Probabilistic models of image motion for recognition of dynamic content in video

G. Piriou, P. Bouthemy, N. Peyrard
IRISA / INRIA
Campus de Beaulieu
35042 Rennes cedex, France

J.-F. Yao
IRMAR & IRISA/INRIA
Campus de Beaulieu
35042 Rennes cedex, France

Abstract

We present new probabilistic motion models of interest for the detection of meaningful 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. Two solutions are investigated for the residual motion. The motion models (both for camera motion and scene motion) associated to pre-identified classes of meaningful events are learned from a training set of video samples. The 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 sport videos.

1 Introduction

One of the actual challenges in computer vision is to somehow approach the "semantic" content of video documents while dealing with physical image signals and numerical measurements. One objective can be to handle tasks such as video summarization, video retrieval or video surveillance. The main difficulty lies in the detection of "semantic concepts" from low-level features. The characteristics of a semantic event has 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.

Different kinds of video features have already been considered in several approaches. In [1], the pixel chrominance components in the image are used to select keyframes maximally distinct and carrying the most information. [2] introduces statistical models for components of the video structure to classify video sequences into different genres (sports, news, movies, commercials, documentaries, ...). Recently, in [3], a semantic classification method based on SVM ("Support Vector Machine") using a motion pattern descriptor has been described.

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 [4] for human action change detection. In [5], 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 [6], 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. In [7], two simple low-level motion features are used to characterize the activity level of video sequences.

In this paper, we propose new probabilistic motion models of particular interest for the detection of meaningful dynamic events. The motion information is captured through low-level motion measurements which convey more elaborated motion information than those used in [5], while still locally computable contrary to optic flow. They can be efficiently and reliabily computed in any video whatever its genre and its content. Our approach consists in handling separately the scene motion (i.e., the residual image motion) and the camera motion (i.e., the dominant image motion) in a sequence. Indeed, these two sources of motion bring important, different but complementary, information which have to be taken into account for event detection or classification.

We have investigated two different probabilistic motion models to specify the residual motion information. With the first model, we evaluate temporal cooccurrences of the local motion-related measurements which are first quantized. The resulting cooccurrence matrix is then viewed as a 2D histogram and represented by a 2D Gaussian mixture model. On the other hand, the second model directly exploits the local motion measurements, along with their temporal contrasts in order to capture the temporal motion evolution. Their histograms are computed over the video segment and are represented by a specific mixture model.

An original probabilistic model is also proposed to cope with the camera motion. It exploits 2D histograms of velocity vectors issued from the estimated affine motion models accounting for the image dominant motion.

We apply this 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 segments among the successive segments of the processed video. It involves two pre-learned groups representing respectively "possibly important dynamic content" and " definitively not important dynamic content". The second step is a classification stage to recognize the relevant events (in terms of dynamic content) among the segments selected after the first step. Such a two-step process allows us to restrict the recognition issue on a limited and pertinent set of classes, to save computation time and to make the overall detection more robust and efficient.

The paper is organized as follows. In Section 2, we briefly present the motion measurements used. Section 3 describes the two statistical modelings of scene motion in a video that we have explored. Section 4 is dedicated to the probabilistic model for the camera motion. We present in Section 5 the full scheme for dynamic event detection. Experiments on sport videos are reported in Section 6 and Section 7 contains concluding remarks.

2 Motion Measurements

As stating above, we are investigating the probabilistic modeling of the motion content of a video. Such a modeling enables to derive a parsimonious motion representation while coping with errors in the motion measurements and with variability in a given kind 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 to specify and learn them from the image data. Let us also stress that we aim at recognizing "broad" event classes and not particular "quantitative" motions. The proposed framework therefore exploits only low-level motion features for generality and efficiency purposes. Although the motion estimation step is not the purpose of this paper, we have first to briefly describe the motion measurements that we use.

It is possible to characterize the global image motion as proposed in [8], by computing at each pixel a local weighted mean of normal flow magnitude. However, the image motion is actually the sum of two motion sources: the dominant motion (usually assumed to be 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 only with the total motion. Thus, we compensate for the camera motion (more precisely, we cancel the estimated dominant image motion) in the sequence of images, in order to compute local motion-related measurements revealing only the residual image motion.

