussian distribution $G_0 = G(\mu, \Gamma)$ (homogeneity assumption), and H_1 assumes that y^1 and y^2 can be modeled by two different Gaussian distributions $G_1 = G(\mu_1, \Gamma_1)$ and $G_2 = G(\mu_2, \Gamma_2)$ (breakdown assumption). log(GLR) is maximal when the likelihood of H_1 is high, which implies that the likelihood of H_0 is low.

4.3. Boundary selection

The boundary estimation consists in a peak selection procedure. A breakdown criterion [11] is first calculated from the homogeneity curves to obtain a set of dominant peaks. The number of highest dominant peaks selected is fixed in proportion to the total number of peaks, as it can be observed that long streams contain more music pieces. We store all dominant peaks according to descending order and define the parameter :

$$\eta = \frac{\text{number of selected peaks}}{\text{total number of dominant peaks}} \quad (2)$$

We may notice that the η acts as an adaptive threshold selection parameter of the breakdown criterion.

4.3.1. Feature and criteria fusion

The homogeneity breakdown criterion is computed on each feature type (MFCC, Chroma vector, dominant Hue and Lightness), and for each modality. In the second case, the criteria are respectively calculated on the concatenation of the feature vectors of each modality, as we assume their statistical independence. Then, a fusion of the modalities is considered through the linear combination of the normalized criteria obtained for each modality (linear weighted fusion [12]).

Let $\{\phi_A\}$ and $\{\phi_V\}$ be the normalized criteria respectively associated to the audio and visual modalities, and $\lambda \in [0, 1]$ a weighting parameter to tune their relative importance in the segmentation process. The criterion ϕ_{AV} resulting from their combination is defined as :

$$\phi_{AV} = \lambda\phi_A + (1 - \lambda)\phi_V. \quad (3)$$

The criteria are normalized by dividing their values of the criteria according to their 9^{th} decile.

5. EVALUATION

5.1. Evaluation database

The evaluation database consists in 15 concert videos from DVD and TV channels referenced in Table 1. They were annotated manually with the ELAN software [1], setting a segment boundary at the beginning and end of each music piece (appearance/disappearance of singing or instrumental sounds). A period of time between music pieces containing speech, silence or crowd noises is considered as a single segment.

Artist	Title	Year
Amy Winehouse	I Told You I Was Trouble, Live In London	2007
Depeche Mode	Live One Night in Paris	2002
Florence + The Machine	Royal Albert hall	2012
Foo Fighters	Live on Letterman	2011
Genesis	The Way We Walk	1992
Jamiroquai	MTV EXIT FESTIVAL	2011
Keane	Live at the O2 Arena London	2007
KISS	Monster Tour - Live in Zurich	2013
Madonna	Sticky Sweet Tour	2010
Muse	Hullabaloo Live at Zenith (Paris)	2002
Norah Jones	Live in Amsterdam	2007
Simply Red	Live in London	1998
The Cranberries	Beneath the Skin Live in Paris	1999
The Police	Live In Concert At The Tokyo Dome	2008
U2	Go Home : Live from Slane Castle	2003

Table 1. Evaluation database : list of concerts (from DVD and live streaming) considered for evaluation.

5.2. Evaluation metrics

The accuracy of a music piece boundary estimation is measured with the Precision P, Recall R and F-measure F. Be b_A the set of boundaries of the reference segmentation (manually annotated) and b_E the set of estimated ones. They are defined as :

$$P = \frac{|b_E \cap b_A|}{|b_E|} ; R = \frac{|b_E \cap b_A|}{|b_A|} ; F = \frac{2PR}{(P+R)}.$$

We restrict the match of an estimated (resp. reference) boundary to a unique reference (resp. estimated) one, as in [13] with *boundary hit rates*. Considering the granularity of our problem, we consider a tolerance window of τ =10 s.

5.3. Evaluation process

The quality of a criterion is evaluated through a cross-validation process [14] : the dataset is randomly divided in five folds of three concert videos. The system is evaluated on each fold after the tuning of its parameters on the four others. The global performances are obtained by the computation of the average of the performance values obtained for the five folds.

5.4. Implementation details

13 MFCCs (including 0^{th} order) and Chroma vectors of size 12 are regularly extracted from the audio using Yaafe [15], with respective hop sizes of 1024 and 2048 points, and an analysis window size of 2048 points for the MFCCs [2]. These features are then expressed at the timescale with period $\Delta t = 0.5$ s by taking the mean of the vectors contained in every window of duration 1 s centered on a multiple of 0.5 s.

