Content-Based Video Segmentation using Statistical Motion Models

N. Peyrard and P. Bouthemy
IRISA/INRIA
Campus universitaire de Beaulieu 35042 Rennes Cedex, France
e-mail: {npeyrard,bouthemy}@irisa.fr

Abstract

We present in this paper an original approach for content-based video segmentation using motion information. The method is generic and does not require any knowledge about the type of the processed video. Its relies on the analysis of the temporal evolution of the dynamic content of the video. The motion content is characterised by a probabilistic Gibbsian modelling of the distribution of local motion-related measurements. The designed statistical framework provides a well formalised similarity measure according to motion activity that we exploit to derive criteria for segmentation decision. Then, the considered merging criteria are sequentially applied between every two successive temporal units of the video to progressively form homogeneous segments in term of motion content. Experiments on real video documents demonstrate the ability of the proposed approach to provide a concise and meaningful overview of a video.

Keywords: video segmentation, probabilistic motion characterisation, Gibbs models, merging decision criteria.

1 Introduction

Due to the explosion in the amount of digital video documents produced, the actual challenge for computer vision is the development of efficient tools for their exploitation. It implies tasks such as video indexing, browsing or search. One of the main operations at the basis of these tasks is video segmentation. Replacing a long video by a small number of representative segments provides a synthetic description of the document which can be exploited for domestic (preview and selection of recorded TV programs) and professional domains (audio-visual archives consultation). One level of video segmentation is the identification of the elementary shots in the video document [2, 3]. However, it only reveals the technical acquisition process of the video (video shooting and editing) and is not adapted for the objectives in sight here: segmenting the video into meaningful segments. Since a shot is not explicitly related to the video content, it can contain more than one event or sometimes a unique event can be spread on several shots. A solution to rectify over-segmentation can be to merge successive elementary shots of similar content [9], leading to more meaningful video segments. However, a non detection of a content change cannot be recovered. Our goal is to design an automatic, generic and simple

BMVC 2002 doi:10.5244/C.16.51

method for content-based video segmentation into coherent temporal segments which can offer a compact and meaningful representation of the video in terms of temporal events. In this order, we do not consider any preliminary segmentation of the video into elementary shots. Conversely, we adopt an approach based on motion content analysis for direct temporal video segmentation, since a variation in motion activity in an image sequence is a strong indicator of an event change. One way to characterise the dynamic content of temporal units of the processed video would be to consider parametric motion models (e.g. 2D affine of quadratic motion models). However, the dynamic situations which can be described by such models are quite limited. The study described in [8] for human action boundaries detection in a video is based on moving object segmentation and temporal discontinuities detection in the optical flow, by analysing the evolution of the most significant coefficients of the Singular Value Decomposition of the set of successive flow fields. In [10], in order to cluster temporal dynamic events, the latters are captured by the local spatial and temporal intensity gradients at different temporal scales. A distance between events is then built, based on the comparison of the empirical histograms of these features. As [10], we propose to exploit low-level motion information straightforwardly extracted from the images intensities. We consider a non parametrical approach and adopt the statistical motion models introduced in [5], specified from temporal cooccurrences of local motion measurements related to the normal velocity. These motion models handle a wide range of dynamic contents and provide a general characterisation of these contents in terms of motion activity. When specifying this way what the characteristics shared by coherent temporal units are, no a priori knowledge on the processed video is required. Besides, these motion models are expressed in a a well-funded probabilistic framework which allow us to properly design a motion-based similarity measure between video units. We then determine homogeneous video segments in a sequential way, by analysing the temporal variations of the motion information. To this end, we investigate two merging decision criteria, relying on a distance between the involved statistical motion models, to sequentially decide whether the successive temporal video units should be merged into an homogeneous segment or not. We compare their behaviour and performances on real video documents.