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

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

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 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 [9]. 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 vector $\mathbf{w}_{\hat{\theta}_\mathbf{t}}(\mathbf{p})$ is available at each time t and for each pixel p. Then, the local motion-related measurement $v_{res}(p,t)$ is defined as the local mean of normal residual flows weighted by the square of the norm of the spatial intensity gradient. The normal residual flows are computed from the Displaced Frame Difference ($DFD_{\hat{\theta}_t}$) given by the estimated dominant motion. We finally get:

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

where $DFD_{\hat{\theta}_t}(q) = I(q + \mathbf{w}_{\hat{\theta}_\mathbf{t}}(\mathbf{q}), \mathbf{t+1}) - \mathbf{I}(\mathbf{q}, \mathbf{t})$. $\mathcal{F}(p)$ is a local spatial window centered in pixel p. $\nabla I(q,t)$ is the spatial intensity gradient of pixel q at time t. η^2 is a predetermined constant related to the noise level. Such measurements have already been used for instance for the detection of independent moving objects in case of a mobile camera. Figure 2 displays three images of an athletics TV program, the corresponding maps of dominant motion support and the corresponding maps of local motion-related measurements v_{res}.

3 Probabilistic models of scene motion

We have explored two types of statistics derived from the motion measurements defined in (2). The first one exploits temporal cooccurrences of the motion measurements which have to be first quantized. This choice enables to capture not only the motion magnitude but also the global temporal evolution of the motion magnitude. The computation of the cooccurrences being expensive in practice, we have defined an alternative approach where we directly consider the local motion-related measurements and their temporal gradients (contrasts). These two options for scene motion characterization and the respective designed probabilistic motion models are presented in the following two subsections.

3.1 GMM from temporal cooccurrences

With the first approach, the measures defined by (2) are quantized on a set Λ, so that for a video segment of length T and of spatial image support $\mathcal{R}$, the motion content is represented by the set of quantized local motion measurements $y = \{y(p,t), p \in \mathcal{R}, t = 1 \dots T\}$. The temporal cooccurrences distribution $\Gamma(y)$ of these quantities is a matrix $\{\Gamma(\nu,\nu'|y)\}_{(\nu,\nu')\in\Lambda^2}$ which is defined as follows:

$$\Gamma(\nu,\nu'|y) = \sum_{t=1}^{T-1} \sum_{p\in\mathcal{R}} \delta(\nu, y(p,t)) \cdot \delta(\nu', y(p,t+1)),$$

where $\delta(i,j)$ is the Kronecker symbol (equal to 1 if $i = j$ and to zero otherwise). The temporal cooccurrences matrix $\Gamma(y)$ is then considered as a 2D empirical histogram, and we model it by a 2D Gaussian mixture model (GMM). The log-likelihood of the sequence y is thus given by:

$$\ln P_{cooc.}(y) = \textstyle\sum_{(\nu,\nu')\in\Lambda^2} \Gamma(\nu,\nu'|y) \ln q(\nu,\nu') \quad (3)$$
$$\text{with } q(\nu,\nu') = \textstyle\sum_{k=1}^{K} \pi_k \phi(\nu,\nu'; m_k, \Sigma_k),$$

where K is the number of components in the mixture model and $\phi(\nu,\nu'; m_k, \Sigma_k)$ is the 2D Gaussian density function with mean vector m_k and covariance matrix Σ_k. The number of components K is determined with the Integrated Completed Likelihood criterion (ICL, [10]), and the Maximum Likelihood (ML) estimate of the model parameters is approximated using the Expectation-Maximisation algorithm.

3.2 DGMM from temporal contrasts

As an alternative to the computation of temporal cooccurrences, the temporal contrasts Δv_{res} of local motion-related measurements are considered. The contrasts are defined as the temporal difference of the variables v_{res} given by (2):

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

We have computed the histograms of these expressions over different video segments and it has been found to be quite similar to a Gaussian distribution except a usually prominent peak at zero. Therefore, we model the temporal contrasts distribution by a specific mixture model with density:

$$P_{\Delta v_{res}}(\gamma) = \beta\delta_0(\gamma) + (1-\beta)\phi(\gamma; 0, \sigma^2)\mathbf{1}_{\gamma\neq 0} \quad (5)$$

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 $\phi(\gamma; 0, \sigma^2)$ is the Gaussian density function with mean 0 and variance σ^2. The parameters β and σ^2 are estimated using the Maximum Likelihood criterion. Nevertheless, if we consider only the temporal contrasts Δv_{res}, the absolute motion magnitude would be lost. Consequently, the addition of the local motion-related measurements is required. 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 here truncated to take into account only the positive values since by definition $v_{res}(p,t) \geq 0$. The mixture weight and the variance of the truncated Gaussian distribution are evaluated using the ML criterion. The global probabilistic residual motion model is then defined as the product of the two described models as follows :

$$P_{contr.} = P_{\Delta v_{res}} . P_{v_{res}} \quad (6)$$

Let us notice that in that case we do not need to quantize v_{res} and Δv_{res}, and we directly deal with the computed real values. Naturally, this model does not allow us to capture how the motion information is spatially distributed, but it is not necessary for the objective we consider here. In the sequel, this Dirac Gaussian mixture model will be referred to as DGMM.