1. `http://tla.mpi.nl/tools/tla-tools/elan/`

2. Other MFCC and Chroma vectors extraction parameters are set as the default ones in Yaafe.

Criterion	$F(\%)$	$P(\%)$	$R(\%)$	η for all folds (%)
ϕ_M	44.13	46.91	42.53	2.00, 1.60, 1.85, 1.85, 1.40
ϕ_C	50.40	59.45	45.34	2.25, 1.45, 1.85, 1.60, 1.05
ϕ_A	**52.86**	65.02	45.54	1.85, 1.15, 1.25, 1.35, 1.00
ϕ_H	24.60	19.30	36.62	6.85, 6.05, 2.50, 5.05, 6.35
ϕ_L	**29.44**	31.95	30.36	3.10, 0.90, 3.00, 2.70, 2.60
ϕ_V	27.36	21.57	44.18	2.85, 4.80, 4.30, 6.20, 10.35

Table 2. Average performances on 15 concerts obtained from the cross evaluation process for the homogeneity breakdown criteria ϕ_M (MFCC), ϕ_C (Chroma), ϕ_A (MFCC and Chroma vectors concatenated), ϕ_H (dominant Hue), ϕ_L (dominant Lightness) and ϕ_V (dominant Hue and Lightness concatenated). The values of the peak selection parameter η, considered with a resolution of 0.05, are related for the five folds.

Fold	$F(\%)$	$P(\%)$	$R(\%)$	λ	η (%)
1	61.47	66.92	57.21	0.9	1.75
2	46.04	66.29	35.49	0.8	0.85
3	53.33	66.74	45.06	0.9	1.20
4	58.30	72.14	49.17	0.9	1.50
5	55.98	77.97	44.39	0.9	1.05
Average	**55.02**	70.01	46.26	-	-

Table 3. Performances from the cross-validation of the multimodal criterion ϕ_{AV} on the 15 concert videos, with associated values of weight λ and peak selection parameter η obtained are related for the five folds.

The video stream is sampled at a period of Δt =0.5 s. Hue and Lightness histograms were generated from each sampled frame using the OpenCV open-source library [3].

The audio and visual features are analyzed through the computation of the criterion described in paragraph 4.2 with an analysis window empirically set to 60 s [4].

5.5. Performances per feature type and modality

Table 2 gathers the average performances obtained for the different criteria by cross-validation.

The average F-measure obtained with ϕ_C overpasses the one for ϕ_M, which provides the use of Chroma vectors compared to MFCCs. However, it must be noted that for some concerts, the use of ϕ_M overcomes ϕ_C such as the Simply Red concert where ϕ_M obtains $F = 53.97\%$ compared to $F = 47.06\%$ for ϕ_C. This fact exhibits that they can provide complementary information for music piece segmentation. Their joint use through concatenation (ϕ_A) brings a slight improvement by over 2% of the average F-measure.

Results obtained with video properties are more modest. The average F-measure of ϕ_L is better than the one obtained by ϕ_H over 5%, but the concatenation of dominant Hue and Lightness (ϕ_V) don't improve the performances. The relative performance associated to ϕ_L and ϕ_H can vary according to the musical stream : for example, ϕ_H obtains $F = 39.34\%$ and ϕ_L leads to $F = 28.99\%$ in the case of Amy Winehouse's concert.

5.6. Performances for combined modalities

Table 3 exhibits the results of the cross-validation of the criterion ϕ_{AV} resulting from the linear combination of ϕ_A and ϕ_V. The average F-measure obtained by ϕ_{AV} is slightly better than for ϕ_A over 2%. It can be noted that each training step tunes λ around 0.8 and 0.9, which shows the prevalence of the audio criterion compared to the visual one. Therefore the use of dominant Hue and Lightness does not improve the boundary estimation in a significant way. An exception can be noted with Depeche Mode, where the cross-validations give$F = 34.92\%$ for ϕ_A, $F = 36.70\%$ for ϕ_V, and $F = 53.93\%$ for ϕ_{AV} with $\lambda = 0.9$ and $\eta = 1.75\%$. The values of η remain stable, around 1% on the considered dataset.