The paper is organized as follows. In section 2, we outline the statistical modelling of motion activity within image sequences that we exploit. Section 3 is dedicated to the definition of our merging decision criteria. The proposed sequential content-based segmentation method is presented in Section 4. Experimental results are reported and commented in Section 5. Concluding remarks (Section 6) end this paper.

2 Statistical motion modelling

The proposed method for content-based video segmentation relies on a probabilistic modelling of the dynamic content within temporal units of the processed video. In order to handle a wide range of dynamic situations (outdoor and indoor scenes, character scenes, sport scenes, ...), we benefit from a general notion of motion activity and we exploit the statistical motion models recently introduced in [5]. This framework for motion activity modelling has already been successfully applied to motion recognition [4] and motion-based video retrieval [5]. In this Section, we briefly outline its main characteristics. The motion activity models are identified from the analysis of the distribution of local motion-

related measurements. More specifically, for a given pixel p and at a given time k, the normal flow is straightforwardly computed from the spatio-temporal intensity gradient in the images. Then a continuous local motion measure is computed as a weighted mean, over a small spatial window, of the normal flow magnitude, in order to obtain a more reliable motion information. These measures are then quantized on a set Λ of discrete values. A causal Gibbs probabilistic distribution can represent the temporal cooccurrences of the quantized local motion measurements $\{y(k,p), k = 1 \dots K, p = 1 \dots \mathcal{R}\}$, where $\mathcal{R}$ is the spatial image support and K is the length of the sequence. More precisely, given an images sequence, we compute the associated sequence $y = \{y(k), k = 1 \dots K\}$ of local motion quantities maps (one motion map $y(k) = \{y(k,p), p = 1 \dots \mathcal{R}\}$ is computed from two successive images). The temporal cooccurrences distribution $\Gamma(y)$ of the sequence y is a matrix $\{\Gamma(\nu, \nu'|y)\}_{(\nu,\nu')\in\Lambda^2}$ defined by

$$\Gamma(\nu, \nu'|y) = \sum_{k=1}^{K-1} \sum_{p\in\mathcal{R}} \delta(\nu, y(k,p)) \cdot \delta(\nu', y_{k+1}(p)), \tag{1}$$

where $\delta(i,j)$ is the Kronecker symbol (equal to 1 if $i = j$ and to zero otherwise). Given a temporal Gibbsian model $\mathcal{M}$ specified by its potentials $\Psi_{\mathcal{M}} = \{\Psi_{\mathcal{M}}(\nu, \nu')\}_{(\nu,\nu')\in\Lambda^2}$, the likelihood of the sequence y under the model $\mathcal{M}$ is simply evaluated from the dot product between the potentials $\Psi_{\mathcal{M}}$ and the matrix of the temporal cooccurrences $\Gamma(y)$ [5]:

$$P_{\mathcal{M}}(y) = \frac{1}{Z} \exp\left[\sum_{(\nu,\nu')\in\Lambda^2} \Psi_{\mathcal{M}}(\nu, \nu') \cdot \Gamma(\nu, \nu'|y)\right], \tag{2}$$

with the normalisation constraint (to ensure potentials unicity) $\sum_{\nu\in\Lambda} \exp\left[\Psi_{\mathcal{M}}(\nu, \nu')\right] = 1$.

The appealing characteristic of these models is that, due to their causal aspect, the normalisation constant Z is explicitly known and tractable. Furthermore, it is independant of the model $\mathcal{M}$. Thus the probability (2) is exactly determined and available for any sequence y and model $\mathcal{M}$. The model estimation is achieved according to the maximum likelihood (ML) principle. It is easy to see that the temporal model (2) is actually equivalent to a product of $\mathcal{R}$ independent and identically distributed Markov chains defined by the transition matrix $T = \{exp(\Psi_{\mathcal{M}}(\nu, \nu'))\}_{(\nu,\nu')\in\Lambda^2}$. Thus, the ML estimate is determined by the empirical estimate of T, and given an observed sequence y, the estimated potentials $\Psi_{\widehat{\mathcal{M}}}$ are deduced from the cooccurrences distribution $\Gamma(y)$ as follows:

$$\Psi_{\widehat{\mathcal{M}}}(\nu, \nu') = \ln\left(\frac{\Gamma(\nu, \nu'|y)}{\sum_{\nu''\in\Lambda} \Gamma(\nu'', \nu'|y)}\right). \tag{3}$$