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. It could be possible to characterize directly the camera motion by the parameter vector θ defined in Section 2 and to represent its distribution over the sequence by a probabilistic model. The main difficulty in that case is to propose a valid probabilistic model. Indeed, if the distribution of the two translation parameters a_1 and a_4 can be easily described (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 are not constant anymore and which are not of the same nature. For this reason, we propose to build the map of the camera motion vectors obtained at each pixel of the image once the affine motion model is estimated, and to exploit these measurements as a 2D histogram. More precisely, at each time t, the motion parameters θ_t of the camera motion model (1) are estimated and the vectors $\mathbf{w}_{\hat{\theta}_t}(\mathbf{p})$ are computed for each point p of the image support. The values of the horizontal and vertical components of $\mathbf{w}_{\hat{\theta}_t}(\mathbf{p})$ 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. The number of components of the mixture and their parameters are estimated in a similar way as explained in subsection 3.1.

5 Event detection algorithm

We exploit now the designed probabilistic models of motion content for the task of event detection. We proceed in two

steps.
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 modeling and recognition issues. 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 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 1D models are used which saves computation. To this end, a motion model is learned off-line in a training stage for each group of segments. Then, the sorting consists in assigning the label "play" or "no play" to each segment of the processed video using the ML criterion. In practice, because of the large diversity of content in "play" or "no play" video segments in some videos, it can be useful to learn several models per group.
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 permits to characterize more precisely a specific event. For a given genre of video document, an off-line training step is again required. A residual motion model $\mathcal{M}^j_{res}$ (cooccurrence-based GMM or contrast-based DGMM) and a camera motion model $\mathcal{M}^j_{cam}$ have to be estimated from a given training set of video samples, for each type j of event to detect. Let $\{s_0, \ldots, s_N\}$ be the previously selected video segments. $\{z_0, \ldots, z_N\}$ are the corresponding motion measurements. If we consider the GMM approach described in subsection 3.1, $z_i = y_i$ where y_i are the quantized local motion-related measurements for segment s_i. If we consider the DGMM approach (subsection 3.2), $z_i = (\gamma_i, x_i)$ where γ_i are the temporal gradients of the local motion-related measurements x_i for segment s_i. The video segments retained after the first step are then labeled with one of the J learned models of dynamic events according to the Maximum Likelihood criterion. Thus, the label l_i of the segment s_i is defined as follows :

$$l_i = \arg\max_{j=1,\ldots,J} P_{\mathcal{M}^j_{res}}(z_i) \times P_{\mathcal{M}^j_{cam}}(w_i) \qquad (7)$$

where w_i represents the motion vectors corresponding to the estimated 2D affine motion models for the segment s_i, and $P_{\mathcal{M}^j_{res}}$ is either given by $P_{cooc.}$ (3) or by $P_{contr.}$ (6) according to the chosen option.

Figure 1: *Skating video:* Left to right and top to bottom: audience, scores, skating and different dance figures.

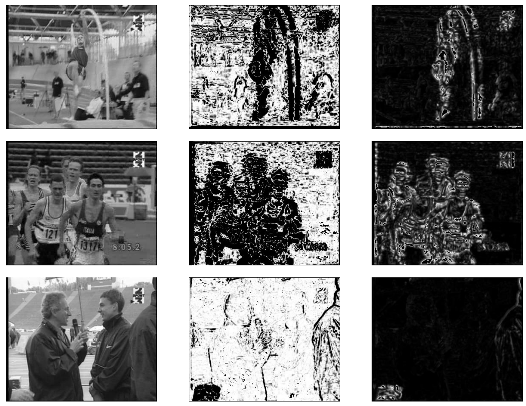

Figure 2: *Athletics video:* Three images at different time instants and their corresponding maps of dominant motion support (in white) and of local related-motion measurements v_{res} (zero-value in black). Top to bottom: pole vault, close-up of track race and interview.

6 Experimental results

In subsection 6.1, we give the results of the first step of the designed method for two different sport programs. Results of event detection are shown and commented in subsection 6.2. Finally, we report experimental comparisons in subsection 6.3. We have carried out experiments on several video programs. Due to page limitation, we report here results obtained on two different sport videos.

6.1 Selecting video segments

The first video is a figure skating (dance) TV program. We want to distinguish between "play" segments which correspond to skating (simple skating motion, artistic effects, dance movements) and "no play" segments involving low-

	P	R
GMM	1	0.83
DGMM	0.95	0.90

Table 1: *Skating video:* Results of the first step of the event detection algorithm for the two considered models (GMM and DGMM). P = precision rate and R = recall rate.