5.7. Influence of crowd noises and speech : case study

The influence of crowd noises and speech between songs is studied for the concert of Norah Jones. Its audio track have been extracted and edited to remove these markers as well as the song introductions to build a continuous musical audio stream. The analysis of the full audio track lead to $F = 53.57\%$ for ϕ_A, $F = 53.57\%$ for ϕ_M, and $F = 51.85\%$ for ϕ_C. The edited audio track gives $F = 76.92\%$ for ϕ_A, $F = 46.15\%$ for ϕ_M, and $F = 83.33\%$ for ϕ_C. As we could expect, ϕ_M is more competent in finding the order between crowd noises and speech, but ϕ_C tend to segment music pieces successfully. These values show that our approach is more efficient on a stream without interruption between music pieces.

6. CONCLUSION

In this article, we focused on the estimation of temporal boundaries of music pieces within audio-visual musical streams. The presented approach measured the presence of segment boundaries by detecting statistical breakdowns of musical and visual properties over time. This approach, based on the calculation of a generalized likelihood ratio, was evaluated considering separated and combined features, and exhibited the efficiency of using tonal features such as Chroma vectors in complement of timbral features such as MFCCs. The joint analysis of dominant Hue and Lightness through the linear combination of the associated criteria did not brought a significative improvement in the estimation of boundaries. However, this straightforward combination could be improved as future work by exploring possible dependencies between these modalities, e.g. using copula models.

3. `http://opencv.org/`

4. A slight Gaussian noise have been artificially added to the features in order to avoid sequences of repeated feature vectors, in particular when a silence period occurs at the end of the recording

7. REFERENCES

[1] Y. van Houten, U. Naci, B. Freiburg, R. Eggermont, S. Schuurman, D. Hollander, J. Reitsma, M. Markslag, J. Kniest, M. Veenstra, and A. Hanjalic, "The MultimediaN concert-video browser," in *IEEE International Conference of Multimedia and Expo (ICME)*, 2005, pp. 1561–1564.

[2] L. Agnihotri, N. Dimitrova, and J. R. Kender, "Design and evaluation of a music video summarization system," in *IEEE International Conference of Multimedia and Expo*. IEEE, 2004, pp. 1943–1946.

[3] R.W. Ferguson III, "Automatic segmentation of concert recordings," M.S. thesis, Mc Gill University, 2004.

[4] M. Marolt, "Probabilistic segmentation and labeling of ethnomusicological field recordings," in *Proceedings of the 10h International Society for Music Information Retrieval (ISMIR)*, 2009, pp. 75–80.

[5] C.G.M. Snoek, M. Worring, A. W. M. Smeulders, and B. Freiburg, "The role of visual content and style for concert video indexing," in *IEEE International Conference of Multimedia and Expo (ICME)*, 2007, pp. 252–255.

[6] Y. Kompatsiaris, B. Merialdo, and S. Lian, Eds., *TV Content Analysis : Techniques and Applications*, chapter 6 : TV program structuring techniques, CRC Press, 2012.

[7] Frédéric Bimbot, Emmanuel Deruty, Gabriel Sargent, and Emmanuel Vincent, "Semiotic Structure Labeling of Music Pieces : Concepts, Methods and Annotation Conventions ," in *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, Oct. 2012, pp. 235–240.

[8] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proceedings of the 2nd International Symposium on Music Information Retrieval (ISMIR)*, 2000.

[9] J. Paulus, M. Muller, and A. Klapuri, "Audio-based music structure analysis," in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, August 2010, pp. 625–636.

[10] M. A. Bartsch and G. H. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," in *IEEE Transactions on multimedia*, February 2005, vol. 7, pp. 96–104.

[11] M. Seck, R. Blouet, and F. Bimbot, "The IRISA/ELISA Speaker Detection and Tracking Systems for the NIST'99 Evaluation Campaign," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 154–171, Jan. 2000.

[12] K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis : a survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, April 2010.

[13] D. Turnbull, G. Lanckriet, E. Pampalk, and M. Goto, "A supervised approach for detecting boundaries in music using difference features and boosting," *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, pp. 51–54, 2007.

[14] Rupert G. Miller, "The jackknife – a review," *Biometrika*, vol. 61, no. 1, pp. 1–15, April 1974.

[15] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software," in *Proceedings of the 11th International Society for Music Information Retrieval (ISMIR)*, 2010, pp. 441–446.