Audio, Video and Audio-Visual Signatures for Short Video Clip Detection: Experiments on Trecvid2003

Benjamin Senechal[(1)], Denis Pellerin[(1)], Laurent Besacier[(2)], Isabelle Simand[(3)], Stéphane Brès[(3)]

[1] LIS, 46 av. F. Viallet, 38031 Grenoble Cedex, France
[2] CLIPS-IMAG, BP 53, 38041 Grenoble Cedex 9, France
[3] LIRIS, bat. J. Verne, INSA, 69621 Villeurbanne Cedex, France
{Benjamin.Senechal,Denis.Pellerin}@lis.inpg.fr, Laurent.Besacier@imag.fr

ABSTRACT

In this paper, we present the association of audio and video signatures for short video clip detection. First, we present an audio signature based on the spectral flatness measure. Then we describe a spatio-temporal video signature, based on the evolution of gray level centroids over time. The major contribution of this work is the association of these two signatures in a so-called audio-visual signature by late integration of similarity measures obtained on both modalities. Our experiments conducted on a large video database (28Gb / 34h extracted from TRECVID2003) show that our audio-visual signature is more robust than the audio-only or video-only signatures, and also permits better detection of video clips of shorter duration (about 2 seconds).

1. INTRODUCTION

The main goal of the TREC Video Retrieval Evaluation (TRECVID) is to promote progress in content-based retrieval from digital videos via open metrics-based evaluation [6]. For this, many features, extracted from audio, video or text obtained with automatic speech recognition (ASR), can be used. While automatic speech recognition can be very useful to extract keywords and perform topic detection on a video stream, this technology has the drawback of not being easily transposable from one language to another. As an alternative, simpler technologies, consisting in detecting pertinent key sounds, so-called jingles, can reveal very interesting semantic information. For example, detecting automatically a particular jingle on a TV broadcast program can announce a particular sequence like the weather forecast.

Jingle detection can be related to a recent domain also named 'audio monitoring' or audio fingerprinting [2]. It is generally based on the creation of a compact representation (signature) of each multimedia document in a database. The particularity of audio fingerprinting is that the search is generally based on a strong matching criteria instead of a similarity one. A similar process can also be applied to the video channel of a multimedia document [3]. In this case, we call it 'video fingerprinting'.

In this paper, we discuss the interest of the association of audio and video signatures for short video clip detection. We first propose an audio signature which has the advantage of being relatively compact, since it uses only 8 acoustic parameters per frame, corresponding to the spectral flatness feature which is part of the MPEG-7 low level description [1]. We also describe a spatio-temporal video signature, based on the evolution of gray level centroids over time. This signature is compact (8 coefficients per frame). The major contribution of this paper is the proposal of an audio-visual signature which combines both audio and video signatures. To our knowledge, no real audio-visual signature has been proposed in the literature, except in [4] where a cross-modal association was proposed for multimedia content processing. All three audio, video and audio-visual signatures are systematically evaluated on a same 34-hour database which is part of the TREC2003 video database.

In section 2, we describe the audio signature, as well as the general video clip detection system. Section 3 is dedicated to the description of the video signature proposed while section 4 presents the audio-visual signature. Section 5 describes the experiments and results and section 6 concludes this work.

2. AUDIO SIGNATURE

Each reference video clip was described by low level descriptors based on a spectral analysis while dissimilarity was measured between the target video clips and the whole video test set with a Euclidian distance. In the next subsections, we describe the audio signature more precisely.

2.1. Feature Extraction

Our low level descriptors were 8 coefficients corresponding to the spectral flatness feature computed on 8 frequency bands. This spectral flatness feature is part of the MPEG-7 low level description and has shown to give

interesting results for audio fingerprinting [1]. We have observed that the spectral flatness measure (SFM) is more efficient that the simple spectral coefficients by channel. Indeed, the right minima are easier to detect on the distance curve (Fig. 2.a) because the noise is reduced. Moreover, the number of coefficients necessary is much lower. As a comparison, 29 spectral parameters are used in [5] for jingle detection. The formula of the SFM for the frequency band k is given by *equation 1*, where $c(i)$ are filter bank coefficients corresponding to the spectral energy computed on n different frequency subbands inside frequency band k.

$$SFM_k = \frac{\sqrt[n]{\prod_{i=1}^{i=n} c(i)}}{\frac{1}{n}\sum_{i=1}^{i=n} c(i)} \qquad (1)$$

The so-called SFM is a function which is related to the tonality aspect of the audio signal and can therefore be used as a discriminating criterion between different audio signals.

