From Video Shot Clustering to Sequence Segmentation

Emmanuel Veneau, Rémi Ronfard, Patrick Bouthemy

From Video Shot Clustering to Sequence Segmentation

Emmanuel Veneau, Rémi Ronfard
Institut National de l'Audiovisuel
4, avenue de l'Europe
94366 Bry-sur-Marne cedex, France
{eveneau,rronfard}@ina.fr

Patrick Bouthemy
IRISA/INRIA
Campus Universitaire de Beaulieu
35042 Rennes cedex, France
bouthemy@irisa.fr

Abstract

Segmenting video documents into sequences from elementary shots to supply an appropriate higher level description of the video is a challenging task. This paper presents a two-stage method. First, we build a binary agglomerative hierarchical time-constrained shot clustering. Second, based on the cophenetic criterion, *a* breaking distance *between shots is computed to detect sequence changes. Various options are implemented and compared. Real experiments have proved that the proposed criterion can be efficiently used to achieve appropriate segmentation into sequences.*

1 Introduction

Browsing and querying data in video documents requires to first extract and organize information from the audio and video tracks. The first step in building a structured description is to segment the video document into elementary shots which are usually defined as the smallest continuous units of a video document. Numerous methods for shot segmentation have been proposed (e.g., see [3]). Nevertheless, shots are often not the relevant level to describe pertinent events, and are too numerous to enable efficient indexing or browsing.

The grouping of shots into higher-level segments has been investigated through various methods which can be gathered into three main families. The first one is based on the principle of the Scene Transition Graph (STG) [9], which can be formulated in a continuous way [7], or according to alternate versions [4]. Methods of the second family [1, 2] use explicit models of video documents or rules related to editing techniques and film theory. In the third family [5, 8], emphasis is put on the joint use of features extracted from audio, video and textual information. These methods achieve shot grouping more or less through a combination of the segmentations performed for each track.

We present a method based on a so-called *cophenetic criterion* which belongs to the first family. The sequel is organized as follows. Section 2 describes our method involving an agglomerative binary hierarchy and the use of the cophenetic matrix. Section 3 specifies the various options we have implemented with respect to extracted features, distance between features, hierarchy updating, and temporal constraints. Experimental results are reported in Section 4, and Section 5 contains concluding remarks.

2 Binary hierarchy for describing shot similarity

We assume that a segmentation of the video into shots is available, where each shot is represented by one or more extracted keyframes. The information representing a shot (except its duration) is given by the (average) signature computed from the corresponding keyframes. We build a spatio-temporal evaluation of shot similarity through a binary agglomerative hierarchical time-constrained clustering.

2.1 Binary agglomerative hierarchical time-constrained clustering

To build a hierarchy following usual methods [10], we need to define a similarity measure s between shots, and a distance between shot clusters, called index of dissimilarity δ. The temporal constraint, as defined in [9], involves a temporal distance d_t. We introduce a temporal weighting function W accounting for a general model for the temporal constraint. The formal definitions of all these functions will be given in Section 3. The time-constrained distance $\tilde{d}$ between shots is defined (assuming that similarity is normalized between 0 and 100) by :

$$\tilde{d}(i,j) = \begin{cases} 100 - s(i,j) \times W(i,j) & \text{if } d_t(i,j) \leq \Delta T \\ \infty & \text{otherwise} \end{cases} \quad (1)$$

0-7695-0750-6/00 $10.00 © 2000 IEEE

where i and j designate two shots and ΔT is the maximal temporal interval for considering any interaction between shots.

At the beginning of the process, each shot forms a cluster, and the time-constrained dissimilarity index $\tilde{\delta}$ between clusters is then the time-constrained distance $\tilde{d}$ between shots. A symmetric time-constrained $N \times N$ proximity matrix $\tilde{\mathcal{D}} = [\tilde{d}(i,j)]$ is considered [6], using $\tilde{\delta}$ to evaluate the dissimilarity between clusters. The hierarchy is built by merging the two closest clusters at each step. The matrix $\tilde{\mathcal{D}}$ is updated according to the index of dissimilarity $\tilde{\delta}$ to take into account each newly created cluster. This is iterated until the proximity matrix contains only infinite values. The resulting binary time-constrained hierarchy supplies a description of the spatio-temporal proximity of the extracted shots.