	P	R
GMM	0.79	0.94
DGMM	0.84	0.94

Table 2: *Athletics video:* Results of the first step of the event detection algorithm for the two considered models (GMM and DGMM). $P =$ precision rate and $R =$ recall rate.

level activity (views of the audience, static shot like waving at the beginning and the end of each show, skaters waiting for the scores) as illustrated in Figure 1. The first 23 minutes of the video (two shows) are used as the training set and the last 9 minutes (one show) form the test set. The video segments of the test set are sorted as described in section 5. Here, each group ("play", "no play") is represented by several motion models; then the ML criterion involves to maximise also over the different models of each group. The precision rate P and the recall rate R are defined as follows:

$$P = \frac{\#\text{correct}}{\#\text{correct} + \#\text{intrusive}} \text{ and } R = \frac{\#\text{correct}}{\#\text{correct} + \#\text{missed}},$$

where $\#$correct is the number of video segments labeled as "play" segments and which effectively belong to this group, $\#$intrusive is the number of "no play" segments labeled as "play" segments and $\#$missed is the number of "play" segments labeled as "no play" segments. Table 1 contains the first-step results obtained respectively with the GMM and DGMM models for the residual motion. Results are comparable and quite satisfactory.
We have also processed one athletics TV program which is formed by 25500 images. The training set is 10 minutes long and the test set is 7 minutes long. 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. Some representative images of this video are displayed on Figure 2. We present in Table 2 the first-step results obtained on that video. Again, satisfactory results are obtained.

6.2 Detecting relevant events

The aim is now to detect the relevant events of the athletics video among the segments selected as "play" segments (16 segments selected from the first step when using the GMM model and 17 when using the DGMM model for the residual motion). For this second step, we introduce the probabilistic camera motion model. The 2D histograms of the estimated camera motion vectors for different classes are plotted on Figure 3. The four events we try to detect are the following: pole vault (*Pv*), replay of pole vault (*Pr*), wide-shots of track race (*Rw*) and close-up of track race (*Rc*). Let us point out that the class "Replay of pole vault" contains the run-up and the jump, whereas the class "pole vault" contains only the jump. On Figure 4, the processed video is represented by a time-line exhibiting the duration of video segments. Figure 4 represents in a combined way the results of the two-step event detection method, also reported in a different and separate way by Table 2 (first step) and Table 3 (second step). "No play" is displayed in grey and a color is associated to each event class. The first row represents the ground-truth. The second and the third ones show the results obtained respectively using GMM model and DGMM model as for the residual motion. The detection also involves the camera motion model. ¿From Figure 4 and Table 2 and 3, we can infer that the majority of events are appropriately detected. Let us note that the intrusive segments appear on the line *NP* in Table 3. The classification errors concern two segments (two short segments at the end of the video sequence) belonging to the class "wide-shot of track race". The misclassification is due to the fact that the first segment involves a scene which is between wide-angle shot and close-up, and the second one is quite similar to the run-up of pole vault in terms of movement.

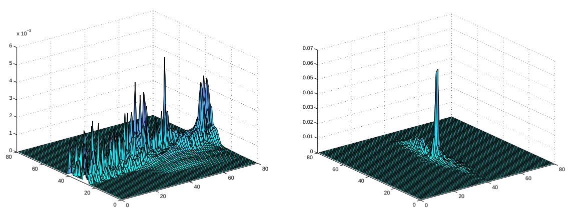

Figure 3: *Athletics video:* 2D histograms of the estimated camera motion vectors. Left: pole vault, right: wide-shot of track race.

	Assigned label			
	Pv	Pr	Rw	Rc
Pv	2	0	0	0
Pr	0	2	0	0
Rw	0	1	3	1
Rc	0	0	0	6
NP	0	1	0	0

	Assigned label			
	Pv	Pr	Rw	Rc
Pv	2	0	0	0
Pr	0	2	0	0
Rw	0	1	4	1
Rc	0	0	0	6
NP	0	1	0	0

Table 3: *Athletics video:* Classification matrix obtained with the two-step event detection method. Left: with the GMM model. Right: with the DGMM model.

