Learned probabilistic image motion models for event detection in videos

Gwenaëlle Piriou [1], Patrick Bouthemy [1], Jian-Feng Yao [1,2]
[1] IRISA/INRIA, [2] IRMAR
Campus universitaire de Beaulieu,
35042 Rennes cedex, France

Abstract

We present new probabilistic motion models of interest for the detection of relevant dynamic contents (or events) in videos. We separately handle the dominant image motion assumed to be due to the camera motion and the residual image motion related to scene motion. These two motion components are then represented by different probabilistic models which are further recombined for the event detection task. The motion models associated to pre-identified classes of meaningful events are learned from a training set of video samples. The event detection scheme proceeds in two steps which exploit different kinds of information and allow us to progressively select the video segments of interest using Maximum Likelihood (ML) criteria. The efficiency of the proposed approach is demonstrated on sports videos.

1. Introduction

Approaching the "semantic" content of video documents while dealing with physical image signals and numerical measurements is a high challenge in computer vision. The characteristics of a semantic event have to be expressed in terms of low-level video primitives which have to be sufficiently discriminant.
Different kinds of video features have already been considered in several approaches. In [7] statistical models are introduced for components of the video structure to classify video sequences into different genres. Recently, in [5], a semantic classification method based on SVM ("Support Vector Machine") using a motion pattern descriptor has been described. In [9], 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 [8], a principal component representation of activity parameters (such as translation, rotation ...) learned from a set of examples is introduced. The considered application was the recognition of particular human motions, assuming an initial segmentation of the body.
Different approaches have been recently investigated to detect highlights in sports videos. In [2], the authors deal with soccer videos and mainly exploit dominant colour information. In [4], the spectator excitement is modeled and derived from three variables related to audio-video features in order to extract the most interesting video segments.
In this paper, we focus on motion information and we propose new probabilistic image motion models useful for the detection of dynamic events. The motion information is captured through low-level motion measurements easily computable in any video. Our approach consists in separately modeling the camera motion (i.e., the dominant image motion) and the scene motion (i.e., the residual image motion). These two sources of motion bring important and complementary information.
We apply the designed statistical framework to the detection of relevant events in a video following a two-step approach. The first step consists of a pre-selection of candidate video segments. The second step is a classification stage to recognize the relevant events (in terms of dynamic content) among the pre-selected segments. Such a two-step process allows us to save computation time and to make the overall detection more robust and efficient.
The paper is organized as follows. Section 2 briefly presents the motion measurements we used. Sections 3 and 4 describe the statistical modeling of scene motion and of camera motion respectively. Section 5 is concerned with dynamic event detection. Experiments on sports videos are reported in Section 6. Section 7 contains the conclusion.

2. Motion measurements

A probabilistic modeling of the motion content of a video enables to derive a parsimonious motion representation while coping with errors in the motion measurements and with variability in a given type of motion content. Furthermore, no analytical motion models are available to account for the variety of dynamic contents to be found in videos. We have then to specify and learn motion models from the image data. Let us also note that we aim at rec-

ognizing "broad" event classes and not particular "quantitative" motions. It is possible to characterize the full image motion as proposed in [3], 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 components: the dominant motion (usually due to camera motion) and the residual motion (related to the scene motion). We believe that more information can be recovered when dealing with these two motions separately rather than with the total motion only.
The dominant image motion can be represented by a deterministic 2D affine motion model (which is a usual choice):

$$w_\theta(p) = (a_1 + a_2 x + a_3 y\,,\, a_4 + a_5 x + a_6 y)^T\,, \qquad (1)$$

where $\theta = (a_i, i = 1, \ldots, 6)$ is the model parameter vector and $p = (x, y)$ is an image point. This motion model can 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 multi-resolution algorithm described in [6]. Let us point out that the motion model parameters are directly computed from the spatio-temporal derivatives of the intensity function. Then, the corresponding motion vector $w_{\hat{\theta}_t}(p)$ is available at any pixel p and time t.
The residual motion measurement v_{res} is defined as the local mean of normal residual flow magnitudes $|v_n|$ (weighted by the square of the norm of the spatial intensity gradient):

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

with $v_n(q,t) = \frac{I(q,t) - I(q + w_{\hat{\theta}_t}(q), t+1)}{\|\nabla I(q,t)\|}$. $\mathcal{F}(p)$ is a local spatial window centered in pixel p. ∇I is the spatial intensity gradient. η^2 is a constant related to the noise level.
Figure 1 displays two images of an athletics TV program with the maps of the estimated dominant motion vectors and the maps of residual motion measurements v_{res}.

