Extraction of Semantic Dynamic Content from Videos with Probabilistic Motion Models

Gwenaëlle Piriou[1], Patrick Bouthemy[1], and Jian-Feng Yao[1,2]

[1] IRISA/INRIA,
[2] IRMAR,
Campus universitaire de Beaulieu, 35042 Rennes cedex, France
{Gwenaelle.Piriou,Patrick.Bouthemy,Jian-Feng.Yao}@irisa.fr

Abstract. The exploitation of video data requires to extract information at a rather semantic level, and then, methods able to infer "concepts" from low-level video features. We adopt a statistical approach and we focus on motion information. Because of the diversity of dynamic video content (even for a given type of events), we have to design appropriate motion models and learn them from videos. We have defined original and parsimonious probabilistic motion models, both for the dominant image motion (camera motion) and the residual image motion (scene motion). These models are learnt off-line. Motion measurements include affine motion models to capture the camera motion, and local motion features for scene motion. The two-step event detection scheme consists in pre-selecting the video segments of potential interest, and then in recognizing the specified events among the pre-selected segments, the recognition being stated as a classification problem. We report accurate results on several sports videos.

1 Introduction and Related Work

Exploiting the tremendous amount of multimedia data, and specifically video data, requires to develop methods able to extract information at a rather semantic level. Video summarization, video retrieval or video surveillance are examples of applications. Inferring concepts from low-level video features is a highly challenging problem. The characteristics of a semantic event have to be expressed in terms of video primitives (color, texture, motion, shape ...) sufficiently discriminant w.r.t. content. This remains an open problem at the source of active research activities.

In [9], statistical models for components of the video structure are introduced to classify video sequences into different genres. The analysis of image motion is widely exploited for the segmentation of videos into meaningful units or for event recognition. Efficient motion characterization can be derived from the optical flow, as in [8] for human action change detection. In [11], the authors use very simple local spatio-temporal measurements, i.e., histograms of the spatial and temporal intensity gradients, to cluster temporal dynamic events. In [10], a principal component representation of activity parameters (such as translation,

T. Pajdla and J. Matas (Eds.): ECCV 2004, LNCS 3023, pp. 145–157, 2004.

rotation ...) learnt from a set of examples is introduced. The considered application was the recognition of particular human motions, assuming an initial segmentation of the body.

In [2], video abstraction relies on a measure of fidelity of a set of key-frames based on color descriptors and a measure of summarizability derived from MPEG-7 descriptors. In [6], spatio-temporal slices extracted in the volume formed by the image sequence are exploited both for clustering and retrieving video shots. Sport videos are receiving specific attention due to the economical importance of sport TV programs and to future services to be designed in that context. Different approaches have been recently investigated to detect highlights in sport videos. Dominant colour information is used in [3].

In this paper, we tackle the problem of inferring concepts from low-level video features and we follow a statistical approach involving modeling, (supervised) learning and classification issues. Such an attempt was recently undertaken for static images in [5]. We are dealing here with concepts related to events in videos, more precisely, to dynamic content. Therefore, we focus on motion information. Since no analytical motion models are available to account for the diversity of dynamic contents to be found in videos, we have to specify and learn them from the image data. To this end, we introduce new probabilistic motion models. Such a probabilistic modelling allows us to derive a parsimonious motion representation while coping with errors in the motion measurements and with variability in motion appearence for a given type of event. We handle in a distinct way the scene motion (i.e., the residual image motion) and the camera motion (i.e., the dominant image motion) since these two sources of motion bring important and complementary information. As for motion measurements, we consider, on one hand, parametric motion models to capture the camera motion, and on the other hand, local motion features to account for the scene motion. They convey more information than those used in [11], while still easily computable contrary to optic flow. They can be efficiently and reliabily computed in any video whatever its genre and its content.