Therefore, the use of these statistical motion activity models appears simple and efficient. The computation of the temporal cooccurrences $\Gamma(y)$ can be realised in a parallel scheme. Once the temporal cooccurrences distribution is available, the model estimation is straightforward and the evaluation of the likelihood (2) requires only the computation of dot products between the model potentials and the cooccurrence coefficients, which are tasks of low computational time.

3 Decision criteria for merging temporal video units

In this Section, we consider the issue of stating whether or not two temporal video units can be considered as homogeneous in terms of motion activity, in order to decide if they could reasonably be merged in an unique representative segment. To achieve this goal we exploit the statistical framework associated to the designed motion models. We use the Kullback-Leibler divergence [1] to build the dynamic-content similarity measure, and we propose criteria for merging decision based on this similarity measure.

Let us consider two sequences of motion quantities maps y and z and the corresponding estimated motion models $\mathcal{M}_y$ and $\mathcal{M}_z$. The Kullback-Leibler divergence between models $\mathcal{M}_y$ and $\mathcal{M}_z$ is defined as:

$$KL(\mathcal{M}_z||\mathcal{M}_y) = \int log(P_{\mathcal{M}_y}(u)/P_{\mathcal{M}_z}(u))dP_{\mathcal{M}_y}(u). \quad (4)$$

In practice, this quantity is not available and an approximation is necessary. Each transition from $y(k-1,p)$ to $y(k,p)$ beeing a realisation of the Markov chain related to $P_{\mathcal{M}_y}$, we use the following Monte-Carlo approximation of expression (4) (see [5] for more details):

$$KL(\mathcal{M}_z||\mathcal{M}_y) \quad \approx \quad log\left(P_{\mathcal{M}_y}(y)/P_{\mathcal{M}_z}(y)\right) \approx (\Psi_{\mathcal{M}_y} - \Psi_{\mathcal{M}_z}) \bullet \Gamma(y), \quad (5)$$

since the normalisation constant Z is model independent. As a result, the Kullback-Leibler divergence has an easy interpretation: it evaluates the information loss when substituting model $\mathcal{M}_z$ for model $\mathcal{M}_y$ to describe y.

Given two successive motion map sequences, y_t and y_{t+1}, a first criterion for the merging decision is thus the Kullback-Leibler divergence $KL(\mathcal{M}_{y_t}||\mathcal{M}_{y_{t+1}})$ between models $\mathcal{M}_{y_t}$ and $\mathcal{M}_{y_{t+1}}$, that we will denote $KL(y_t, y_{t+1})$ in the following. As mentioned above, this criterion expresses how well the model $\mathcal{M}_{y_t}$ estimated on the current sequence can fit the next coming sequence. However, when considering the issue of merging the two sequences y_t and y_{t+1} based on their respective dynamic contents, a symmetrical criterion would seem more natural. This could be the symmetrical version of $KL(y_t, y_{t+1})$, however, it is important to notice that actually three statistical motion models are involved in the merging process: $\mathcal{M}_{y_t}$, $\mathcal{M}_{y_{t+1}}$ and the model $\mathcal{M}_{y_t \cup y_{t+1}}$ that would describe the new sequence $y_t \cup y_{t+1}$ resulting from the merging of y_t and y_{t+1}. The decision of merging should be taken if the model $\mathcal{M}_{y_t \cup y_{t+1}}$ can fit well both y_t and y_{t+1}. We thus propose the following criterion, combining the three models and likely to give a more meaningful representation of the considered merging problem:

$$C_m(y_t, y_{t+1}) = \frac{1}{2}[KL(\mathcal{M}_{y_t \cup y_{t+1}}||\mathcal{M}_{y_t}) + KL(\mathcal{M}_{y_t \cup y_{t+1}}||\mathcal{M}_{y_{t+1}})]. \quad (6)$$