3. Probabilistic model of scene motion

We describe now the probabilistic model of scene motion derived from statistics on the local residual motion measurements expressed by relation (2). The 1D histograms of these measurements computed over different video segments show two degrees of freedom. In fact, they present usually a prominent peak at zero and a continuous component part which can be modeled by a distribution of the exponential family. We have opted here for the exponential distribution to represent the continuous part. Therefore, we model the distribution of the local residual motion measurements within a video segment by a specific mixture model with density:

$$f(z) = \alpha\delta_0(z) + (1-\alpha)\phi_\beta(z), \qquad (3)$$

where z holds for $v_{res}(p,t)$, α is the mixture weight, δ_0 denotes the Dirac function at 0 and $\phi_\beta(z) = \beta e^{-\beta z}$ for $z > 0$. α and β are estimated using the ML criterion.
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:

$$\Delta v_{res}(p,t) = v_{res}(p,t+1) - v_{res}(p,t). \qquad (4)$$

They are modeled by a mixture $g(z')$ of a Dirac function at 0 and a zero-mean Gaussian distribution, where z' holds for $\Delta v_{res}(p,t)$. The mixture weight and the variance of the Gaussian distribution are still evaluated using the ML criterion. The density of the full probabilistic residual motion model is then simply defined as:

$$h^{res}(z,z') = f(z).g(z'). \qquad (5)$$

The probabilistic residual motion model is completely specified by four parameters which can be easily estimated. It accounts for global statistics accumulated over both the image grid and time (i.e., over all the frames of the video segment). It can be considered as a global occurrence model.

4. Probabilistic model of camera 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. We first investigated to characterize the camera motion directly by the parameter vector θ defined in Section 2 and to represent its distribution over the sequence by a probabilistic model. However, it was difficult to design a convenient probabilistic model. We propose instead to consider an equivalent representation by the 2D motion vectors $w_{\theta_t}(p)$, and to exploit them as a 2D histogram. More precisely, at each time

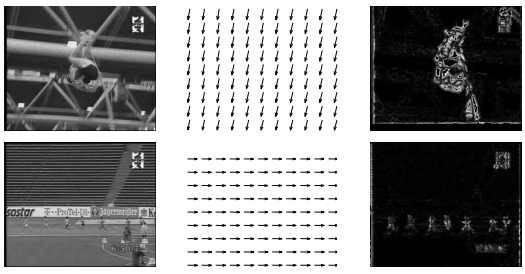

Figure 1. Two images at different time instants (involving an upward-tilt camera motion and a left panning one) of the Athletics video (provided by INA) and their corresponding maps of the estimated dominant motion fields and of residual motion measurements v_{res} (zero-value in black).

t, the motion parameters θ_t of the camera motion model (1) are estimated and the vectors $w_{\hat{\theta}_t}(p)$ are computed for each point p of the image support. The values of the horizontal and vertical components of $w_{\hat{\theta}_t}(p)$ 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 γ^{cam} of 2D Gaussian distributions. The number of components of the mixture is determined with the Integrated Completed Likelihood criterion (ICL, [1]) and their parameters are estimated using the Expectation-Maximization algorithm (EM).

5. Event detection algorithm

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. To segment the video, we can use either a shot change detection technique or a motion-based video segmentation method.

The first step of the event detection algorithm 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 sports 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 1D models are used which saves computation. To this end, a set of residual motion models (due to the content diversity) is learned off-line for each group of segments in an unsupervised way using an ascendant hierarchical classification technique. Then, the sorting consists in assigning the label "play" or "no play" to each segment of the processed video using the ML criterion.

The second step of the proposed scheme consists in retrieving several specific events among the previously selected segments. Contrary to the first step, the two kinds of motion information (residual and camera motion) are required since the combination allows us to characterize more precisely a specific event. An off-line training step is again required. A residual motion model with density h_j^{res} and a camera motion model with density γ_j^{cam} have to be estimated, from a training set of video samples, for each type j of event to detect (thanks to the initial sorting step, we can now restrict these modeling and learning steps to the event classes of interest only). Let $\{s_0, \ldots, s_N\}$ be the previously selected video segments. The video segments retained after the first step are labeled with one of the J learned models of dynamic events according to the ML criterion. Let $L_j(s_i)$ be the likelihood of the segment s_i related to the learned models for the event j:

$$L_j(s_i) = \prod_{(p,t)\in s_i} h_j^{res}(z_{(p,t)}, z'_{(p,t)}). \prod_{(p,t)\in s_i} \gamma_j^{cam}(w_{\hat{\theta}_t}(p)) \quad (6)$$

Figure 2. Image samples extracted from the skating video (top) and from the tennis video (bottom). (Videos provided by INA).

	Athletics	Skating	Tennis
Training set	15000	34500	63000
Test set	10500	13500	18000

Table 1. Number of images in the training set and in the test set of the processed videos (for the first step).

