Multimedia Tools and Applications, 26, 259–276, 2005

Motion-Based Selection of Relevant Video Segments for Video Summarization

NATHALIE PEYRARD nathalie.peyrard@avignon.inra.fr
INRA, Domaine Saint Paul Site Agroparc, 84914 Avignon Cedex 9, France

PATRICK BOUTHEMY patrick.bouthemy@irisa.fr
IRISA/INRIA, Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France

Abstract. We present a method for motion-based video segmentation and segment classification as a step towards video summarization. The sequential segmentation of the video is performed by detecting changes in the dominant image motion, assumed to be related to camera motion and represented by a 2D affine model. The detection is achieved by analysing the temporal variations of some coefficients of the 2D affine model (robustly) estimated. The obtained video segments supply reasonable temporal units to be further classified. For the second stage, we adopt a statistical representation of the residual motion content of the video scene, relying on the distribution of temporal co-occurrences of local motion-related measurements. Pre-identified classes of dynamic events are learned off-line from a training set of video samples of the genre of interest. Each video segment is then classified according to a Maximum Likelihood criterion. Finally, excerpts of the relevant classes can be selected for video summarization. Experiments regarding the two steps of the method are presented on different video genres leading to very encouraging results while only low-level motion information is considered.

Keywords: video segmentation, probabilistic motion modelling, supervised event classification

1. Introduction

Replacing a long video by a small number of representative segments provides a synthetic description of the document, which can be exploited for numerous applications including both home video and professional usages. However, the construction of video summary remains an open problem at the source of active research activities. The main difficulty obviously relies on the detection of semantic events from low-level information. Several approaches have been developed involving different video information and different representations. For instance, [21] proposed a strategy for video summarization and browsing by selecting key-frames maximally distinct and which carry the more information, based on the chrominance components of each pixel in the image. The video elementary unit considered in this method is simply the frame, which can be restrictive when trying to detect temporal semantic events. In [12], the authors present a generic method based on the modelling of user attention. Visual as well as audio information are combined to provide a user attention curve whose maxima define video segments likely to be of interest for the viewer. The combination of audio and image information is also exploited in [14] for video summarization. In this paper, we consider the task of selecting relevant temporal segments in a video and we adopt an approach based on motion-content analysis. It is obvious that

the use of complementary information, such as color or audio, would lead to better results. The aim here is not to fully solve the problem but to explore the potentiality of motion information for the specified task.

When dealing with video summarization, the first step usually consists in partitioning the video into elementary temporal segments. Such elementary units can be shots [1, 3, 4, 13, 20] which reveal the technical acquisition and editing processes of the video. We believe that a content-based segmentation, relying on the analysis of the evolution of the motion content in the sequence of images, should be more suited to the extraction of relevant dynamic events. In a previous approach [16], we have exploited the global motion in a video and a distance between probabilistic motion models to perform the video segmentation. In this paper we propose a simpler method based on the camera motion only. Indeed, a single shot can involve different successive camera motions and a camera motion change usually traduces a change in the activity content of the depicted scene. This segmentation is performed by detecting changes in the temporal evolution of coefficients of the 2D affine model representing the dominant motion in the images, the latter being assumed to be due to camera motion. Such a model has often been considered to estimate the camera motion, for instance in [13] using the MPEG motion vector of compressed video stream.

In a second stage, we apply a supervised classification algorithm based on the characterization of the residual image motion. Indeed, if the dominant image motion is due to camera motion, the residual motion (i.e., after subtracting the dominant motion in the global image motion) can be related to the projection of the scene motion. Once temporal units of the processed video are identified, one way to characterize their dynamic content would be to consider again parametric motion models (e.g. 2D affine or quadratic motion models). However, the dynamic situations which can be described by such models are too restricted. They are suitable for modelling camera motion but no more for modelling general scene motion. Several motion characterizations have been investigated. In [1], motion-related features are derived from the computation of the optical flow field. Based on these features, a method for video indexing is proposed. The study described in [18] for the detection of a sequence of home activities in a video relies on segmenting moving objects and detecting temporal discontinuities in the successive optical flow fields. It involves the analysis of the evolution of the most significant coefficients of the singular value decomposition (SVD) of the set of successive flow fields. Still dealing with video content characterization but in the context of video classification into genres, statistical models are introduced by [20] to describe two components of the video structure: shot duration and shot activity. In [22], in order to cluster temporal dynamic events, the latter are captured by the computed histograms of 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. Recently in [11], the authors have proposed motion pattern descriptors extracted from motion vector fields and have exploited support vector machines (SVM) for classification of video clips into semantic categories. For an objective very close to video summarization, shots overview, the method in [5] relies on the nonlinear temporal modelling of wavelet-based motion features. As [22], we propose to exploit low-level motion measurements but conveying more elaborated motion information, while still easily computable compared to the optic flow. These local measurements are straightforwardly extracted from the images

intensities and are exploited with statistical motion models as introduced in [8]. These models are specified from the temporal co-occurrences of the quantized local motion-related measurements and can handle a wide range of dynamic contents. Exploiting this statistical framework, we propose to label each video segment according to learned classes of dynamic events, using a Maximum Likelihood criterion. Then, only the segments associated to classes defined as relevant in terms of dynamic events are selected, and excerpts of these significant segments could be further exploited for video summarization.

Section 2 describes the temporal video segmentation stage based on camera motion, and the behaviour of the resulting algorithm is illustrated on videos of different genres. In Section 3, we present the classification stage relying on a probabilistic motion modelling, and the global two-step method is applied for the recognition of relevant events in two sports videos. Section 4 contains concluding remarks.

2. Temporal video segmentation

In this section, we present the first stage of our method for selecting relevant segments in a given video. It consists in performing a sequential segmentation of the video into homogeneous segments in terms of camera motion. Its performance is illustrated on three real video documents. In particular, we compare the results with those obtained with other types of segmentation (segmentation into shots or based on global motion content) to confirm the suitability of the segmentation presented in this paper for the objective in sight, i.e., the detection of particular dynamic events in a video.