This criterion is now symmetrical. It evaluates the information loss when deciding to merge y_t and y_{t+1} and to represent them by a single motion model. In term of computational amount, the use of $C_m(y_t, y_{t+1})$ when comparing each two successive temporal units of a video involves the computation of $\mathcal{M}_{y_t \cup y_{t+1}}$ at each comparison, while the use of $KL(y_t, y_{t+1})$ necessitates this evaluation only after a merging decision, for model updating. However, the model $\mathcal{M}_{y_t \cup y_{t+1}}$ can be easily (and fastly) estimated by adding the (purely) temporal cooccurrences matrix $\Gamma(y_t)$ and $\Gamma(y_{t+1})$ to obtain $\Gamma(y_t \cup y_{t+1})$, from which the model potentials are derived as in expression (3). In practice, the cost

corresponding to the two criteria are quite similar. In the next section, we describe more precisely our method for content-based video segmentation.

4 Temporal video segmentation based on motion information

Our method for motion-based video segmentation relies on the analysis of the dynamic content of successive temporal units of the considered video document, through the statistical motion models associated to each of these units. The method provides a temporal video segmentation into homogeneous segments according to motion content as well as a characterisation of these segments by an associated statistical motion model. Given a video, we first need to define a sequence of temporal units of the video, $\{u_t\}_{t\in[1,T]}$, which is the input of our temporal segmentation method. A temporal unit can be defined by its first and its last image. Temporal units of two successive motion maps, computed from three successive images, is the minimal length L we can consider, since the statistical motion activity models are specified from temporal cooccurrences of the motion-related measurements. In the following, we will denote y_t the sequence of motion maps and $\mathcal{M}_t$ the motion model estimated from the temporal unit u_t.

In a previous work [6], a hierarchical batch approach had been investigated, using the symmetrical version of $KL(y_t, y_{t+1})$, but we prefer here a sequential approach, which is much less time consuming and whose implementation is less demanding in terms of memory space. Each step of our algorithm consists in deciding if the homogeneous segment currently built should encompass the next temporal unit. Let us denote h_n and $\mathcal{M}_n^{hom}$ respectively the sequence of motion quantities maps and the motion model associated to this homogeneous segment. At each iteration the sequence of motion quantities maps y_t and the motion model $\mathcal{M}_t$ associated to the next temporal unit are computed. Then, the similarity between the two involved sequences of motion maps, h_n and y_t, is evaluated through the computation of the considered merging criterion C (note that if this criterion is $C_m(h_n, y_t)$ given by expression (6), it is also necessary to estimate the motion model $\mathcal{M}_{h_n \cup y_t}$ associated to the merged segment $h_n \cup y_t$). If the value of C is lower than a given threshold, then y_t is incorporated to the current homogeneous segment and the motion model $\mathcal{M}_n^{hom}$ is updated. Otherwise, the current homogeneous segment h_n is ended at y_{t-1} and a new homogeneous segment h_{n+1} is initialised, corresponding at that point to the single unit y_t. Then, the process is iterated with unit y_{t+1} (see Figure 1). The algorithm supplies as output a sequence of homogeneous temporal video segments $\{h_n\}_{n\in[1,N]}$ and the sequence of the associated motion models $\{\mathcal{M}_n^{hom}\}_{n\in[1,N]}$. Each model $\mathcal{M}_n^{hom}$ corresponds to the maximum likelihood estimator for the sequence h_n.

The method performances are studied in the following section.