We have designed a two-step event detection method to restrict the recognition issue to a limited and pertinent set of classes since probabilistic motion models have to be learnt for each class of event to be recognized. This allows us to simplify the learning stage, to save computation time and to make the overall detection more robust and efficient. The first step consists in selecting candidate segments of potential interest in the processed video. Typically, for sport videos, it involves to select the "play" segments. The second step handles the recognition of the relevant events (in terms of dynamic content) among the segments selected after the first step and is stated as a classification problem.

The remainder of the paper is organized as follows. In Section 2, we briefly present the motion measurements we use. Section 3 is concerned with the probabilistic models introduced to represent the dominant image motion and the residual motion. We describe in Section 4 the two-step event detection method, while the learning stage is detailed in Section 5. Experiments on sports videos are reported in Section 6, and Section 7 contains concluding remarks.

2 Motion Measurements

Let us first briefly describe the motion measurements that we use. Let us point out that the choice of these measurements is motivated by the goal we are pursuing, that is the recognition of important events in videos. This task is intended as a rather qualitative characterization which does not require a full estimation of the image motion.

It is possible to characterize the image motion as proposed in [4], by computing at each pixel a local weighted mean of the normal flow magnitude. However, the image motion is actually the sum of two motion sources: the dominant motion (which can be usually assumed to be due to camera motion) and the residual motion (which is then related to the independent moving objects in the scene, which will be referred to as the scene motion in the sequel). More information can be recovered if we explicitly consider these two motion components rather than the total motion only. Thus, we first compute the camera motion (more precisely, we estimate the dominant image motion) between successive images of the sequence. Then, we cancel the camera motion (i.e., we compensate for the estimated dominant image motion), which allows us to compute local motion-related measurements revealing the residual image motion only.

The dominant image motion is represented by a deterministic 2D affine motion model which is a usual choice:

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

where $\theta = (a_i, i = 1, \ldots, 6)$ is the model parameter vector and $p = (x, y)$ is an image point. This simple motion model can correctly handle different camera motions such as panning, zooming, tracking, (including of course static shots). Different methods are available to estimate such a motion model. We use the robust real-time multiresolution algorithm described in [7]. Let us point out that the motion model parameters are directly computed from the spatio-temporal derivatives of the intensity function. Thus, the camera-motion flow vector $\mathbf{w}_{\hat{\theta}_t}(p)$ is available at each time t and for each pixel p.

Then, the residual motion measurement $v_{res}(p,t)$ is defined as the local mean of the magnitude of normal residual flows weighted by the square of the norm of the spatial intensity gradient. The normal residual flow magnitude is given by the absolute value of the Displaced Frame Difference $DFD_{\hat{\theta}_t}$, evaluated with the estimated dominant motion, and divided by the norm of the image spatial gradient. We finally get:

$$v_{res}(p,t) = \frac{\sum_{q \in \mathcal{F}(p)} \|\nabla I(q,t)\| . |DFD_{\hat{\theta}_t}(q)|}{\max\left(\eta^2, \sum_{q \in \mathcal{F}(q)} \|\nabla I(q,t)\|^2\right)}, \tag{2}$$

where $DFD_{\hat{\theta}_t}(q) = I(q + \mathbf{w}_{\hat{\theta}_t}(q), t+1) - I(q,t)$. $\mathcal{F}(p)$ is a local spatial window centered in pixel p (typically a 3×3 window). $\nabla I(q,t)$ is the spatial intensity gradient of pixel q at time t. η^2 is a predetermined constant related to the noise

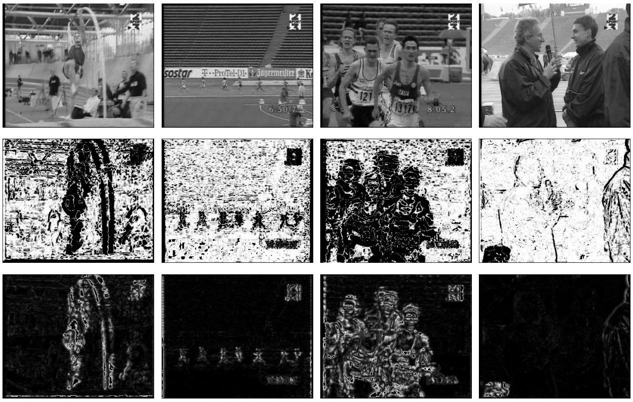

Fig. 1. *Athletics video:* First row: four images of the video. Second row: the corresponding maps of dominant image motion supports (inliers in white, outliers in black). Third row: local residual motion measurements v_{res} (zero-value in black).