2.1. Camera motion modelling and estimation

The video segmentation relies on the analysis of the dominant image motion computed between two successive frames of the video. The dominant image motion is assumed to correspond to the apparent motion of the background induced by the 3D camera motion. It is possible to represent the projection of the 3D motion field (relative to the camera) of the static background by a 2D parametric motion model (assuming a shallow environment, or accounting for its main part if it involves different depth layers). For example, we can deal with the affine motion model defined at image point $p = (x, y)$ by:

$$\mathbf{w}_\theta(p) = \begin{pmatrix} a_1 + a_2 x + a_3 y \\ a_4 + a_5 x + a_6 y \end{pmatrix}, \tag{1}$$

where $\theta = (a_1, a_2, a_3, a_4, a_5, a_6)$ is the motion model parameter vector. Such a model can handle different camera motions: panning, zooming, tracking (including of course static shots). For more complex situations, a quadratic model can be used but we will restrict ourselves here to the affine model. It forms a good trade-off between the relevance of the motion representation and the complexity of the estimation. The model parameters θ of the dominant image motion are then estimated using the gradient-based robust multiresolution algorithm designed in [15]. More precisely, the robustness is ensured by the minimization

of a M-estimator criterion. The constraint is derived from the assumption of brightness constancy and the parameter estimator of the affine motion model between frame $I(k)$ and frame $I(k+1)$ is defined as:

$$\hat{\theta} = \arg\min_{\theta} \sum_{p \in \mathcal{R}} \rho(DFD_{\theta}(p)),$$

where $DFD_{\theta}(p) = I(p + \mathbf{w}_{\theta}(p), k+1) - I(p,k))$ is the Displaced Frame Difference and $\mathcal{R}$ is the spatial image grid. The M-estimator ρ is chosen as a hard-redescending function. The minimization takes advantage of a multiresolution framework and an incremental scheme based on the Gauss-Newton method. It is implemented as an iteratively reweighted least-squares technique. This method yields an accurate estimation of the dominant motion between two images even if other secondary motions are present.

2.2. Detection of camera motion changes

In order to detect changes in the camera motion, we analyse the temporal evolution of the two translation coefficients a_1 and a_4 of the affine model (1). In general, a change in camera motion induces a jump in the evolution of these two signals. To detect such ruptures we propose to apply a Hinkley test on each signal, $a_1(k)$ and $a_4(k)$. This statistical test is issued from likelihood ratio tests, evaluating the "no change hypothesis" (no change between frames $k-1$ and k) versus the "change hypothesis". It provides a simple and efficient means to detect jumps in the mean of a signal and it is known to be robust since cumulative (see [2]). Since the direction of the change in the mean of the signal is unknown, in practice, two tests are performed in parallel, to look for respectively a decrease and an increase in the mean. Let us consider first the signal $a_1(k)$ and the case of testing for a decrease. Let μ_0 be the value of the mean before the change occurs and $\frac{j_{\min}}{2}$ the, a priori chosen, minimum jump magnitude. The sequence D_n defined as follows:

$$D_0 = 0, \quad D_n = \sum_{k=1}^{n} \left(a_1(k) - \mu_0 + \frac{j_{\min}}{2} \right)$$

represents the cumulative sum of the differences between signal a_1 and $\mu_0 - \frac{j_{\min}}{2}$. A jump is detected when $d_n - D_n > \lambda$, with $d_n = \max_{0 \le k \le n} D_k$ and λ a predefined threshold. Intuitively, it means that a change in the mean is detected when a value of a_1 is significatively smaller than $\mu_0 - \frac{j_{\min}}{2}$ and the phenomenum is not isolated. The mean μ_0 is estimated on-line and reinitialised after each jump detection. In the case of testing for an increase, the test performed is defined by:

$$U_0 = 0, \quad U_n = \sum_{k=1}^{n} \left(a_1(k) - \mu_0 - \frac{j_{\min}}{2} \right)$$
$$u_n = \min_{0 \le k \le n} U_k, \quad \text{alarm if } U_n - u_n > \lambda$$

When a change is detected, the jump location is given by the last k satisfying $d_k - D_k = 0$ or $U_k - u_k = 0$, the variable D_n is reinitialised to 0 and the next search starts from the instant (image) following the detected jump location.

Similarly and in parallel, the Hinkley test is performed for detecting ruptures in signal a_4, corresponding to the two detection rules:

$$M_0 = 0, \quad M_n = \sum_{k=1}^{n} \left(a_4(k) - \mu_0 + \frac{j_{\min}}{2} \right)$$
$$m_n = \min_{0 \le k \le n} M_k, \text{ alarm if } m_n - M_n > \lambda,$$
$$R_0 = 0, \quad R_n = \sum_{k=1}^{n} \left(a_4(k) - \mu_0 - \frac{j_{\min}}{2} \right)$$
$$r_n = \min_{0 \le k \le n} R_k, \text{ alarm if } R_n - r_n > \lambda$$

Note that for (perfect) zooming and for a static shot, the coefficients a_1 and a_4 are supposed to be zero. Thus, if two successive shots are two static shots, two pure zooming motions or one is a static shot and the other one involves a pure zooming motion, no change would occur in a_1 and a_4 values. In practice, if the two shots are separated by a cut transition, (high) erroneous measures of a_1 and a_4 are observed at the cut location and the motion change is detected all the same. Besides, the method could be completed by the analysis of the temporal evolution of other parameters of model (1). More precisely, in the case of a pure zoom, the two diagonal coefficients a_2 and a_6 are supposed to be equal in theory and they often exhibit a rather constant slope over the time interval corresponding to the zooming motion. Consequently, performing a Hinkley test on the temporal derivate of the signal $a_2(t)$ should allow us to detect changes between two zooms or between a zoom and a static shot. In practice, it seems more reasonable to work with the divergence parameter $div = \frac{1}{2}(a_2 + a_6)$ for stability reasons.