It is also important to note that our frequency scale is not linear but logarithmic (we actually use the MEL scale, widely used in speech processing). The parameters are extracted on 64ms windows with a window shift of 33.375ms, corresponding approximately to the video frame rate. This window shift analysis will allow both audio and video signatures to be aligned which is particularly important to design the audio-visual signature.

2.2. Short video clip detection

The video clip detection principle is the same for the audio, video or audio-visual signature. It is described more precisely in the following subsections.

2.2.1. Distance computation

A reference video clip (the *request*) is characterized by a sequence of M 8-dimensional vectors which is called the signature of the clip. The size M is the number of analysis frames and corresponds here to the number of images on the video clip, because of the window shift value chosen. The detection consists in finding this sequence in the data flow (the *database*), which is also transformed in a large sequence of N 8-dimensional vectors (N>>M). The signature and the data flow are compared using an Euclidean distance, we then obtain N-M+1 distance values given by:

$$d_h = \sum_{i=1}^{i=M}\sum_{j=1}^{j=8} |a_{(i-1+h)j} - b_{ij}| \quad , 1 \leq h \leq N\text{-}M\text{+}1 \quad (2)$$

where a_{kj}, is the j-th SFM coefficient of vector k in the data flow signature (k=1..N, j=1..8),

and b_{ij}, is the j-th SFM coefficient of vector i in the video clip request signature (i=1..M, j=1..8).

2.2.2. Minima detection on the distance curve

The goal of this step is to find minima (valleys) on the distance curve computed in the previous step. An adaptative threshold is first applied. It is function of the mean and standard deviation of the distance curve. Then, the final video clip detection decision is made according to the width of the valleys detected, since it was shown in [5] that the wider valleys may correspond to false alarms.

3. VIDEO SIGNATURE

We have chosen to characterize each video by calculating the gray level centroid of each frame of the video. The coordinates (c_x, c_y) of the centroid are calculated as the weighted sum of the pixel positions, the weights being the gray values of the pixels:

$$c_x = \frac{\sum_{i=0}^{size\_1}(L_i.x_i)}{\sum_{i=0}^{size\_1} L_i} \qquad c_y = \frac{\sum_{i=0}^{size\_1}(L_i.y_i)}{\sum_{i=0}^{size\_1} L_i} \qquad (3)$$

where size is the number of pixels considered in the image, and L_i the gray level or the luminance of the pixel. Then, these coordinates are standardized to a single frame size. The resulting point is obviously not significant for one frame, but becomes characteristic for several successive frames, due to its movement during a few seconds. As a consequence, it leads to a spatio-temporal signature (Fig. 1).

As a video can be related to a complex scene, a single centroid is not sufficient to characterize it. We thus propose to divide the frames into several sub-images. Based on our experiments, dividing a frame into four quarters is the best compromise between the complexity of the signature and its discriminative power. For each quarter, a centroid is calculated. This allows us to keep minimal spatial information on the gray level repartition.

The computation of the centroids' location is biased toward brighter areas in order to amplify the centroids' movements. This is done by a look-up table, without any additional cost *(Eq. 4)*. Then, the motion is less uniform and easier to be discriminated.

$$f(L) = \frac{L^3}{255^2} + 1 \qquad (4)$$

where L is the original gray level between 0 and 255.

The signature of an image is composed of the eight values of the four centroids' coordinates. The signature of a video consists of the signatures of its images.

We have sub-sampled the image horizontally and vertically to 1 pixel out of 8 to speed up the computation. This selection is founded on the redundancy due to the gray level correlation of neighboring pixels.

The unicity of the signature cannot of course be proved but we shall assume this unicity as a consequence of the complexity of the signature: *8.N* pixel locations, where *N*

is the number of frames. For short video clips, the (video and audio) signatures are compact (for example, *N=150* for a 5s video clip duration). For long video clips, the signature could be calculated after a temporal sub-sampling to reduce the frame number and then the size of the signature.
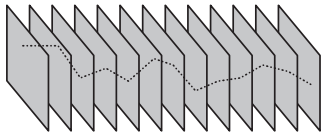

Figure 1: Example of motion for a barycenter on a few frames.