level. Such measurements have already been used for instance for the detection of independent moving objects in case of a mobile camera. Figure 1 respectively displays images of an athletic TV program, the corresponding maps of dominant motion support (i.e., the points belonging to the image parts undergoing the estimated dominant motion) and the corresponding maps of residual motion measurements. This example shows that the camera motion is reliably captured even in case of multiple moving elements in the scene since the static background is correctly included in the inliers. It also indicates that the scene motion is correctly accounted for by the residual motion measurements. From relation (2), it can be straightforwardly noted that we only get information related to motion magnitude, and consequently, we lose the motion direction. As demonstrated by the results reported later, this is not a shortcoming since we aim at detecting events, i.e., at determining "qualitative" motion classes. Furthermore, it allows us to manipulate scalar measurements.

3 Probabilistic Modelling of Motion

The proposed method for the detection of important dynamic events relies on the probabilistic modelling of the motion content in a video. Indeed, the large diversity of video contents leads us to favor a probabilistic approach which moreover allows us to formulate the problem of event recognition within a Bayesian framework. Due to the different, nature of the information brought by the residual motion (scene motion) and by the dominant motion (camera motion), two different probabilistic models are defined.

3.1 Residual Motion

We first describe the probabilistic model of scene motion derived from statistics on the local residual motion measurements expressed by relation (2). The histograms of these measurements computed over different video segments were found to be similar to a zero-mean Gaussian distribution (a truncated version since we deal with positive values only, by definition $v_{res}(p,t) \geq 0$) except a usually prominent peak at zero. Therefore, we model the distribution of the local residual motion measurements within a video segment by a specific mixture model involving a truncated Gaussian distribution and a Dirac distribution. It can be written as:

$$f_{v_{res}}(\gamma) = \beta\delta_0(\gamma) + (1-\beta)\phi_t(\gamma; 0, \sigma^2)\mathbf{I}_{\gamma\neq 0}(\gamma), \tag{3}$$

where β is the mixture weight, δ_0 denotes the Dirac function at 0 ($\delta_0(\gamma) = 1$ if $\gamma = 0$ and $\delta_0(\gamma) = 0$ otherwise) and $\phi_t(\gamma; 0, \sigma^2)$ denotes the truncated Gaussian density function with mean 0 and variance σ^2. $\mathbf{I}$ denotes the indicator function ($\mathbf{I}_{\gamma\neq 0} = 1$ if $\gamma \neq 0$ and $\mathbf{I}_{\gamma\neq 0} = 0$ otherwise). Parameters β and σ^2 are estimated using the Maximum Likelihood criterion (ML). In order to capture not only the instantaneous motion information but also its temporal evolution over the video segment, the temporal contrasts Δv_{res} of the local residual motion measurements are also considered: $\Delta v_{res}(p,t) = v_{res}(p,t+1) - v_{res}(p,t)$. They are also modeled by a mixture model of a Dirac function at 0 and a zero-mean Gaussian distribution, but the Gaussian distribution is not truncated here. The mixture weight and the variance of the Gaussian distribution are again evaluated using the ML criterion.

The full probabilistic residual motion model is then defined as the product of these two models as follows: $P_{\mathcal{M}_{res}}(v_{res}, \Delta v_{res}) = P(v_{res}).P(\Delta v_{res})$ The probabilistic residual motion model is completely specified by four parameters only which are moreover easily computable. Obviously, this model does not allow us to capture how the motion information is spatially distributed in the image plane, but this is not necessary for the objective we consider here.