2.3. Results

We have carried out experiments on three real video documents of different genres (a movie, a documentary and a sport program). For each example, we compare the result of the automatic segmentation based on camera motion with a manually-made segmentation according to the same criterion. We compare also with a manual segmentation into shots and with the method proposed in [16]. With the latter, homogeneous video segments are built in a sequential way by analysing the temporal variations of the global motion (i.e., residual plus dominant motion) in the sequence of images. The global motion of successive temporal units of the video is described by a statistical motion model as introduced in [8] (see also Section 3.1). Then, a merging decision criterion is considered, 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.

Figure 1. Irisa video. Representative images of the video.

2.3.1. Irisa video. The first document, the *Irisa* video, is an excerpt of a video document presenting the institute IRISA. Some representative images of this video are displayed on figure 1. The processed excerpt contains 710 frames and involves different camera motions and several cuts and dissolves. The manually-made ground truth segmentation in terms of camera motion changes is given in figure 2 (top ribbon). The goal of the proposed video segmentation method is only to detect changes in camera motion and not to identify the nature of this motion. However, the latter can help understanding the behaviour of the segmentation method. The successive camera motions observed in the *Irisa* video are the following: static, (dissolve transition), left pan, static, (dissolve transition), static, (cut), complex motion, (cut), static, left pan, combined left pan and down tilt, left pan, (dissolve transition), left pan, left pan, right pan, (cut), static, left pan, static. Results of our temporal video segmentation method are displayed on figure 2. The three cuts and most of the camera motion changes within a shot have been recovered. The dissolve transitions (indicated by

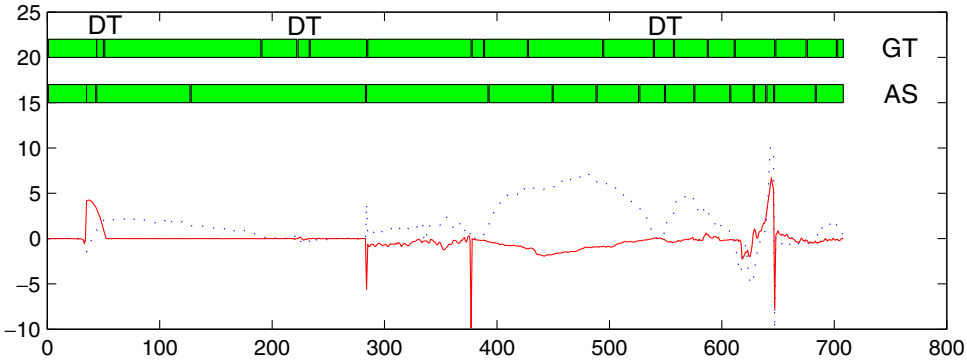

Figure 2. *Irisa* video. Bottom: temporal evolution of the translation parameters a_1 (dot line) and a_4 (full line). Top: Automatic Segmentation based on these parameters (with $j_{\min} = 1$ and $\lambda = 10$) (AS ribbon) and Ground Truth (manually derived) segmentation of the video in terms of camera motion (GT ribbon) with localisation of the dissolve transitions (DT).

the mark "DT" in figure 2) are poorly recovered due to the particularly slow evolution in parameters a_1 and a_4 they lead to. Wrong values are observed around image 650. This corresponds to a passage with majoritary black images where the dominant motion model (1) is poorly estimated.

2.3.2. Athletics video. The second document, the *Athletics* video, is part of a TV sport program corresponding to an athletics meeting. The second example is an excerpt of 1416 frames. This video is composed of five main successive activities: a long jump event (A1), a TV program advertisement (A2), a pole vault (A3), a high jump (A4) and again a pole vault (A5). In addition, each event contains several camera motion changes. This succession of events can be recovered from the analysis of the two signals $a_1(t)$ and $a_4(t)$ (see figure 3, bottom). For instance, during each pole vault the camera follows the athlet motion: the run-up, the ascent and then the descent after getting over the bar. The camera motion is thus successively a backward zooming motion, and an upward and a downward tilt. The two pole vaults appear clearly around image 2400 and image 2750, with successively a plateau around zero and then a fast increase and a fast decrease of a_4. In the case of the high jump, the camera motion is not that large. High or very noisy values of the two parameters are estimated during the TV program advertisement, due to many special effects which make difficult the estimation of the camera motion within the corresponding images. During the long jump (activity A1), the camera is first static the instants before the athlet starts running, then the camera motion is a pan while following the run-up. During the jump, the camera motion is successively a backward and a forward zooming. Then, the camera follows the athlet in a right and a left pan. These two successive pans can be recovered from the evolution of signal $a_1(t)$ before and after image 1800 approximatively. However, because of light

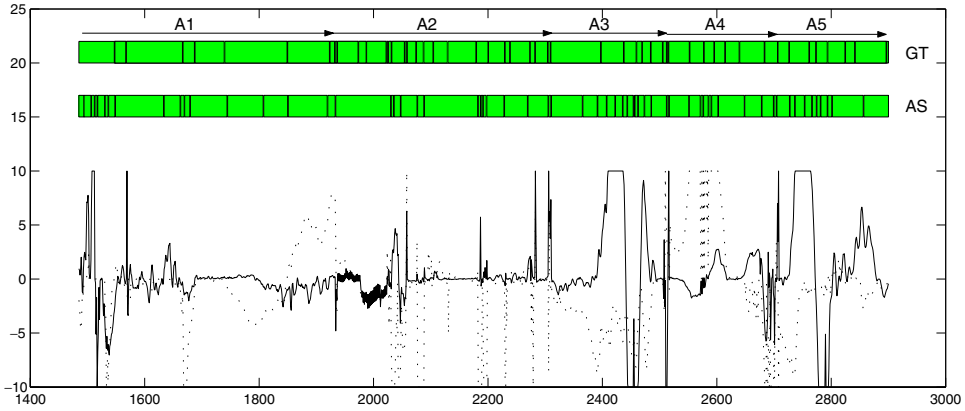

Figure 3. *Athletics* video. Bottom: temporal evolution of the translation parameters a_1 (dot line) and a_4 (full line). Top: Automatic Segmentation based on these parameters (with $j_{\min} = 1$ and $\lambda = 20$) (AS ribbon) and Ground Truth (manually derived) segmentation of the video in terms of camera motion (GT ribbon) with localisation of the successive main activities, long jump event (A1), TV program advertisement (A2), pole vault (A3), high jump (A4) and pole vault (A5).