4. AUDIO-VISUAL DETECTION

The audio-visual method for video clip detection consists in using the signatures above. As the time step is synchronized between the video and the audio signatures, we propose to simply normalize both audio and video distance curves, and to add them together. Thus, our A/V fusion is not performed at the feature level, but at the level of the distance curves (late integration). After that, the *minima detection* process can be applied as it is done for audio or video signatures. Figure 2 presents an example of distances between a video clip and the data flow for audio, video and audio-visual signatures. The appearance of the three curves is similar. In this example, we observe three clear minima (valleys) which correspond to the same video clip repeated three times. For the audio-visual signature (Fig. 2c), the sum of the audio and video distance curves reduces the variance of the distance curve and facilitates the detection of the minima (illustrated by dotted lines on Fig. 2).

5. EXPERIMENTS AND RESULTS

5.1. Database and performance metrics

In our experiments we have used the TRECVID2003 database. This database is made up of TV broadcast news coming from ABC and CNN channels. We have selected 7 video clips which appear frequently in the database: six of them announce a particular news sequence, and the last one is an advertisement. The video clip detection performance measure metrics are recall R and precision P and are defined as follows.

-Recall: the percentage of the total relevant documents in a database retrieved by the search.

-Precision: the percentage of relevant documents in relation to the number of documents retrieved.

5.2. Results for audio, video and audio-visual signatures with a real ground truth

We have selected 69 videos (43 from CNN and 26 from ABC, in the MPEG format at 256kbps layer 2) from TRECVID2003 which represents a total duration of about 34 hours and 28 Gb of data. We have also selected video

clips which are present a great number of times in our video base. A real ground truth, corresponding to the time location of all the video clip occurences in the 69 videos was obtained by manual annotation.
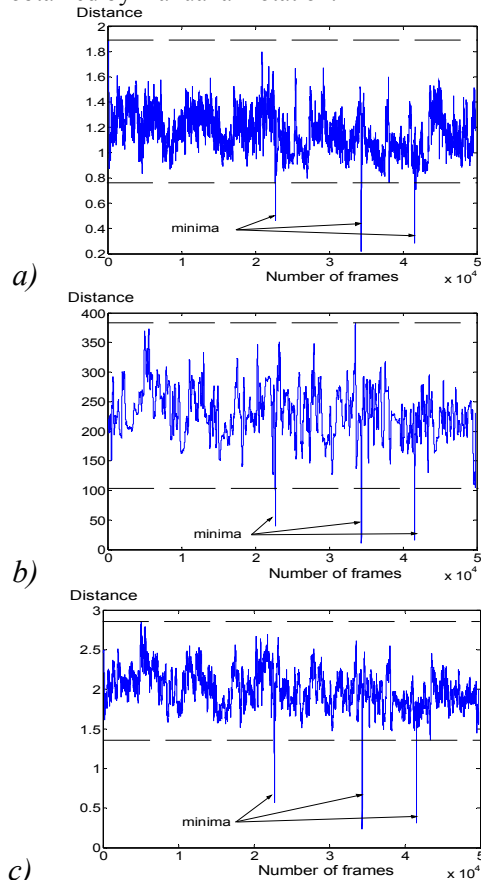

Figure 2: Example of Euclidean distances between a video clip and the data flow for a) audio and b) video signatures. c) Euclidean distance for audio-visual signature obtained by addition of curves a) and b) after normalization.

5.2.1. Case of video clips longer than 5s

Table 1 presents the detection performances for audio, video and audio-visual signatures in the case of video clips between 5 and 10s duration. For all the occurences of video clips 1 to 3, the audio stream is unchanged, whereas the video stream presents some variations (for instance, some images with journalists change for each opening) . In these cases, recall and precision ratios are high with audio signatures and lower for video ones. For all the occurences of video clip 4, it is the contrary: the video stream is unchanged, whereas the audio stream presents some variations (some words change according to the day). In these cases, recall and precision ratios are high with video signatures and lower for audio ones. We can note that, except for video clip 1 (which presents important variability on the video channel and a variable

duration), audio-visual signatures give results close to the best ones obtained with audio or video signature. For video clips 5 and 6, the audio and the video streams are unchanged. Thus, recall and precision ratios are high with audio, video and obviously audio-visual signatures.

Table 1. Case of video clips > 5s: recall R and precision P ratios with audio, video and audio-visual signatures.