3.2 Dominant Image Motion

We have to design a probabilistic model of the camera motion to combine it with the probabilistic model of the residual motion in the recognition process. A first choice could be to characterize the camera motion by the motion parameter vector θ defined in Section 2 and to represent its distribution over the video segment by a probabilistic model. However, if the distribution of the two translation parameters a_1 and a_4 could be easily inferred (these two parameters are likely to be constant within a video segment so that a Gaussian mixture could reasonably be used, the task becomes more difficult when dealing with the other parameters which may be not constant anymore over a segment.

We propose instead to consider another mathematical representation of the estimated motion models, that is the camera-motion flow vectors and to consider the 2D histogram of these vectors. At each time t, the motion parameters θ_t of the camera motion model (1) are available and the vectors $\mathbf{w}_{\hat{\theta}_t}(p)$ can be computed at any point p of the image plane (we consider the points of the image grid in practice). The values of the horizontal and vertical components of $\mathbf{w}_{\hat{\theta}_t}(p)$ are then finely quantized, and we form the empirical 2D histogram of their distribution over the considered video segment. Finally, this histogram is represented by a mixture model of 2D Gaussian distributions. Let us point out that this modeling can involve several different global motions for events of the same type filmed in different ways. The number of components of the mixture is determined with the Integrated Completed Likelihood criterion (ICL, [1]) and the mixture model parameters are estimated using the Expectation-Maximisation (EM) algorithm.

4 Event Detection Algorithm

We now exploit the designed probabilistic models of motion content for the task of event detection in video. We have to learn the concepts of dynamic content to be involved in the event detection task.

We suppose that the videos to be processed are segmented into homogeneous temporal units. This preliminary step is out of the scope of this paper which focuses on the motion modelling, learning and recognition issues. To segment the video, we can use either a shot change detection technique or a motion-based temporal video segmentation method. Let $\{s_i\}_{i=1,\cdots,N}$ be the partition of the processed video into homogeneous temporal segments.

4.1 Selecting Video Segments

The first step of our event detection method permits to sort the video segments in two groups, the first group contains the segments likely to contain the relevant events, the second one is formed by the video segments to be definitively discarded. Typically, if we consider sport videos, we try to first distinguish between "play" and "no play" segments. This step is based only on the residual motion which accounts for the scene motion, therefore only single-variable probabilistic models are used, which saves computation. To this end, several motion models are learnt off-line in a training stage for each of the two groups of segments. This will be detailed in Section 5. We denote by $\{\mathcal{M}^{1,n}_{res}, 1 \leq n \leq N_1\}$ the residual motion models learnt for the "play" group and by $\{\mathcal{M}^{2,n}_{res}, 1 \leq n \leq N_2\}$ the residual motion models learnt for the "no play" group. Then, the sorting consists in assigning the label ζ_i, whose value can be 1 for "play" or 2 for "no play", to each segment s_i of the processed video using the ML criterion defined as follows:

$$\zeta_i = \arg\max_{k=1,2} \left[\max_{1 \leq n \leq N_k} P_{\mathcal{M}^{k,n}_{res}}(z_i) \right] \tag{4}$$

$z_i = \{v_{res\,i}, \mathit{\Delta} v_{res\,i}\}$ denote the local residual motion measurements and their temporal contrasts for the video segment s_i.

4.2 Detecting Relevant Events

Problem statement. The second step of the proposed method effectively deals with the detection of the events of interest within the previously selected segments. Contrary to the first step, the two kinds of motion information (scene motion and camera motion) are exploited, since their combination permits to more precisely characterize a specific event. For a given genre of video document, an off-line training stage is required to learn the dynamic content concepts involved in the event detection task. As explained in Section 5, a residual motion model $\mathcal{M}^j_{res}$ and a camera motion model $\mathcal{M}^j_{cam}$ have to be estimated from a given training set of video samples, for each event j to be retrieved. The detection is performed in two sub-steps. First, we assign to each pre-selected segment the label of one of the event classes introduced in the considered task. This issue is stated as a classification problem which avoids the need of detection thresholds and allows all the considered events to be extracted in a single process. Since false segments might be included in the pre-selected segments, a validation step is subsequently applied to confirm or not the assigned labels.