variations, the parameters a_1 and a_4 are poorly estimated in the beginning of the activity. Globally, one can observe that most of the camera motion changes are recovered by the automatic segmentation method (low rate of non-detection). This is not the case however for the part A2 of the video, for the reason previously mentioned. The automatic method tends to oversegment the video in parts A3, A4, and A5, compared to the manually-made segmentation. However, it corresponds to the very moment when the athlets accomplish their jump and thus to a complex and quite fast camera motion. Besides, let us point out that building a ground-truth segmentation of the camera motion is also not straightforward and might be questionable in complex situations.

2.3.3. Avengers video. The last video processed, the *Avengers* video, is a sequence of images from the TV series "Avengers". The example contains 3496 frames. Three types of scenes form the excerpt: corridor scenes (T1), balcony scenes (T2) and street and cars scenes (T3). The manually-made segmentation in terms of camera motion as well as the automatic segmentation are plotted in figure 4 (respectively ribbon c1 and c2). Here again, most of the ruptures are detected. The automatic segmentation leads to an oversegmentation in the first third of the video, probably due to the successive slight increases and decreases in the velocity of the camera following the actors in the corridor scenes. Such changes are difficult to detect manually.

2.3.4. Comparison and conclusion. With Figures 4–6, we are able to compare the content structure of a video recovered when segmenting according to shot changes, global image motion changes and camera motion changes. These examples illustrate the fact that a

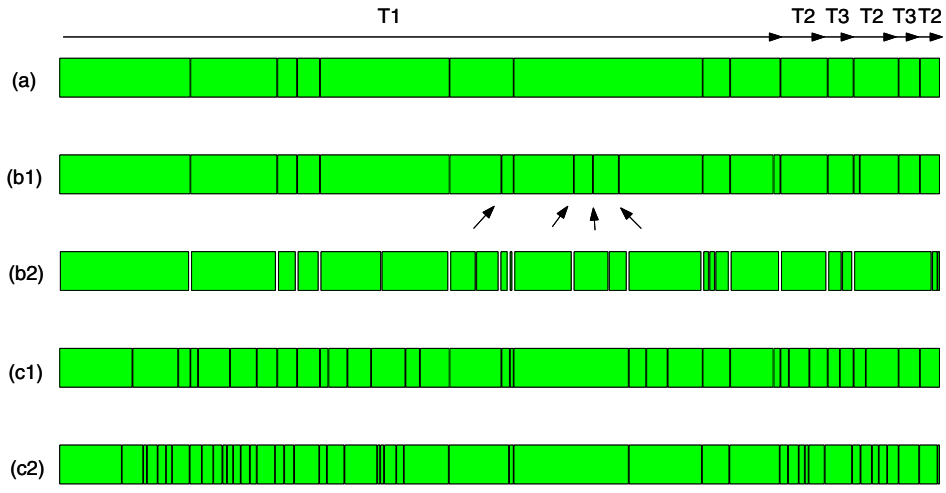

Figure 4. *Avengers* video. (a) manually-made segmentation into shots with successive scenes types (corridor scenes (T1), balcony scenes (T2) and street and cars scenes (T3)), (b1) and (b2) respectively manually-made and automatic segmentation based on global motion, (c1) and (c2) respectively manually-made and automatic segmentation based on camera motion ($j_{\min} = 1$ and $\lambda = 10$).

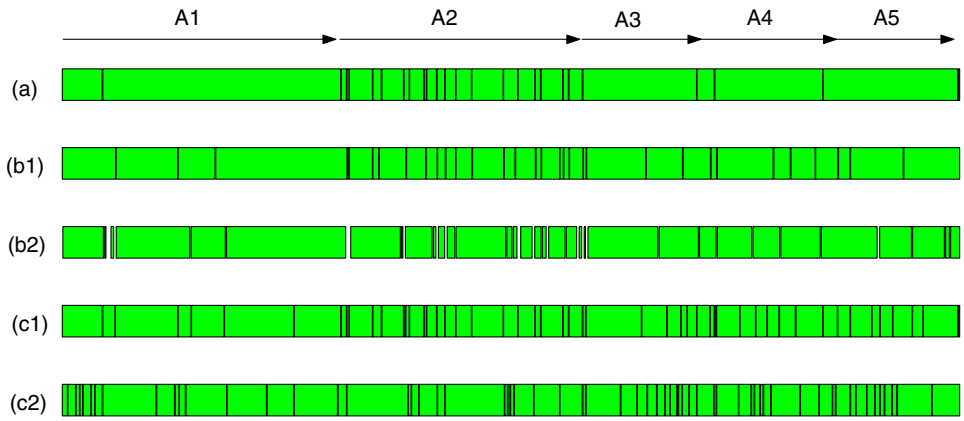

Figure 5. *Irisa* video. (a) manually-made segmentation into shots, (b1) and (b2) respectively manually-made and automatic segmentation based on global motion, (c1) and (c2) respectively manually-made and automatic segmentation based on camera motion ($j_{\min} = 1$ and $\lambda = 10$).