	Dura-tion (s)	Occu-rences	Audio R(%) P(%)	Video R(%) P(%)	Audio-visual R(%) P(%)
1: ABC-start	10.2	26	80.7 100	57.6 100	65.3 100
2: CNN-dollar	7.2	38	100 100	65.7 100	100 100
3: CNN-topstories	5.2	40	100 100	90 100	100 100
4: CNN-sport	8.5	40	10 100	97.5 90.6	87.5 100
5: CNN-headline	5.2	120	100 99.1	100 100	100 100
6: advert-merri	8	3	100 100	100 100	100 100
Total		**267**	**84.6** **99.5**	**88.7** **98.3**	**94.7** **100**

5.2.2. Case of shorter video clips (<5s)

Table 2. Case of shorter video clips: recall R and precision P ratios with audio, video and audio-visual signatures.

	Dura-tion (s)	Occu-rences	Audio R(%) P(%)	Video R(%) P(%)	Audio-visual R(%) P(%)
7: sCNN-topstories	3	40	97.5 65	57.5 85.7	85 97.1
8: sCNN-sport	3	40	100 100	92.5 71.1	92.5 86
9: sCNN-headline	2	120	100 47.7	97.5 100	98.3 98.3
10: ABC-closerlook	2	16	100 32	6.25 100	100 69
Total		**216**	**99.5** **53.5**	**81.9** **88**	**94.9** **92.7**

Table 2 presents the detection performances for audio, video and audio-visual signatures in the case of shorter video clips. Clips 7, 8 and 9 have been made by shortening respectively clips 3, 4 and 5. In this case, the audio and video signatures are more sensitive to variations on respectively audio or video streams. Consequently, the quality of results for audio and video signatures generally decreases compared to *table 1*. Nevertheless, we can note an improvement of audio results for video clip 8 (versus video clip 4) because during the short duration clip the audio stream remains unchanged. On the contrary, very good results are obtained with audio-visual signatures. In particular, we observe that the association of audio and video signatures strongly increases precision (clips 7). Finally, the few wrong clips retrieved look like the initial

request. Video clip 10 does not exist in a longer version. In that case, audio and video signatures alone give poor results for detection, while the audio-visual signature is efficient.

5.2.3. Change of compression format

Table 3 presents the detection performances for audio, video and audio-visual signatures in the case of compression of the audio or video requests. It shows that the results remain good in spite of compression.

Table 3. Case of compressed video clips: recall R and precision P ratios with audio (MP3 96kbps), video (DivX 500kbps) and audio-visual (MP3 96kbps and DivX 500kbps) signatures.

	Dura-tion (s)	Occu-rences	Audio R(%) P(%)	Video R(%) P(%)	Audio-visual R(%) P(%)
1: ABC-start	10.2	26	84.6 100	57.6 100	65.3 100
5: CNN-headline	5.2	120	100 98.3	100 100	100 100
10: ABC-closerlook	2	16	100 41	6.25 100	100 84.2
Total		**162**	**97.5** **86.3**	**84** **100**	**94.4** **98.1**

6. CONCLUSION AND FUTURE WORK

In this paper, we have presented the association of audio and video signatures for short video clip detection. Our experiments conducted on a large video database (28Gb / 34h extracted from TRECVID2003) have shown that our audio-visual signature is more robust than the audio-only or video-only signatures. We have also shown that the audio, video and audio-visual signatures are robust against some change of compression format (MP3 and DivX).

7. REFERENCES

[1] E. Allamanche, J. Herre, O. Hellmuth, B. Fröba, T. Kastner, and M. Cremer, "Content-based identification of Audio Material Using MPEG-7 Low Level Description", International Symposium on Music Information Retrieval (ISMIR'2001), Indiana, USA, October 2001.

[2] P. Cano, E. Batlle, T. Kalker, and J. Haitsma. "A review of algorithms for audio fingerprinting", International Workshop on Multimedia Signal Processing (MMSP'02), St Thomas, US Virgin Islands, December 2002.

[3] S-C.S. Cheung and A. Zakhor, "Efficient video similarity measurement with video signature". IEEE Trans. on Circuits and Systems for Video Technology, 13(1), pp 59-74, 2003.

[4] D. Li, N. Dimitrova, M. Li, I. K. Sethi, "Multimedia content processing through cross-modal association", Proceedings of ACM Multimedia 2003, Berkeley, CA, USA,. pp 604-611, November 2003.

[5] J. Pinquier and R. André-Obrecht, "Jingle detection and identification in audio documents", ICASSP'2004, Montréal, Canada, May 2004.

[6] A. Smeaton, W. Kraaij and P. Over, "TRECVID 2003 – An introduction", 12th Text Retrieval Conference, USA, 2003.