5 Experiment results

We have carried out experiments on two different video documents. The *Athletics* video is part of a TV sport program corresponding to an athletics meeting and the *Avengers* video is a film strip of the TV serie “Avengers”. The examples processed contain respectively 1416 and 3496 frames. Representative images of the two video sequences are displayed

Initialisation:	compute y_1 and $\mathcal{M}_1$ $h_1 \leftarrow y_1$ $\mathcal{M}_1^{hom} \leftarrow \mathcal{M}_1$
Body:	for $t = 2$ to $t = T$ 1. compute y_t and $\mathcal{M}_t$ 2. compute $C(h_n, y_t)$ (and thus if necessary $\mathcal{M}_{h_n \cup y_t}$) 3. **if** $C(h_n, y_t) < threshold$ **then** merge : compute $\mathcal{M}_{h_n \cup y_t}$ if not already computed in step 2, update $\mathcal{M}_n^{hom} \leftarrow \mathcal{M}_{h_n \cup y_t}$ and $h_n \leftarrow h_n \cup y_t$ **else** create new segment: $h_{n+1} \leftarrow y_t$ and $\mathcal{M}_{n+1}^{hom} \leftarrow \mathcal{M}_t$

Figure 1: Content-based temporal segmentation algorithm using statistical motion models.

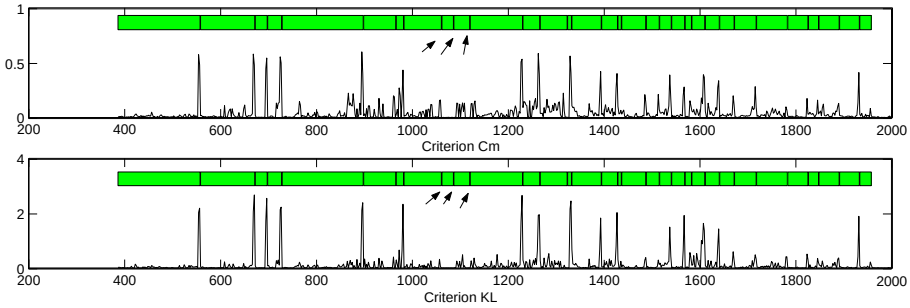


Figure 2: *Avengers* video: computed values of expressions $C_m(y_t, y_{t+1})$ (top row) and $KL(y_t, y_{t+1})$ (bottom row), between every two successive temporal units y_t and y_{t+1} of the video, and motion changes ground truth (top bar in each row).

on Figures 3-a and 4-a.

Before evaluating the performance of our temporal segmentation method, we first compare the ability of the two merging criteria $KL(y_t, y_{t+1})$ and $C_m(y_t, y_{t+1})$ to enhance motion-content dissimilarity. To that end, we have computed these criteria between every successive temporal units y_t and y_{t+1} of a video (without any merging yet at this stage). Let us stress that this does not evaluate the segmentation process itself since the latter considers the current homogeneous segment h_n and the next temporal unit y_{t+1}. We have plotted on Figure 2 the "instantaneous" values of the two criteria for the *Avengers* video (we have set $L = 2$ for these experiments). We have compared the obtained values to a manually-made ground truth of motion-content changes in the processed video. A peak in the criteria values is almost always observed when such a change occurs. It seems that the criterion $KL(y_t, y_{t+1})$ is more sensible to these changes than the criterion $C_m(y_t, y_{t+1})$ and consequently that incorporating the motion model of the merged seg-

ment $y_t \cup y_{t+1}$ smoothes the response of the criterion. Qualitatively, it means that even if a unit y_{t+1} can be poorly represented by the previous motion model $\mathcal{M}_t$, y_t and y_{t+1} can still be relatively well represented by the motion model corresponding to their fusion. Most of the motion-content changes in the *Avengers* video correspond to shot changes and the others (as the ones designated by the arrows on Figure 2) correspond to a dynamic evolution within the same shot. Obviously recovering the latters is a more difficult task because the change is less abrupt. The direct effect is a lower peak value of the criterion for this sort of break (see Figure 2).