2.2 Cophenetic dissimilarity criterion

In [6], another proximity matrix $\mathcal{D}_c$, called *cophenetic* matrix, is proposed to capture the structure of the hierarchy. We will use the time-constrained version $\tilde{\mathcal{D}}_c$ of this matrix to define a criterion for the segmentation of the video into sequences. The *cophenetic* matrix is expressed as $\tilde{\mathcal{D}}_c = [\tilde{d}_c(i,j)]$, where $\tilde{d}_c$ is the so-called *clustering distance* defined by :

$$\tilde{d}_c(i,j) = \max_{p \neq q/(i,j) \in C_p \times C_q} \{\tilde{\delta}(C_p, C_q)\}$$

where $\tilde{\delta}$ is the index of dissimilarity constructed from $\tilde{d}$, and C_p and C_q are two clusters. Assuming that the shot indices follow a temporal order, the *cophenetic* matrix leads to the definition of our criterion for sequence segmentation, called *breaking distance*, calculated between two consecutive shots as : $\tilde{d}_b(i, i+1) = \min_{k \leq i < l} \{\tilde{\mathcal{D}}_c(k,l)\}$.

2.3 Segmentation using the breaking distance

If the breaking distance $\tilde{d}_b$ between consecutive shots exceeds a given threshold τ_c, then a sequence boundary is inserted between these two shots. An example is presented on Fig 1 where two different thresholds to perform segmentation into sequences $\tau_1 = 20$ and $\tau_2 = 45$ are considered. Fig. 2 displays results corresponding to thersholds τ_1 and τ_2.

2.4 Comparison with the STG method

We have formally proved that our method delivers the same segmentation into sequences as the STG method described in [9]. Considering that STG method considers in a binary way inter-shot spacing and implies non-obvious setting of parameters [7], the advantage of our formulation is to smooth the effects of time, in the time-constrained distance, using continuous temporal weighting functions, and to consider a threshold parameter related to sequence segmentation and not to shot clustering. As a consequence, our approach allows one to visualize what the segmentation results are according to the selected threshold value which can then be appropriately tuned by the user. There is no need to rebuild the STG whenever the threshold is changed.

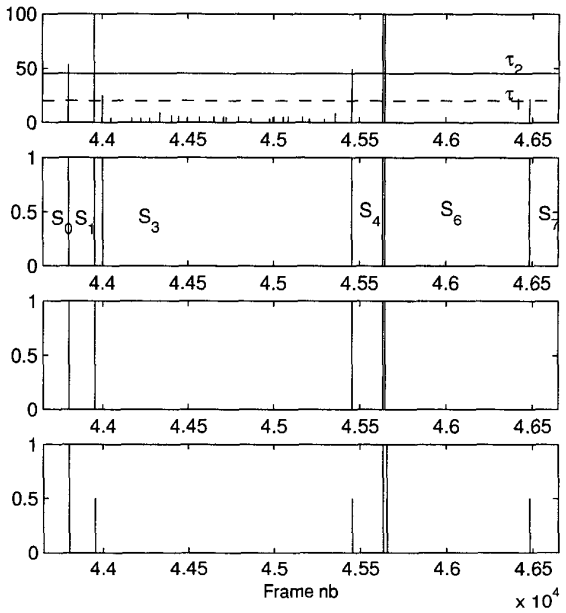

Figure 1. Thresholding the breaking distance values on excerpt 1 of *Avengers* movie (upper row), detected sequence boundaries for τ_1 (upper middle row) and τ_2 (lower middle row), and manual segmentation (lower row)

3 Description of implemented options

3.1 Signatures for shots

We have considered in pratrice three kinds of signatures : shot duration, color and region-based color histograms. Color and region-based color histograms are defined in the (Y, C_b, C_r) space with respectively 16, 4, and 4 levels, and 12 image blocks are considered for region-based histograms. The shot duration gives a relevant information on the rhythm of the action and on the editing work.

3.2 Distances between signatures

Various distances between signatures have been tested. Comparison between histograms can be achieved using his-

togram intersection, euclidian distance, χ_2-distance. The distance chosen between shot durations is the Manhattan distance.

3.3 Updating of the agglomerative binary hierarchy

In order to update the classification hierarchy, two algorithms are available [10] :

- the *Complete Link* method. The index of dissimilarity between clusters is defined by :

$$\tilde{\delta}(C_p, C_q) = \max_{(i,j)\in C_p \times C_q} \{\tilde{d}(i,j)\}$$

- the *Ward's* method. The index of dissimilarity between clusters is given by :

$$\tilde{\delta}(C_p, C_q) = \frac{n_{C_p}.n_{C_q}}{n_{C_p} + n_{C_q}} \tilde{d}(G_{C_p}, G_{C_q})$$

where G_{C_i} is the gravity centre of cluster C_i, n_{C_i} represents either $Cardinal(C_i)$ or $Duration(C_i)$.