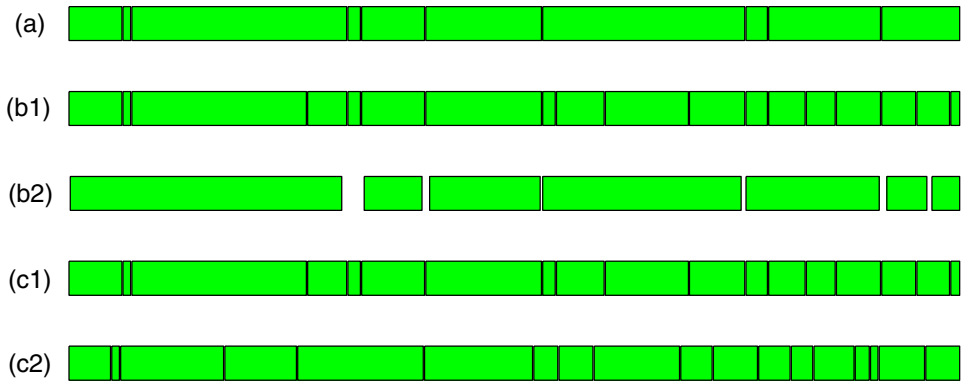

Figure 6. *Athletics* video. (a) manually-made segmentation into shots, (b1) and (b2) respectively manually-made and automatic segmentation based on global motion, (c1) and (c2) respectively manually-made and automatic segmentation based on camera motion ($j_{\min} = 1$ and $\lambda = 20$).

segmentation into shots is usually at a coarser level than a segmentation in terms of motion. Generally, the segmentation based on global image motion seems to be at an intermediate level between segmentation into shots and segmentation based on camera motion. The two methods for video segmenting according to motion variations allow us to recover most of the cuts in an indirect way since, as mentioned before, a cut will cause erroneous motion measurements interpreted like a motion change. However, whatever the method, special effects in videos are imperfectly interpreted and there is still some progress to do at this stage. With the method based on camera motion, we cannot guarantee the detection of all scene motion changes and in particular in the case of a static background with a moving

scene (as illustrated in figure 4 with the four breakings indicated by the arrows). However, the segmentation method based on global motion seems to suffer more from non-detection than the one based on camera motion. Furthermore, it can lead to substantial oversegmentation (blank spaces in ribbon b2 of Figures 4–6 represent blocks of images where the method has detected a change every three images). All this validates our choice for the method based on camera motion.

More generally, we observed for the proposed segmentation method, based on camera motion, few non-detected changes and reasonable oversegmentation. Since the result of the segmentation is used as a preliminary step to initialise the classification, oversegmentation will not be a drawback. In all our experiments, the ability of the method to provide homogeneous segments in terms of camera motion has been proved, which is the main requirement to carry on with the second step.

3. Classification of the video segments

The second step consists in classifying and selecting the segments, supplied by the segmentation step, according to their dynamic content. We deal now with the real motion content of the scene depicted by the video: the residual image motion. We adopt a statistical representation of the residual motion content as presented below. The efficiency of the proposed method is evaluated on two sports videos about which categories of relevant events can be explicitly defined.

3.1. Statistical motion model

To extract meaningful dynamic events, we rely on the probabilistic modelling of the residual motion content of a video. In order to handle a wide range of dynamic situations (outdoor and indoor scenes, sports scenes, ...), we resort to a general notion of motion activity and we exploit the statistical motion models recently introduced in [8]. The motion-related measurements considered are low-level measurements related to the normal flow. They convey a more elaborated local motion information than the local histograms of the spatio-temporal intensity gradient proposed in [22], while still locally and easily computable contrary to the full optic flow. These measurements can indeed be efficiently and reliabily computed for any video whatever its genre and content. This framework for motion activity modelling has already been successfully applied to motion recognition [7] and motion-based video retrieval [8]. In this section, we outline its main characteristics.

The motion activity model is 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 residual normal flow $\mathbf{v}_n(p,k)$ is computed as follows:

$$\mathbf{v}_n(p,k) = \frac{-DFD_{\hat{\theta}_k}(p,k)}{||\nabla I(p,k)||} \cdot \frac{\nabla I(p,k)}{||\nabla I(p,k)||},$$

where $\hat{\theta}_k$ is the estimated parameter of the 2D affine camera motion model, $DFD_{\hat{\theta}_k}(p,k)$ is the Displaced Frame Difference (see Section 2.1) and $\nabla I(p,k)$ is the spatial intensity

gradient. Then, a continuous local motion measure is computed as a weighted mean, over a small spatial window, of the residual normal flow magnitude in order to obtain a more reliable motion information. Indeed, the accuracy of the evaluation of the residual normal flow is highly dependent on the norm of the spatial intensity gradient, and this accuracy increases with $||\nabla I(p,k)||$. Thus, the continuous local motion measure considered is:

$$v(p,k) = \frac{\sum_{q\in\mathcal{F}(p)} ||\nabla I(q,k)||^2 . |\mathbf{v}_n(q,k)|}{N_{\mathcal{F}(p)} \max\left(G^2_{\min}, G^2_{\text{moy}}(p)\right)}, \tag{2}$$

where $\mathcal{F}(p)$ is a local window centered in p and of size $N_{\mathcal{F}(p)}$ (typically 3×3 pixels), $G^2_{\text{moy}}(p)$ is the normalization factor and the constant $G_{\min}$ avoids a null denominator in uniform areas. This measure can a priori be computed for any video type.

The continuous (positive) local motion-related measure (2) is then quantized on a set Λ of discrete values, leading to the measure $y(p,k)$. A causal Gibbs probabilistic distribution [9] can represent the temporal co-occurrences of the quantized local motion measurements $\{y(p,k),\ p = 1 \ldots |\mathcal{R}|, k = 1 \ldots K\}$, where $\mathcal{R}$ is the spatial image grid and K is the length of the sequence. More precisely, given an image sequence, we compute the associated sequence $y = \{y(k), k = 1 \ldots K\}$ of local motion quantities maps (one motion map $y(k) = \{y(p,k),\ p = 1 \ldots |\mathcal{R}|\}$ is computed from two successive images).