When using the two criteria to perform the temporal segmentation of a given video, the difference in the response intensity tends to disappear. Indeed, we are now working with homogeneous segments h_n growing after each merging decision. As a result, the possible deviation brought by a new unit has less weight and the second term in expression (6), $KL(\mathcal{M}_{h_n}||\mathcal{M}_{h_n \cup y_t})$, decreases to zero as the size of h_n increases. Thus, the two criteria $KL(h_n, y_t)$ and $C_m(h_n, y_t)$ tend to give the same answer. As an illustration, Figures 3-b and 4-b display the results obtained with our temporal segmentation algorithm according to the decision criterion used and for the two video examples (due to page limitation we present the results for a excerpt of the complete videos). For comparison convenience, for both criteria, the threshold was choosen such that the number of resulting homogeneous segments was similar to the one in the manually-made ground truth. (Note that in practice, the strategy would be to apply the segmentation algorithm for a given threshold.) The obtained segmentations demonstrate the efficiency of the proposed method since all the key events have been well captured. On the *Avengers* example, all the motion changes are recovered and especially the dynamic changes within a shot (designed by the arrows). Nevertheless, motion-based segmentation remains subjective since it may exist several levels in motion discrimination depending on the objective in sight. As an illustration, the *Athletics* video is composed of five main successive activities (see Figure 4-a): a long jump event, a TV program advertisement, a pole vault, a high jump and again a pole vault. Both segmentation methods permit to isolate these different events. In addition, each event contains several motion changes (in particular during the TV program advertisement which includes many special effects) and we can observe that most of them are recovered with our segmentation method (Figure 4-b). In the two examples, a sequence of key images, one for each homogeneous segment detected by the motion-based segmentation can provide a concise and significant overview of the content of the processed video sequences.

The segmentation method does not allow to explicitly select the total number N of homogeneous segments. This number depends on the threshold value adopted in the merging decision. We have studied the sensibility of our motion-based segmentation method to the threshold setting. Figure 5 illustrates, for the *Avengers* video, the evolution of the number of resulting homogeneous segments when the threshold value increases, and for the two choices of decision criterion. We can observe that the two criteria present a similar evolution of the number of segments. The criterion $KL(h_n, y_{t+1})$ seems to lead to more segments for a given threshold. This is due to the fact that when a motion change occurs within a temporal unit u_t, a segment composed of this single unit tends to be formed, and this happens more often when using $KL(h_n, y_{t+1})$. Figure 5 shows that after a first phase of fast decrease of N a certain range of threshold values can be considered without affecting the segmentation results.

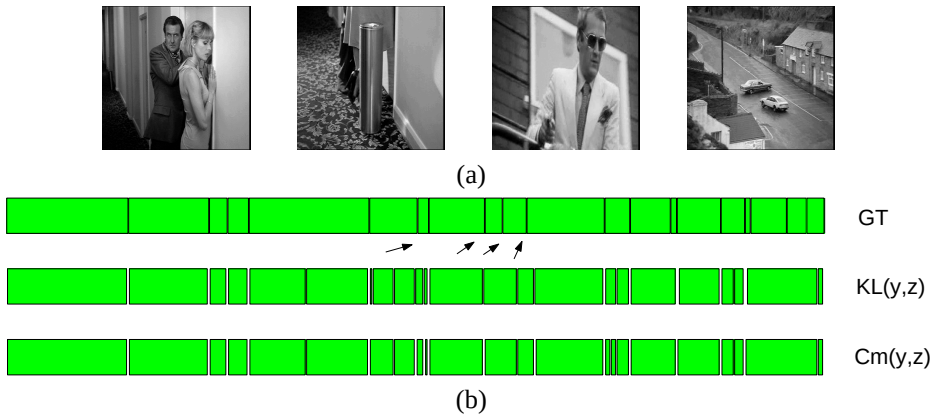


Figure 3: Excerpt of the *Avengers* video: (a), representative images of the video document, (b) top to bottom, manually-made ground truth, segmentation obtained with criterion $KL(h_n, y_{t+1})$, segmentation obtained with criterion $C_m(h_n, y_{t+1})$.