In both cases, the Lance and William formula, given by $\tilde{\delta}(A \cup B, C) = a_1\tilde{\delta}(A,C) + a_2\tilde{\delta}(B,C) + a_3\tilde{\delta}(A,B) + a_4|\tilde{\delta}(A,C) - \tilde{\delta}(B,C)|$, is used to update the proximity matrix. We have $a_1 = a_2 = a_4 = \frac{1}{2}$, $a_3 = 0$ for the *Complete Link* method, and $a_1 = \frac{n_A+n_C}{n_{A\cup B}+n_C}$, $a_2 = \frac{n_B+n_C}{n_{A\cup B}+n_C}$, $a_3 = 0$, $a_4 = \frac{n_C}{n_{A\cup B}+n_C}$ for the *Ward's* method.

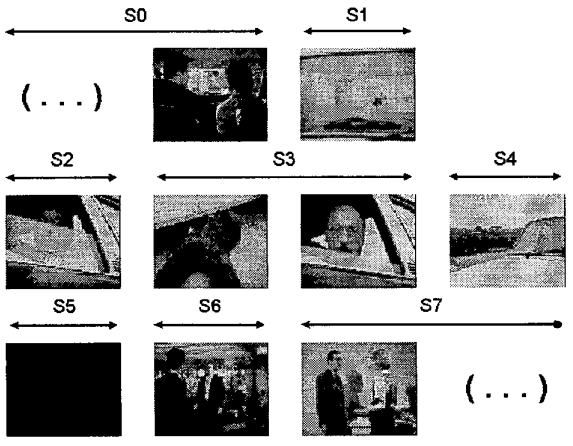

Figure 2. Obtained sequence segmentation on excerpt 1 of *Avengers* movie for threshold τ_1. S_3 is an angle / reverse angle sequence. S_5 is a fade out / fade in effect.

3.4 Temporal weighting function

The temporal weighting function is used to constrain the distance and the index of dissimilarity as introduced in equation 1. In [9], only one type of temporal weighting function was proposed, i.e. rectangular function which is not smooth. We have tested three smooth functions : linear, parabolic, and sinusoidal.

4 Experimental results

We have evaluated our method on a three hour video corpus. We report here results on four excerpts of two minutes. Three excerpts are taken from *Avengers* movies to evaluate the segmentation into sequences in different contexts. The first one comprises an angle / reverse angle editing effect and a transition with a dissolve effect. The second one includes a set change, and the third one involves color and rhythm changes. Obtained segmentations can be compared with a hand segmentation acting as ground truth. In plots displayed in Figures 1, 3 and 4, main sequence changes are represented by a value of 1 and secondary changes by a value of 0.5. The last excerpt is extracted from a news program to test the relevance of the built hierarchy.

Among the implemented options, three sets of descriptors and functions are selected : (O_1) color histograms intersection, rectangular temporal weighting function, and Complete Link method, (O_2) color histograms intersection, parabolic temporal weighting function, and Ward's method based on cluster duration, (O_3) Manhattan distance on shots duration, parabolic weighting function, and Ward's method based on cluster duration.

Results obtained on the news program excerpt show that the clustering distance $\tilde{d}_c$ provides a correct description of the similarity between shots at different levels, even if the information distribution is not homogeneous in the various levels of the hierarchy. An adaptive thresholding applied to breaking distance values would be nevertheless necessary to avoid heterogeneous results. Tests have shown that the best video segmentation into sequences is found using option set O_2.

In the processed excerpts, most of the sequence changes were correctly detected, when the proper options were selected. On Fig.1, we can point out that, using τ_1 and option O_1, all changes are detected with only one false alarm, the angle / reverse angle effect is recognized. Selecting the threshold value is nevertheless a rather critical issue. On excerpt 2, with a relevant threshold, we extract all the correct boundaries with option O_1, with only one false alarm (Fig. 3). Using option O_2 false alarms and missed detections increase on excerpt 2. The color and rhythm changes in excerpt 3 (Fig. 4) have been better detected using option