Video segment labeling. We consider only the segments s_i which have been selected as "play" segments after the first step described above. For each video segment s_i, $z_i = \{v_{res\ i}, \Delta v_{res\ i}\}$ are the residual motion measurements and their temporal contrasts, and w_i represent the motion vectors corresponding to the 2D affine motion models estimated between successive images over the video segment s_i.

The video segments are then labeled with one of the J learnt classes of dynamic events according to the ML criterion. More precisely, the label l_i assigned to the segment s_i takes its value in the label set $\{1, \ldots, J\}$ and is defined as follows :

$$l_i = \arg \max_{j=1,\ldots,J} P_{\mathcal{M}^j_{res}}(z_i) \times P_{\mathcal{M}^j_{cam}}(w_i) \tag{5}$$

Prior on the classes could be introduced in (5) leading to a MAP criterion.

Event label validation. By applying (5), we can label all the segments supplied by the first selection step. However, we have to consider that there might be "no play" segments wrongly labeled as "play" after the first selection step. We call them "intruders". These segments are forced to be assigned one of the event classes using relation (5), which creates false detection. As a consequence, we propose a validation test, involving only residual motion models. It consists in testing for each segment s_i the hypotheses defined by:

$$\begin{cases} H_0 : \text{"}s_i \text{ really belongs to the class } l_i \text{ determined by (5)"} \\ H_1 : \text{"}s_i \text{ is labeled as } l_i\text{, whereas it is an intruder segment"} \end{cases}$$

To this end, a set of models $\overline{\mathcal{M}}^j_{res}$ has to be specified and estimated to represent the intruder segments. This will be explained in Section 5.

The likelihood test to choose between this two hypotheses, is given by:

$$\text{if } \frac{P_{\mathcal{M}^j_{res}}(z_i)}{P_{\overline{\mathcal{M}^j_{res}}}(z_i)} < \varepsilon,\ H_1 \text{ is accepted ; else, } H_0 \text{ is accepted.}$$

In this way, we can correct some misclassifications resulting from the imperfect output of the first selection step, by discarding the video segments which are rejected by the likelihood test.

5 Learning the Dynamic Content Concepts

For a given video genre, a training step is performed off-line in order to learn the residual motion models and the dominant motion models needed by the event detection method. Let us note that we have to divide the training set in two sub-sets. The first one is used to learn the motion models required by steps 1 and 2 of the event detection algorithm, while the second one allows us to learn the intruder motion models.

Learning the residual motion models for the two-group selection step. As the first selection step involves the scene motion only, we have to learn residual motion models as specified in subsection 3.1. Because of the large diversity of video contents in the two groups “play” and “no play”, we have to represent each group by several motion models. We apply the ascendant hierarchical classification (AHC) technique, on one hand, to the “play” group, and on the other hand, to the “no play” group of the training set. The overall procedure is defined as follows.
Step 0: A residual motion model is estimated for each video segment belonging to the training set of the considered group. At this early stage, each segment forms a cluster. *Step 1:* The two clusters (either composed of one segment or of several segments) found as the nearest w.r.t the symmetrized Kullback-Leibler distance between their corresponding residual motion models, are merged in the same cluster. The expression of this distance between two residual motion models $\mathcal{M}^1_{res}$ and $\mathcal{M}^2_{res}$ is $d(\mathcal{M}^1_{res}, \mathcal{M}^2_{res}) = \frac{1}{2}(d_K(\mathcal{M}^1_{res}, \mathcal{M}^2_{res}) + d_K(\mathcal{M}^2_{res}, \mathcal{M}^1_{res}))$, where $d_K(\mathcal{M}^1_{res}, \mathcal{M}^2_{res}) = d_K(f^1_{v_{res}}, f^2_{v_{res}}) + d_K(f^1_{\Delta v_{res}}, f^2_{\Delta v_{res}})$. The expression of the Kullback-Leibler distance between the density functions $f^1_{v_{res}}$ with parameters (β_1, σ_1), and $f^2_{v_{res}}$ with parameters (β_2, σ_2), of the residual motion measurements is given by:

$$d_K(f^1_{v_{res}}, f^2_{v_{res}}) = \beta_1 ln\left(\frac{\beta_1}{\beta_2}\right) + (1-\beta_1) ln\left(\frac{\sigma_2(1-\beta_1)}{\sigma_1(1-\beta_2)}\right) + \frac{1-\beta_1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} - 1\right).$$

The Kullback-Leibler distance between the density functions $f^1_{\Delta v_{res}}$ and $f^2_{\Delta v_{res}}$ of the temporal contrasts can be similarly written. A residual motion model is then estimated for the obtained new cluster. We iterate until the stopping criterion is satisfied. *Stopping criterion:* We stop if the maximum of the symmetrized Kullback-Leibler distances between two clusters is lower than a certain percentage of the maximum of the distances computed at step 0.

At this stage, the load of manually labelling the video segments of the training set is kept low. Indeed, we just need to sort the video segments into the two groups “play” and “no play”. At the end, each group is represented by a (small) set of clusters (depending on the heterogeneity of the video segment contents of the group) and their associated residual motion models, both obtained in an automatic way.

Learning the motion models of the event classes. Camera motion models and residual motion models representing the different event classes to be recognized are required for the second step of our detection method. They are estimated from the same training set as the one used to learn residual motion models involved in the selection step. We first need a manual labelling of the “play” segments of the training set according to the events to detect. For each event class, a camera motion model is estimated from the video segments representing the considered event as explained at the end of subsection 3.2. Similarly, the four parameters of the residual motion models for each event class are estimated using the ML criterion.

Learning of intruder motion models. We have also to determine motion models, from the second subset of the training set, to represent the intruder segments. It is important to consider a different set of video segments than the one used to learn the models involved in the first steps of the detection method. The selection step is applied to the second subset of the training set. The intruder segments are then determined (since we have the ground truth on that training set) and submitted to the classification step of the method. Finally, we get a subset of intruder segments associated to each predefined event j, which allows us to estimate the associated residual motion model previously denoted by $\overline{\mathcal{M}}^j_{res}$.

6 Experimental Results

We have applied the described method on sports videos which involve complex contents while being easily specified. Moreover, events or highlights can be naturally related to motion information in that context. We report here results obtained on athletics and tennis videos.

6.1 Experimental Comparison

First, we have carried out an experimental comparison between our statistical approach and a histogram-based technique. In order to evaluate the probabilistic framework we have designed, we consider the same motion measurements for the histogram technique. Thus, the latter involves three histograms: the histogram of residual motion measurements v_{res} (2), the histogram of their temporal contrasts Δv_{res}, and the 2D histogram of the camera-motion flow vectors (subsection 3.2). Each event j is then represented by three histograms: $H^j_{v_{res}}$, $H^j_{\Delta v_{res}}$ and H^j_{cam}.

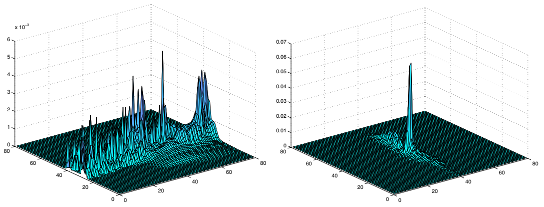

Fig. 2. *Athletics video:* 2D histograms of the camera-motion flow vectors. Left: for a pole vault shot, right: for a long-shot of track race.