The temporal co-occurrences distribution $\Gamma(y)$ of the sequence y is a matrix $\{\Gamma(\nu, \nu'|y)\}_{(\nu,\nu')\in\Lambda^2}$ defined by:

$$\Gamma(\nu, \nu' \mid y) = \sum_{k=2}^{K} \sum_{p\in\mathcal{R}} \delta(\nu, y(p,k)) \cdot \delta(\nu', y(p,k-1)), \tag{3}$$

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

$$\Psi_{\mathcal{M}} \bullet \Gamma(y) = \sum_{(\nu,\nu')\in\Lambda^2} \Psi_{\mathcal{M}}(\nu, \nu') \cdot \Gamma(\nu, \nu'|y), \tag{4}$$

between the potentials $\Psi_{\mathcal{M}}$ and the matrix of the temporal co-occurrences $\Gamma(y)$ [8]:

$$P_{\mathcal{M}}(y) = \frac{1}{Z} \exp\left[\Psi_{\mathcal{M}} \bullet \Gamma(y)\right], \tag{5}$$

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

The appealing characteristic of these models is that, due to their causal aspect, the normalization constant Z is explicitly known and tractable. Furthermore, it is independent of the model $\mathcal{M}$. Thus, the probability (5) can be exactly determined and available for any sequence y and model $\mathcal{M}$. The model estimation is achieved according to the Maximum Likelihood criterion. It is easy to see that the temporal model (5) 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 Maximum Likelihood estimate is given by the empirical estimate of T, and for an observed sequence y, the estimated potentials $\Psi_{\widehat{\mathcal{M}}}$ are deduced from the co-occurrences distribution $\Gamma(y)$ as follows:

$$\Psi_{\widehat{\mathcal{M}}}(\nu, \nu') = \ln\left(\frac{\Gamma(\nu, \nu'|y)}{\sum_{\nu''\in\Lambda}\Gamma(\nu'', \nu'|y)}\right). \tag{6}$$

The use of these statistical motion activity models appears simple and efficient. The computation of the temporal co-occurrences $\Gamma(y)$ can be realised in a parallel scheme. Once the temporal co-occurrences distribution is available, the model estimation is straightforward. Besides, the evaluation of the likelihood (5) requires only the computation of dot products between the model potentials and the co-occurrences matrix coefficients, which is of low computation time.

The number of coefficients of this model, equal to $|\Lambda|^2$, is large (usual choices for Λ are quantization in 16 or 32 levels). In practice, it is possible to reduce the model complexity by selecting the more informative potentials, based on the computation of likelihood ratios [8]. However, it is still an open issue to evaluate its impact in terms of motion recognition.

3.2. Supervised classification and selection

For a given type of video document, a training step is performed off-line. It implies to manually define the different classes of motion content present in the video genre (and thus their number G), to extract, in each video of the training base, segments homogeneous in terms of motion activity and to manually label each segment. The choice of G remains a difficult task since there is never an obvious true set of classes but rather a hierarchical structure of classes and sub-classes. For the study presented in this article the extraction, as the labeling of the homogeneous segments of the training base, have been made manually in order to have a training set as perfect as possible. However, this is very tedious and time consuming. Ideally, one would expect an unsupervised classification method (i.e. automatic extraction of homogeneous segments in the training set and clustering, the naming of the different resulting classes being still manually-made) but this remains a difficult objective. An intermediate solution could be to use the automatic segmentation method presented in the previous section to extract homogeneous segments from the video base and then to manually label them.

In our supervised scheme, the probabilistic models are trained as follows. Once the G classes are defined, for each class g the Maximum Likelihood estimators $\hat{\Psi}_g$ of the potentials of the causal Gibbs model $\mathcal{M}_g$ describing the class g are evaluated according to

$$\hat{\Psi}_g = \arg\max_{\Psi} \prod_{s\in g} P_{\mathcal{M}_\Psi}(y_s),$$

y_s being the quantized motion-related measurements associated with the segment s in class g. It is equivalent to compute the Maximum Likelihood estimates of the potentials

according to Eq. (6), with $\Gamma = \Gamma_g$, Γ_g being the sum of the co-occurrence matrices of all the segments belonging to class g.

Let us consider now the task of extracting meaningful events in a video test. Given the partition $\{s_0, \ldots, s_N\}$ of the video test into homogeneous segments and $\{y_0, \ldots, y_N\}$ the corresponding motion-related measurements, each video segment is labeled with one of the learned classes of dynamic events according to the Maximum Likelihood criterion, as follows:

$$\begin{aligned} l_n &= \arg \max_{1 \le g \le G} P_{\mathcal{M}_g}(y_n), \\ &= \arg \max_{1 \le g \le G} \hat{\Psi}_g \bullet \Gamma(y_n) \end{aligned} \tag{7}$$

where l_n is the label of segment s_n. Then, only the segments associated to classes defined as relevant in terms of dynamic event are selected. Excerpts of these significant segments could be further exploited for video summarization.

3.3. Results

The proposed method for the selection of relevant dynamic events is evaluated on two sport videos about which classes of relevant events can be explicitly defined.

3.3.1. Albertville video. The first video processed corresponds to a (dance) figure skating TV program. The first 23 minutes of the video (displaying two shows) form the training set and the last 9 minutes of the video (one show) form the test set. Each video segment supplied by the camera motion-based segmentation method is then classified according to the Maximum Likelihood criterion (7). This is a good example to evaluate the potentiality of a motion-based video analysis, or more specifically to answer the question: how close to the semantic level can we hope to come starting from a low-level motion information? We will see that the classification specification is an important factor in the performance of a recognition method.

We first considered a classification in two categories, corresponding to *play* and *no play* events. Category *play* corresponds to all skating motions and category *no play* includes scenes of the skaters waiting for their scores, scenes of persons in the audience, and all quasi static scenes. This remains at a coarse level in terms of semantic meaning but can still be of interest for the creation of a video summary. The results of this sorting are reported in Tables 1(a) and (b). Table 1(a) shows that 82% (resp. 87%) of *play* segments (resp. *no play* segments) are correctly detected. With Table 1(b), we can see that among the segments classified as *play* by our algorithm, only 3% are *no play* segments (false alarms). According to the high rates of good classification obtained, a strategy for summarization could then be in a first stage to apply a sorting into *play* and *no play* categories based on the residual motion, and in a second stage to select the pertinent segments within those labeled as *play* based on additional information (such as camera motion, color, audio information, ...).