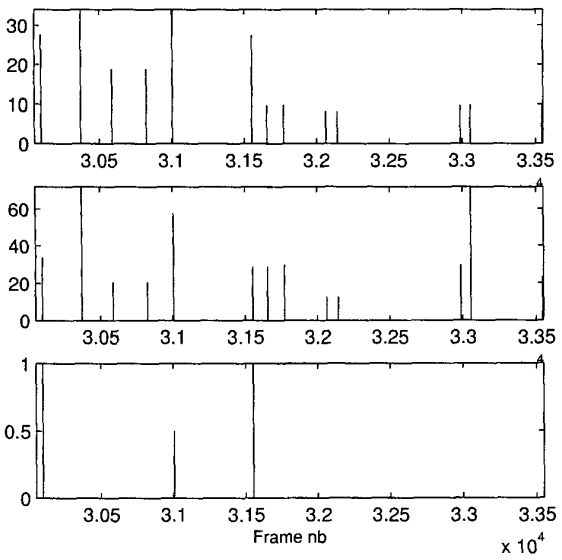

Figure 3. Breaking distance values on excerpt 2 of *Avengers* movie using option O_1 (upper row), option O_3 (middle row), and manual segmentation (lower row)

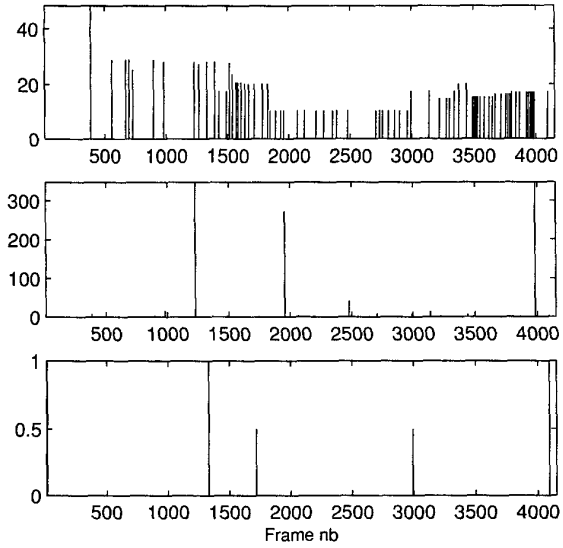

Figure 4. Breaking distance values on excerpt 3 of *Avengers* movie using option O_1 (upper row), option O_3 (middle row), and manual segmentation (lower row)

O_3, rather than O_1. Consequently, how to automatically select the proper option remains an open issue.

5 Conclusion

The method described in this paper, based on the cophenetic matrix, enables to accurately and efficiently segment video documents into sequences by building a binary agglomerative time-constrained hierarchy. We have implemented several versions. Selecting the most appropriate one improved results and gave a better description of the similarity of the shots through the hierarchy. Experiments on a larger base will be conducted in future work for selecting the best parameter set and evaluating alternative thresholding stategies.

References

[1] P. Aigrain, P. Joly, and V. Longueville. Medium knowledge-based macro-segmentation of video into sequences. In M. T. Maybury, editor, *Intelligent Multimedia Information Retrieval*, pages 159–173. AAAI/MIT Press, 1997.

[2] J. Carrive, F. Pachet, and R. Ronfard. Using description logics for indexing audiovisual documents. In ITC-IRST, editor, *Int. Workshop on Description Logics (DL'98)*, pages 116–120, Trento, 1998.

[3] A. Dailianas, R. B. Allen, and P. England. Comparison of automatic video segmentation algorithms. In *SPIE Photonics West*, volume 2615, pages 2–16, Philadelphia, 1995.

[4] A. Hanjalic, R. L. Lagendijk, and J. Biemond. Automatically segmenting movies into logical story units. In *Third Int. Conf. on Visual Information Systems (VISUAL'99)*, volume LNCS 1614, pages 229–236, Amsterdam, 1999.

[5] A. G. Hauptmann and M. A. Smith. Text, speech, and vision for video segmentation : The informedia project. In *AAAI Fall Symposium, Computational Models for Integrating Language and Vision*, Boston, 1995.

[6] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[7] J. R. Kender and B.-L. Yeo. Video scene segmentation via continuous video coherence. Technical report, IBM Research Division, 1997.

[8] R. Lienhart, S. Pfeiffer, and W. Effelsberg. Scene determination based on video and audio features. Technical report, University of Mannheim, November 1998.

[9] M. Yeung, B.-L. Yeo, and B. Liu. Extracting story units from long programs for video browsing and navigation. In *Proc. of IEEE Int. Conf. on Multimedia Computing and Systems*, Tokyo, 1996.

[10] J. Zupan. *Clustering of Large Data Sets*. Chemometrics Research Studies Series. John Wiley & Sons Ltd., 1982.

Acknowledgements Images from the *Avenger* movie, part of the AIM corpus, were reproduced thanks to INA, Department Innovation.