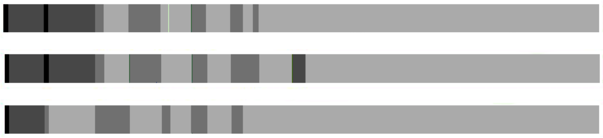

Fig. 3. *Athletics video:* Detection of relevant events: Top row: ground-truth, middle row: results obtained with the probabilistic motion models, bottom row: results obtained with the histogram-based technique. From dark to light shining: pole vault, replay of pole vault, long-shot of track race and close-up of track-race

To compare two histograms, we consider the Euclidian distance, denoted by d_1 for 1D histograms and by d_2 for 2D histograms. Several distances can be considered to compare two histograms, and this issue has to be carefully addressed. However, the computed motion measurements are all real values and we have a huge number of available computed values. We can thus consider a very fine quantization and the resulting histograms are very close to the real continuous distributions. Moreover, the histogram distance is only used to rank the classes. The Euclidean distance is then a reasonable choicewhile easy to compute. These histograms are computed (and stored) for each event j from the training set of video samples. Then, we consider the test set and we compute the three histograms $H^{s_i}_{v_{res}}, H^{s_i}_{\Delta v_{res}}$ and $H^{s_i}_{cam}$, for each video segment s_i of the test set. The classification step is now formulated as assigning the label l_i of the event which minimizes the sum of the distances between histograms:

$$l_i = \arg \min_{j=1,\dots,J} \left(d_1(H^{s_i}_{v_{res}}, H^{j}_{v_{res}}) + d_1(H^{s_i}_{\Delta v_{res}}, H^{j}_{\Delta v_{res}}) + d_2(H^{s_i}_{cam}, H^{j}_{cam}) \right) \tag{6}$$

In order to focus on the classification performance of the two methods, the test set only involves "play" segments. We have processed a part of an athletics TV program which includes jump events and track race shots. The training set is formed by 12500 images and the test set comprises 7800 images. Some representative images of this video are displayed on Figure 1. We want to recognize four events: Pole vault, Replay of pole vault, Long-shots of track race and Close-up of track race. Consequently, we have to learn four residual motion models and four camera motion models for the method based on the probabilistic motion modelling. Figure 2 contains the 2D histograms of the camera-motion flow vectors for two classes. In Figure 3, the processed video is represented by a time line exhibiting the duration of the video segments. The “no play” segments have been in fact withdrawn, and the “play” segments have been concatenated to form the time line. A grey level is associated to each event class. The first row corresponds to the ground truth, the second one and the third one contain the results obtained respectively using the probabilistic motion models and using the histogram technique. These results demonstrate that the statistical framework yields quite satisfactory results and outperforms the histogram-based technique.

6.2 Event Detection Method

We have applied our event detection method to a tennis TV program. The first 42 minutes (63000 images) of the video are used as the training set (22 minutes for the learning of the motion models involved in the two first steps and 20 minutes for the learning of intruder models), and the last 15 minutes (18000 images) form the test set.

Selecting video segments. We want to distinguish between “play” segments involving the two tennis players in action and the “no play” segments including views of the audience, referee shots or shots of the players resting, as illustrated in Figure 4. We only exploit the first subset of the training set to learn the residual motion models that we need for the selection step. 205 video segments of the training set represent “play” segments and 95 are “no play” segments. 31 residual motion clusters and their associated models are supplied by the AHC algorithm for the “play” group, and 9 for the “no play” group. The high number of clusters obtained reveals the diversity of dynamic contents in the two groups of the processed video. Quite satisfactory results are obtained, since the precision rate for the play group is 0.88 and the recall rate is 0.89.

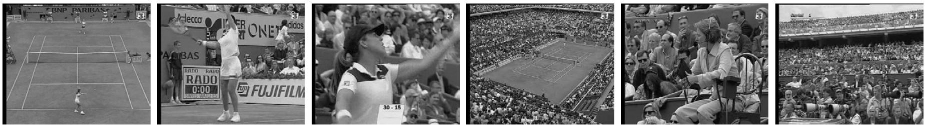

Fig. 4. *Tennis video:* Three image samples extracted from the group of “play” segments and three image samples extracted from the group of “no play” segments.