Figure 4: *Athletics video:* Top: ground-truth, middle: results obtained with GMM model, bottom: results obtained with DGMM model. Grey: "no play", red: pole vault, yellow: replay of pole vault, green: wide-shot of track race, blue: close-up of track-race

6.3 Experimental comparisons

Table 4 contains results when considering the camera motion only. Conversely, Table 5 gives results obtained when

	Assigned label			
	Pv	Pr	Rw	Rc
Pv	2	0	0	0
Pr	0	2	0	0
Rw	0	1	2	2
Rc	1	0	0	5
NP	0	1	0	0

	Assigned label			
	Pv	Pr	Rw	Rc
Pv	2	0	0	0
Pr	0	2	0	0
Rw	0	1	1	4
Rc	1	0	0	5
NP	0	1	0	0

Table 4: *Athletics video:* Classification matrix obtained when considering the camera motion only. Left: on the segments selected after the first step based on the GMM model. Right: on the segments selected after the first step based on the DGMM model.

	Assigned label			
	Pv	Pr	Rw	Rc
Pv	2	0	0	0
Pr	0	2	0	0
Rw	0	0	5	0
Rc	0	0	2	4
NP	0	0	1	0

	Assigned label			
	Pv	Pr	Rw	Rc
Pv	2	0	0	0
Pr	0	2	0	0
Rw	2	3	1	0
Rc	2	0	0	4
NP	0	0	1	0

Table 5: *Athletics video:* Classification matrix obtained when considering the residual motion only. Left: with the GMM model. Right: with the DGMM model.

using the residual motion model only. These two tables demonstrate that the combination of both motions (residual motion and camera motion) yields better results as for event detection.
On the other hand, whereas the GMM model was supposed to capture more information with the cooccurences statistics, Tables 2 and 3 show that the DGMM model finally yields similar results while being less time consuming. Indeed, for the GMM model the computation time is 0.8 sec/image with a Pentium IV 2.4 Ghz, while it is 0.2 sec/image for the DGMM model.

7 Conclusion

In this paper, we have introduced new probabilistic motion models which can be easily learned and computed from the image data and can handle a large variety of dynamic video contents. We explicitly handle the information related respectively to the scene motion and to the camera motion. Two probabilistic models involving different statistical representations of the scene motion have been investigated and compared. We have also introduced an original probabilistic modeling of camera motion. 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.
The proposed two-step method for event detection is general and does not exploit very specific knowledge (related to the video genre, e.g., type of sport) and dedicated solutions. It can thus be successfully applied to a large range of videos. In the same time, due to the considered statistical framework, it is flexible enough to properly introduce prior on the classes (then, skipping to MAP instead of ML criterion) if available, or to incorporate other useful information. A complete event detection scheme should also integrate colour (the dominant colour is useful for instance to account for the presence of the play field or the tennis court in sport videos), or audio features which are also of obvious interest when processing videos. Such developments are indeed in progress for a video summarization application.

Acknowledgments

This work has been partly supported by Brittany Council (Ph-D thesis grant) and by the IST project LAVA. The authors would like to thank INA, Direction de la Recherche, for providing the videos.

References

[1] J. Vermaak, P. Pérez, and M. Gangnet, “Rapid summarization and browsing of video sequences,” *British Machine Vision Conference, Cardiff*, September 2002.

[2] N. Vasconcelos and A. Lippman, “Statistical models of video structure for content analysis and characterization,” *IEEE Trans. on Image Processing*, vol. 9, no. 1, pp. 3–19, Jan. 2000.

[3] Y-F. Ma and H-J. Zhang, “Motion pattern-based video classification retrieval,” *EURASIP Journal on Applied Signal Processing*, vol. 2, pp. 199–208, March 2003.

[4] Y. Rui and P. Anandan, “Segmenting visual actions based on spatio-temporal motion patterns,” *IEEE Int. Conf. on Computer Vision and Pattern Recognition, Hilton Head, SC*, vol. 1, pp. 111–118, 2000.

[5] L. Zelnik-Manor and M. Irani, “Event-based video analysis,” *IEEE Int. Conf. on CVPR, Kauai, Hawaii*, vol. 2, pp. 123–130, December 2001.

[6] Y. Yacoob and J. Black, “Parametrized modeling and recognition of activities,” *Sixth IEEE Int. Conf. on Computer Vision, Bombay, India*, pp. 120–127, 1998.

[7] K. A. Peker, A. A. Alatan, and A. N. Akansu, “Low-level motion activity features for semantic characterization of video,” *IEEE Int. Conf. on Mulimedia and Expo, ICME’00*, 2000.

[8] 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*, vol. 11, no. 4, pp. 393–407, 2002.

[9] J-M. Odobez and P. Bouthemy, “Robust multiresolution estimation of parametric motion models,” *J. of Visual Communication and Image Representation*, vol. 6, no. 4, pp. 348–365, December 1995.

[10] 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*, vol. 22, no. 3, pp. 719–725, 2000.