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

Automatically building high-level segments to structure information in video documents is a challenging task. This paper presents a method based on the cophenetic criterion, *a distance between clustered shots which detects breaks between sequences. It describes and compares various implemented options. Experiments have proved that the proposed criterion can be used for achieving segmentation.*

1 Introduction

Browsing and querying data in video documents requires to structure extracted information from the audio and video flows. The first step in building a structured description of data is to segment the video document. The elementary segment is the shot, which is usually defined as the smallest continuous unit of a video document. Numerous methods for shot segmentation have been proposed (e.g. see [4]). 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 four main families. The first one is based on the principle of the Scene Transition Graph (STG) [11], which can be exploited in a continuous way [9], or according to alternate versions [5]. The methods of the second family [1, 3] use explicit models of video documents or rules related to editing techniques and film theory. In the third family [6, 10], 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 synthesis of the segmentation performed for each media. The fourth family of algorithms relies on statistical techniques as Hidden Markov Models (HMM) and other Bayesian tools [2, 7].

In this paper, we present a method based on a *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 contained in a shot (except its duration) reduces to the (average) signature computed from the corresponding keyframes. We build a spatio-temporal description of shot similarity through a binary agglomerative hierarchical time-constrained clustering.

2.1 Binary agglomerative hierarchical time-constrained clustering

To build a hierarchy following standard methods [12], we require a similarity measure s between shots, and a distance between shot clusters, called index of dissimilarity, δ. The temporal constraint, as defined in [11], involves a temporal distance d_t. We introduce a temporal weighting function W in order to have a general model for the temporal constraint. The formal definitions of these functions will be given in section 3. The time-constrained distance between shots $\tilde{d}$ is defined (assuming that similarity is normalized between 0 and 100) by :

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

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 the time-constrained distance $\tilde{d}$ between shots. A symmetric time-constrained $N \times N$ proximity matrix $\tilde{\mathcal{D}} = [\tilde{d}(i,j)]$ can be defined [8], using $\tilde{\delta}$, as a representation of 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 the newly created cluster. This step is iterated until the proximity matrix contains only infinite values.

The resulting binary time-constrained hierarchy provides a description of the spatio-temporal proximity of shots.

2.2 Cophenetic dissimilarity criterion

In [8], 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 of this matrix $\tilde{\mathcal{D}}_c$ to defined a criterion for sequence segmentation. The *cophenetic* matrix can be expressed as follows : $\tilde{\mathcal{D}}_c = [\tilde{d}_c(i,j)]$, where $\tilde{d}_c$ is a so-called *clustering distance* defined as :

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

where $\tilde{\delta}$ is the index of dissimilarity constructed on $\tilde{d}$, and C_p and C_q are two clusters. Assuming that the shot indices follow a temporal order, the *cophenetic* matrix leads us to the definition of our criterion for 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 $\tau_1 = 20$ and $\tau_2 = 45$ are evaluated to perform two different segmentations in sequences (Fig. 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 [11]. The advantage of our formulation is to allow 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.

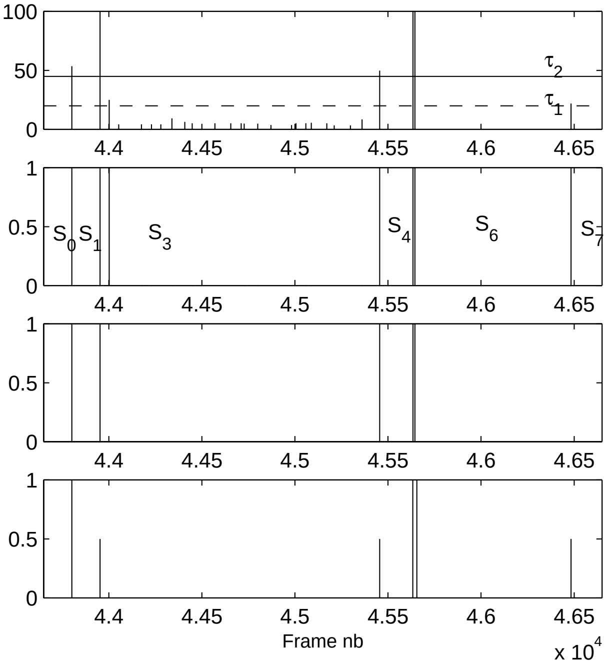

Figure 1. Thresholding the breaking distance values on excerpt 1 of *Avengers* movie (upper row), detected sequence boundaries for τ_1 (upper middle row) and τ_2 (lower middle row), and manual segmentation (lower row)

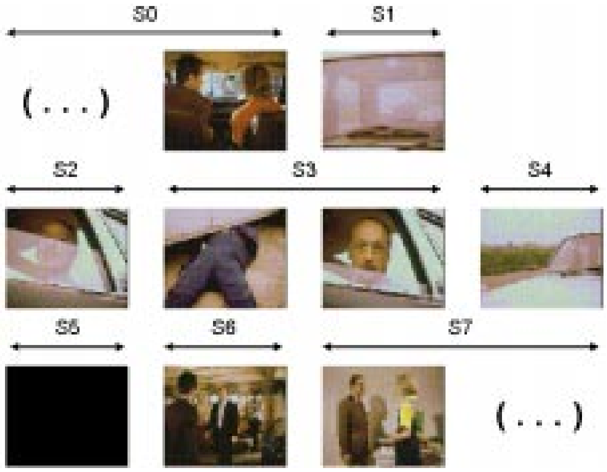

Figure 2. Obtained sequence segmentation on excerpt 1 of *Avengers* movie for threshold τ_1. S_3 is an angle / reverse angle sequence. S_5 is a fade out / fade in effect.