Thus, the label ξ_i of the segment s_i is defined as:

$$\xi_i = \arg \max_{j=1,\ldots,J} L_j(s_i) \quad (7)$$

6. Experimental results

6.1. Pre-selecting video segments

The first processed video is an athletics TV program. The "play" segments are formed by jump events and track race shots and the "no play" segments contain interview shots and large views of the stadium. The second video is a figure skating (dance) TV program. We want here to distinguish between "play" segments which correspond to skating (simple skating motion, artistic effects, dance movements) and "no play" segments involving low-level activity (views of the audience, static shots involving skaters waving at the audience or skaters waiting for the scores). The last video is a tennis TV program. The "play" segments involve the two tennis players in action and the "no play" segments include views of the audience, referee shots or shots of the players resting. Figure 1 and 2 display image samples of these videos and Table 1 gives the number of images which form the training set and the test set for each video. Let us

Video genre	Athletics	Skating	Tennis
P	0.95	0.95	0.86
R	0.95	0.93	0.89

Table 2. Results of the first step of the event detection method for the three processed videos. P: Precision for the play group. R: recall for the play group.

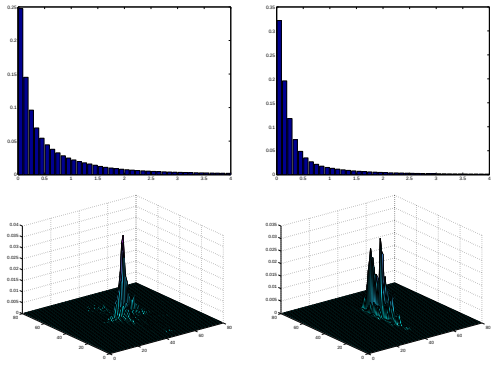

Figure 3. *Tennis video:* Histograms of the local residual motion measurements (top) and 2D histograms of the dominant motion flow vectors (bottom). Left: close-up of serve, right: wide shot of change of side

point out that the training set is formed by the first part of the video while the test set is formed by the last part. For the three considered videos, each group ("play", "no play") is represented by several residual motion models. As shown in Table 2, quite satisfactory results are obtained for the three processed videos.

6.2. Detecting relevant events

The aim is now to detect the relevant events among the segments selected as "play" segments. Due to page limitation, we will focus on the tennis example. 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 for the Change-of-side class too. As a consequence, five residual motion models and five camera motion models have to be learnt. Figure 3 displays 1D histograms of the local residual motion measurements and 2D histograms of the dominant motion flow vectors for two classes. The obtained results of the event detection method are reported in Table 3. Good results are obtained, especially for the rally class. The precision for 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.

	Rally	Serve	Change of side
P	0.91	0.56	0.84
R	0.90	0.69	0.70

Table 3. *Tennis video:* Results of the event detection method (P: precision, R: recall).

7. Conclusion

We have introduced new probabilistic motion models which can be easily learned and computed from the video data. They can handle a large variety of dynamic video contents. We explicitly take into account the information related to the scene motion and to the camera motion respectively. These motion models were proven to be efficient and appropriate for event detection in videos. The proposed method induces a low computation time, and accurate results on sport videos have been reported.
In the same time, due to the designed statistical framework, it is flexible enough to properly introduce prior on the classes if available, or to incorporate other useful information such as colour (e.g., the dominant colour to recognize the presence of the play field or the tennis court), or audio features. Such developments are indeed in progress for a video summarization application.

Acknowledgments

This research was partly supported by the "Région Bretagne" and by the IST European project LAVA.

References

[1] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the Integrated Completed Likelihood. *IEEE Trans. on PAMI*, 22(3):719–725, 2000.
[2] A. Ekin, A. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Int. Trans. on Image Processing*, 12(7):796–807, July 2003.
[3] 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.
[4] A. Hanjalic. Generic approach to highlights extraction from a sport video. *IEEE Int. Conf. on Image Processing, ICIP'03*, Barcelona, September 2003.
[5] Y.-F. Ma and H.-J. Zhang. Motion pattern-based video classification retrieval. *EURASIP Journal on Applied Signal Processing*, 2:199–208, March 2003.
[6] 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.
[7] N. Vasconcelos and A. Lippman. Statistical models of video structure for content analysis and characterization. *IEEE Trans. on Image Processing*, 9(1):3–19, January 2000.
[8] Y. Yacoob and J. Black. Parametrized modeling and recognition of activities. *Sixth IEEE Int. Conf. on Computer Vision, Bombay, India*, pages 120–127, 1998.
[9] L. Zelnik-Manor and M. Irani. Event-based video analysis. *IEEE Int. Conf. on Computer Vision and Pattern Recognition, Kauai, Hawaii*, 2:123–130, December 2001.