6 Concluding remarks

We have presented an original and simple method for automatic content-based video segmentation. This method does not require knowledge about the video genre or a prior segmentation into shots. It relies on the analysis of low-level motion information through statistical motion activity models which can capture and discriminate a large variety of dynamic situations. The temporal segmentation of the video document into homogeneous segments is realised by means of a merging decision criterion applied sequentially along the video. The experiments are encouraging for a low level method. They confirm that considering motion activity models is relevant to analyse video content. It highlights that the proposed sequential method for content-based video segmentation is able to properly discriminate different categories of motion activities, provided that the introduced motion models permit to capture in a flexible way a large range of dynamic situations and that the segmentation criterion is statistically well formalized. The resulting segments are coherent with respect to dynamic content and furthermore provide a meaningful overview of the video.

We have investigated two merging decision criteria based on the Kullback-Leibler divergence. The first one, $KL(y, z)$, only compares the motion models associated with the two concerned video units, while the second one, $C_m(y, z)$, takes also into account the motion model estimated from the potential union of the two video units. We have observed that when used in a sequential segmentation scheme they actually behave similarly. However, theoretically, the criterion $C_m(y, z)$ should express a better formulation of the merging problem since it takes into account the merged model. The observed "smoothing" effect of this criterion is probably due to the motion model used, offering a large number of degrees of freedom (i.e. a high number of parameters corresponding to the Gibbs model potentials). In this case, the merged model can fit well the data and can thus embedd each element of the fusion.

Experiments are currently carried out on a larger video base involving several hours

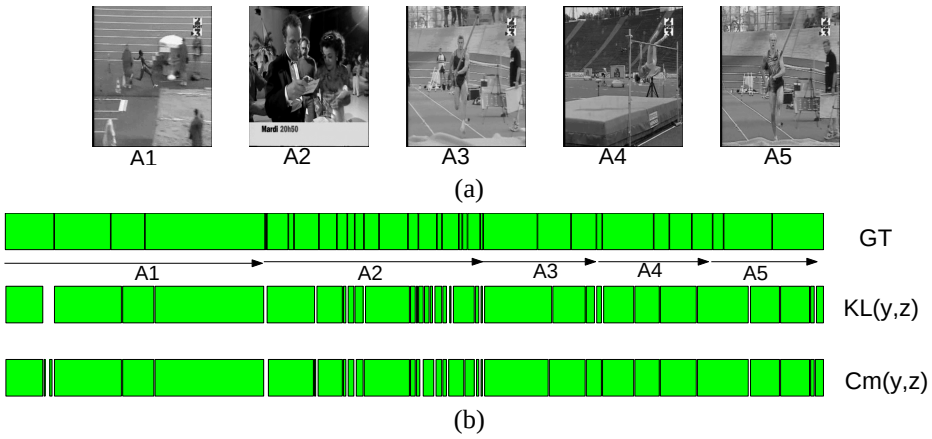


Figure 4: Excerpt of the *Athletics* video: (a), representative images of the video document, (b) top to bottom, manually-made ground truth, segmentation obtained with criterion $KL(h_n, y_{t+1})$, segmentation obtained with criterion $C_m(h_n, y_{t+1})$