3 Description of implemented options

3.1 Signatures for shots

Three kinds of signatures are considered in practice : shot duration, color or region-color histogram. Color and region-color histograms are defined in the (Y, U, V) space with respectively 16, 4, and 4 levels, and 12 image blocks are considered for region-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 histogram 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 [12] :

- 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} may represent either $Card(C_i)$ 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.

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 [11], 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 are evaluating our method on a three hour video corpus. For this communication, four excerpts of two minutes were selected. 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 content change 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, which is weighted as follow : 1 for main changes, 0.5 for secondary changes. The last excerpt is extracted from a news program to test the relevance of the built hierarchy.

Among the implemented options, three sets were selected for their relevant results : (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 clusters duration, (O_3) Manhattan distance on shots duration, parabolic weighting function, and Ward's method based on clusters duration.

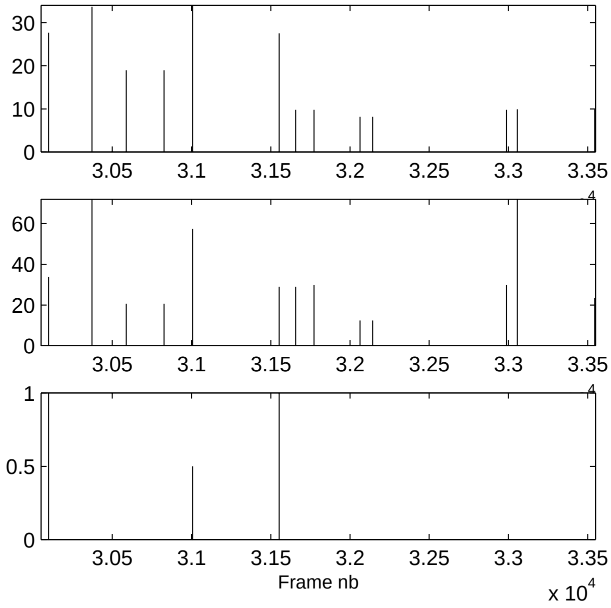

Figure 3. Breaking distance values on excerpt 2 of *Avengers* movie using options set O_1 (upper row), options set O_3 (middle row), and manual segmentation (lower row)

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 description is found using the options set O_2.

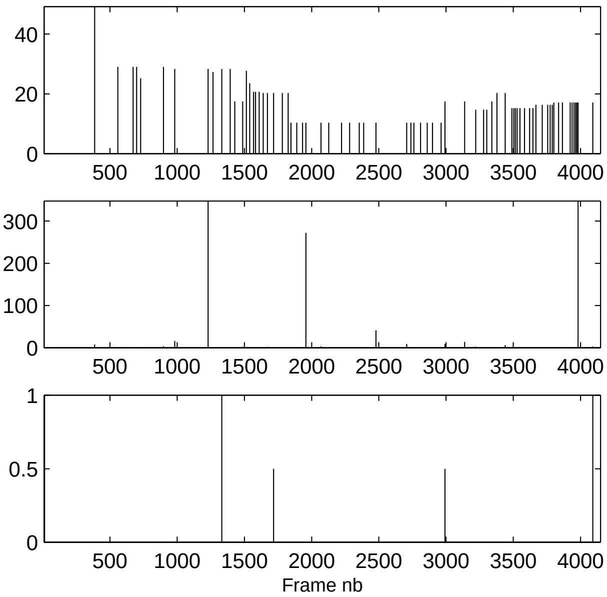

Figure 4. Breaking distance values on excerpt 3 of *Avengers* movie using options set O_1 (upper row), options set O_3 (middle row), and manual segmentation (lower row)

In the processed excerpts, most of the sequence changes were correctly detected, when the proper options were selected. On Fig.1, using τ_1 and the options set O_1, one can see that all changes are detected with only one false alarm, the angle / reverse angle effect is recognized, but that selecting the threshold value is a rather critical issue. On excerpt 2, with a relevant threshold, we can predict all the boundaries with options set O_1, with only one false alarm (Fig. 3). Using options set O_2, relevant for the hierarchy building, false alarms and miss rates increase on excerpt 2. The color and rhythm change in excerpt 3 (Fig. 4) have been better detected using options set O_3, rather than O_1. Consequently, how to automatically select the proper options remains an open issue.

5 Conclusion

The method described in this paper, based on the cophenetic matrix, allows us to determine and visualize the sequence boundaries corresponding to all levels in the binary agglomerative time-constrained hierarchy. We implemented several options. Selecting the most appropriate ones improved our results and gave a better description of the similarity of the shots through the hierarchy. Experiments on a larger scale will be undertaken in future work for selecting the best parameter sets 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. S. Boreczky and L. D. Wilcox. A hidden Markov model framework for video segmentation using audio and image features. In *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'97)*, Seattle, 1997.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] G. Iyengar and A. Lippman. Models for automatic classification of video sequences. In *Photonics West '98, (Storage and Retrieval VI)*, volume SPIE 3312, pages 216–227, San Jose, 1998.

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

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

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

[11] 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.

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

Acknoledgement The images from the *Avenger* movie, part of the AIM corpus, was reproduced thanks to INA, Department Innovation.