Table 1. *Albertville* video. (a) Classification matrix for two pre-defined classes of motion content (the row labels stand for true labels and the column labels stand for detected labels), (b) number of detected segments and false alarms for the class *play*.

	Play	No play
Play	**0.82**	0.18
No play	0.13	**0.87**

(a)

	Play
Total number	319
Detected	261 (82%)
False alarms	5 (3%)

(b)

We then considered a finer level of classification, with five classes of dynamic content defined as: *scores* (Sc, skaters and coach moving in the stand, waiting for the scores), *no motion* (No, static or quasi static scenes), *performance* (Pe, all skating motions), *slow motion* (Sl, replay) and *audience* (Au, persons in the audience moving on the terraces). The corresponding classification results are given in Table 2(a). The discrimination power of the method remains satisfactory, in particular for class *performance* (83% of the segments in this class are detected). Here again we could imagine to build a video summary from segments associated to this class, in particular if considering the low rate of false alarms (see Table 2(b)). Note that category *play* is formed of classes *performance* and *slow motion*, and category *no play* is composed of classes *scores*, *no motion* and *audience*.

Finally, we considered the case of seven classes of dynamic content which are closer to semantic classes than to motion classes: *scores* (Sc, skaters and coach moving in the stand, waiting for the scores), *static* (St, static scene), *waving* (Wa, skaters waving to the audience), *figure* (Fi, artistic effects, mainly dance movements), *skating* (Sk, only simple skating motions), *slow motion* (Sl, replay) and *audience* (Au, persons in the audience moving on the terraces). Class *no motion* has been split into classes *static* and *waving* and class *performance* has been split into classes *figure* and *skating*. The numbers of segments of each class in the test set are reported in Table 3 and the full classification results are gathered in Tables 4(a) and (b). We observe a null recognition rate for the class *waving* (which nevertheless is not really significant since this class is represented by a very small number

Table 2. *Albertville* video. (a) Classification matrix for five classes of motion content (the row labels stand for true labels and the column labels stand for detected labels), (b) number of detected segments and false alarms for the class *performance*.

	Sc	No	Pe	Sl	Au
Sc	**0.61**	0.22	0.17	0	0
No	0	**0.80**	0.10	0	0.10
Pe	0.025	0.10	**0.83**	0.045	0
Sl	0	0.33	0.17	**0.50**	0
Au	0.33	0	0.33	0	**0.33**

(a)

	Pe
Total number	313
Detected	261 (83%)
False alarms	7 (2.6%)

(b)

Table 3. *Albertville* video. Number of segments in each of the seven classes of motion content in the test set.

Sc	St	Wa	Fi	Sk	Sl	Au
18	7	3	30	283	6	6

Table 4. *Albertville* video. (a) Classification matrix for seven classes of motion content (the row labels stand for true labels and the column labels stand for detected labels) and (b) number of detected segments and false alarms for the class *skating*.

	Sc	St	Wa	Fi	Sk	Sl	Au
Sc	**0.61**	0.11	0.11	0	0.17	0	0
St	0	**0.86**	0	0	0	0	0.14
Wa	0	0.67	**0**	0	0.33	0	0
Fi	0	0.07	0.20	**0.20**	0.50	0.03	0
Sk	0.03	0.01	0.07	0.23	**0.62**	0.046	0
Sl	0	0	0.33	0.17	0	**0.50**	0
Au	0.33	0	0	0	0.33	0	**0.33**

(a)

	Sk
Total number	283
Detected	175 (62%)
False alarms	21 (10.7%)

(b)

of segments in the processed video), segments of this type being in majority misclassified in class *static*. These classes are indeed too similar in terms of motion content. The motion of the skaters waving corresponds to a too small part of the image to be highlighted by the model. It appears also that the two classes *figure* and *skating* tend to be mixed up. Since the skaters present a dance show (as opposed to shows with jumps and spins), the artistic movements performed by the skaters are rather smooth and similar to their regular skating motion. This is confirmed by the recognition rate obtained in class *performance* (83%) in the previous experiment, and by the rate obtained in the two classes *figure* and *skating* when considering the two best maximizers of the likelihood (more than 80% in both classes). As a consequence, it seems that the last classification problem is too ambitious, regarding the available and computable motion information. This example points out the difficulty to find the level of semantic meaning effectively reachable, the classification specification being very important.

3.3.2. Athletics video. The second example processed is another excerpt of the *Athletics* video previoulsy presented in Section 2.3. We have chosen a different sequence in order to have a training set large enough for each class of motion content (which was not the case for long jump and high jump). Six different motion classes are observed on this sample: interview shots (Int), large views of the stadium (LV), pole vault (PV), replay of pole volt (Sl, including not only the jump but also the run-up), wide-shots of track race (WR) and close-up of track race (CR). We have selected 10 minutes of the video as a training set and 5 minutes as a test set. The distribution of the segments supplied by the segmentation method among the six motion classes is given in Table 5. The classification results are provided

Table 5. *Athletics* video. Number of segments in each of the six classes of motion content in the test set.

Int	LV	PV	Sl	WR	CR
9	1	13	18	45	25

Table 6. *Athletics* video. Classification matrix for the six classes of motion content.

	Int	LV	PV	Sl	WR	CR
Int	**0.11**	0	0.22	0.11	0.56	0
LV	0	**1**	0	0	0	0
PV	0	0	**0.77**	0	0.23	0
Sl	0	0	0	**1**	0	0
WR	0.25	0	0.02	0.02	**0.69**	0.02
CR	0	0	0.16	0	0.16	**0.68**

by Table 6. Apart for class Int, the obtained results are satisfactory (about 80% in average for the five other classes). Some misclassifications correspond to segments of few images (less than 10), which is probably too short to capture the scene motion. We observe also some misclassifications between class WR and CR, for segments including a progressive transition between a wide shot and a close-up. Let us note that we generally observe an increase of the recognition rate when considering the two first maximisers of the likelihood: for instance we reach 0.92% (resp. 0.98%) for the class pole vault, (resp. wide-shot of track race). These results ensure that a relevant motion information has been extracted with the causal Gibbs models.