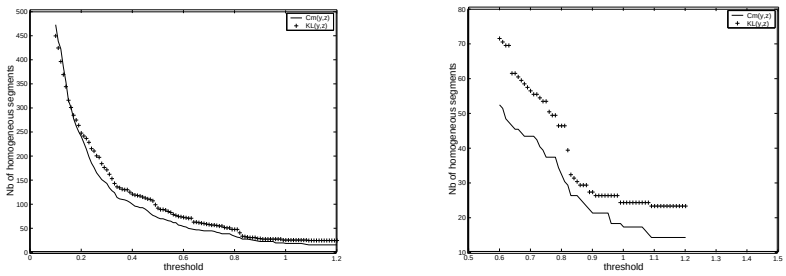
of video. In addition, the proposed segmentation method could be extended in two different ways, which are currently investigated. First, it would be interesting to control the relation between the total number of homogeneous segments and the merging threshold value. Second, the segmentation method relies on dynamic content analysis. Obviously, combining motion information with other informations such as color or audio features would place the segmentation at a more semantic level, as investigated in [7]. The advantage of the proposed statistical framework is that the integration of different information sources would be straigthforward. Our motion-based segmentation method, enriched with complementary video features, can be seen as a first step towards video summarisation. The video summary could be stated as the selection of pertinent segments among the homogeneous segments supplied by the content-based video segmentation stage. Our motion-based video segmentation algorithm can also be adapted to other tasks such as video indexing, temporal motion decomposition (e.g., analysis of sport gesture) or detection of unusual events (e.g., video surveillance).

Acknowledgments The work presented in this paper has been carried out with the financial support of the French Ministery of Industry, in the context of the RNTL project "Domus Videum". The authors acknowledge also INA, Département Innovation, Direction de la Recherche, for providing the videos.

References

[1] M. Basseville. Distance measures for signal processing and pattern recognition. *Signal Processing*, 18(4):349–369, 1989.

[2] J.S. Boreczky and L.A. Rowe. Comparison of video shot boundary detection techniques. In *SPIE Conference on Storage and Retrieval for Image and Video Databas-*



(a) (b)

Figure 5: *Avengers* video: number of homogeneous segments supplied by the motion-based temporal segmentation versus decision threshold, depending on the criterion, $C_m(h_n, y_{t+1})$ (solid line) and $KL(h_n, y_{t+1})$ (plus line), (a) threshold varying between 0.1 and 1.2, (b) zoom of Figure a.

es IV, volume SPIE 2670, pages 170–179, San Jose, January 1996.

[3] P. Bouthemy, M. Gelgon, and F. Ganansia. A unified approach for shot change detection and camera motion characterization. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(7):1030–1044, 1999.

[4] R. Fablet and P. Bouthemy. Non parametric motion recognition using temporal multiscale Gibbs models. In *IEEE International Conference on Computer Vision and Pattern Recognition, CVPR'01*, Kauai, Hawaii, December 2001.

[5] R. Fablet, P. Bouthemy, and P. Pérez. Non-parametric motion characterization using causal probabilistic models for video indexing and retrieval. *IEEE Transactions on Image Processing*, 11(4):393–407, 2002.

[6] C. Lacoste, R. Fablet, P. Bouthemy, and YF. Yao. Création de résumés de vidéos par une approche statistique. In *Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle, RFIA'02*, volume 1, pages 153–162, Angers, January 2002.

[7] J. Nam and H. Tewfik. Dynamic video summarization and visualization. In *7th ACM International Conference on Multimedia, ACM Multimedia'99*, pages 53–56, Orlando, November 1999.

[8] Y. Rui and P. Anandan. Segmenting visual actions based on spatio-temporal motion patterns. In *IEEE International Conference on Computer Vision and Pattern Recognition, CVPR'00*, volume 1, pages 111–118, Hilton Head, SC, June 2000.

[9] M.M Yeung, B.-L. Yeo, and B. Lui. Extracting story units from long programs for video browsing and navigation. In *3rd IEEE International Conference on Multimedia Computing and Systems, ICMCS'96*, pages 296–305, Hiroshima, June 1996.

[10] L. Zelnik-Manor and M. Irani. Event-based video analysis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR'01*, volume 2, pages 123–130, Kauai, Hawaii, December 2001.