Table 1. *Tennis video:* Results of the event detection method based on probabilistic motion models (P: precision, R: recall).

	Rally	Serve	Change of side
P	0.92	0.63	0.85
R	0.89	0.77	0.74

Detecting relevant events. The goal is now to detect the relevant events of the tennis video among the segments selected as “play” segments. For this second step, we introduce the probabilistic camera motion model. The three events we try to detect are the following: Rally, Serve and Change of side. In practice, we consider two sub-classes for the Serve class, which are wide-shot of serve and close-up of serve. Two sub-classes are considered too for the Change-of-side class. As a consequence, five residual motion models and five camera motion models have to be learnt. We have also to determine the residual motion models accounting for the intruder segments for each class. The obtained results are reported in Table 1. Satisfactory results are obtained specially for the rally class. The precision of the serve class is lower than the others. In fact, for the serve class, errors come from the selection step (i.e., some serve segments are wrongly put in the “no play” group, and then, are lost). It appears that a few serve segments are difficult to distinguish from some “no play” segments when using only motion information. However, the proposed statistical framework can easily integrate other information such as color or audio.

7 Conclusion

We have addressed the issue of determining dynamic content concepts from low-level video features with the view to detecting meaningful events in video. We have focused on motion information and designed an original and efficient statistical method. We have introduced new probabilistic motion models representing the scene motion and the camera motion. They can be easily computed from the image sequence and can handle a large variety of dynamic video contents. We have demonstrated that the considered statistical framework outperforms a histogram-based technique. Moreover, it is flexible enough to properly introduce prior on the classes if available, or to incorporate other kinds of video primitives (such as color or audio). The proposed two-step method for event detection is general and does not exploit very specific knowledge (e.g. related to the type of sport) and dedicated solutions. Satisfactory results on sports videos have been reported.

Acknowledgments. This research was supported by “Région Bretagne” (PhD thesis grant) and by the French Ministery of Industry (RNTL Domus Videum project). The authors would like to thank INA, Direction de la Recherche, for providing the videos.

References

1. C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the Integrated Completed Likelihood. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(3):719–725, 2000.
2. A. Divakaran, R. Radhakrishnan, and K.A. Peker. Motion activity-based extraction of key-frame from video shots. *ICIP’02*, Rochester, Sept. 2002.

3. A. Ekin, A.M. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Int. Trans. on Image Processing*, 12(7):796–807, July 2003.
4. R. Fablet, P. Bouthemy, and P. Pérez. Non-parametric motion characterization using causal probabilistic models for video indexing and retrieval. *IEEE Trans. on Image Processing*, 11(4):393–407, 2002.
5. J. Li and J.Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. on PAMI*, 25(9):1075–1088, Sept. 2003.
6. C-W. Ngo, T-C. Pong, and H-J. Zhang. On clustering and retrieval of video shots through temporal slices analysis. *IEEE Trans. Multimedia*, 4(4):446–458, Dec. 2002.
7. J-M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *J. of Visual Comm. and Image Repr.*, 6(4):348–365, Dec. 1995.
8. Y. Rui and P. Anandan. Segmenting visual actions based on spatio-temporal motion patterns. *CVPR'2000*, Hilton Head, SC, 2000.
9. N. Vasconcelos and A. Lippman. Statistical models of video structure for content analysis and characterization. *IEEE Trans. on IP*, 9(1):3–19, Jan. 2000.
10. Y. Yacoob and J. Black. Parametrized modeling and recognition of activities. *Sixth IEEE Int. Conf. on Computer Vision*, Bombay, India, 1998.
11. L. Zelnik-Manor and M. Irani. Event-based video analysis. *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Kauai, Hawaii, Dec. 2001.