4. Concluding remarks

We have presented a two-stage approach for extracting meaningful dynamic events in a video, based on motion-content analysis. The first stage, the temporal segmentation step, is based on the detection of changes in camera motion. This segmentation method is simple and generic. The experiments show that the video segments obtained form reasonable temporal units within which to apply, in a second step, a recognition and selection algorithm. At this stage, our evaluation is purely visual and a quantitative evaluation could be envisaged as the protocol proposed in [10] in the context of content-based image retrieval. The second stage, the classification step, is supervised. The experimental results are encouraging, since the method involves only a low-level video analysis, and confirm the ability of the Gibbsian probabilistic motion models to recognise distinct dynamic activities. At this level, the main difficulty seems to determine a pertinent classification structure of the dynamic content in a video, the trap being to be too ambitious and to look for semantic classes more than motion-related classes. The proposed classification method remains improvable while being flexible. First, no a priori knowledge on the classes has been considered yet, but it can be

easily introduced by considering a Maximum a Posteriori criterion instead of the Maximum Likelihood criterion. Moreover, only the residual image motion (related to scene motion) is exploited during the classification step. The dominant image motion (related to camera motion) could be reintroduced at this stage, these two sources of motion carrying important, different but complementary information. A first probabilistic model has been explored in [17] for characterizing the camera motion and improvements are currently at work. The potential of the approach could also be overpassed by considering not only motion information but other features such as dominant color and audio information, as investigated in [6]. The integration of this information within the proposed statistical framework is quite tractable as presented in [19] for instance. Investigations in this direction are currently in progress.

Acknowledgments

This research was partly supported by the French Ministery of Industry, in the context of the RNTL project "Domus Videum". The authors would like to thank INA, Direction de la Recherche for providing the videos *Athletics, Avengers* and *Albertville*. The work was carried out while N. Peyrard was at IRISA as INRIA post-doctorant.

References

1. E. Ardizzone and M. La Casia, "Video indexing using optical flow field," in Third IEEE International Conference on Image Processing, Lausanne, September 1996.
2. M. Basseville, "Detecting changes in signals and systems—a survey," Automatica, Vol. 24, No. 3, pp. 309–326, 1988.
3. 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 Databases IV, volume SPIE 2670, San Jose, January 1996, pp. 170–179.
4. 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, Vol. 9, No. 7, pp. 1030–1044, 1999.
5. E. Bruno and S. Marchand-Maillet, "Nonlinear temporal modeling for motion-based video overviewing," in Third International workshop on Content-Based Multimedia Indexing, CBMI'2003, Rennes, September 2003.
6. M. Christel, M. Smith, C. Taylor, and D. Winkler, "Evolving video skims into useful multimedia abstractions," in ACM Conference on Human Factors in Computing Systems, CHI'1998, Los Angeles, April 1998.
7. R. Fablet and P. Bouthemy, "Non parametric motion recognition using temporal multiscale Gibbs models," in IEEE Int. Conf. on Computer Vision and Pattern Recognition, Kauai, Hawaii, December 2001.
8. 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, Vol. 11, No. 4, pp. 393–407, 2002.
9. S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 6, pp. 721–741, 1984.
10. W. Liu, Z. Su, S. Li, Y. Sun, and H. Zhang, "Performance evaluation protocol for content-based image retrieval," in CVPR Workshop on Empirical Evaluation Methods in Computer Vision, Hawaii, December 2001.

11. Y.-F. Ma and H.-J. Zhang, "Motion pattern-based video classification retrieval," EURASIP Journal on Applied Signal Processing, Vol. 2, pp. 199–208, 2003.
12. Y.-F. Ma, H.-J. Zhang, and M. Li, "A user attention model for video summarization," in 10th ACM International Conference on Multimedia, Juan-Les-Pins, December 2002.
13. J. Meng and S.-F. Chang, "CVEPS—A compressed video editing and parsing system," in 4th ACM International Conference on Multimedia, Boston, November 1996.
14. J. Nam and H. Tewfik, "Dynamic video summarization and visualization," in 7th ACM International Conference on Multimedia, Orlando, November 1999, pp. 53–56.
15. J-M. Odobez and P. Bouthemy, "Robust multiresolution estimation of parametric motion models," Journal of Visual Communication and Image Representation, Vol. 6, No. 4, pp. 348–365, 1995.
16. N. Peyrard and P. Bouthemy, "Content-based video segmentation using statistical motion models," in British Machine Vision Conference BMVC'2002, Cardiff, September 2002.
17. N. Peyrard and P. Bouthemy, "Detection of meaningful events in videos based on a supervised classification approach," in International Conference on Image Processing, ICIP'2003, Barcelona, September 2003.
18. 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'2000, Hilton Head, SC, June 2000, Vol. 1, pp. 111–118.
19. J. Sánchez, X. Binefa, and J. Kender, "Coupled Markov chains for video contents characterization," in IEEE International Conference on Pattern Recognition, ICPR'2002, Quebec City, August 2002.
20. N. Vasconcelos and A. Lippman, "Statistical models of video structure for content analysis and characterization," IEEE Transactions on Image Processing, Vol. 9, No. 1, pp. 3–19, 2000.
21. J. Vermaak, P. Pérez, and M. Gangnet, "Rapid summarisation and browsing of video sequences," in British Machine Vision Conference, BMVC'2002, Cardiff, September 2002.
22. L. Zelnik-Manor and M. Irani, "Event-based video analysis," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR'2001, Kauai, Hawaii, December 2001, Vol. 2, pp. 123–130.