Extraction of regions of interest based on motion activity for video retrieval with partial query

Ronan Fablet, Patrick Bouthemy

9th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU'2002
Special session on "Intelligent Systems for Video Processing"

Extraction of regions of interest based on motion activity for video retrieval with partial query

Ronan Fablet
Dept of Computer Science
Brown University, Box 1910
Providence, RI02912, USA
rfablet@cs.brown.edu

Patrick Bouthemy
IRISA/INRIA Rennes
Campus universitaire de Beaulieu
35042 Rennes Cedex, France
bouthemy@irisa.fr

Abstract

This paper presents an original approach to extract regions of interest in image sequences according to motion information. It relies on motion activity analysis from non-parametric probabilistic models of local motion information. It can handle a wide range of dynamic scenes from rigid motion situations to non-rigid motion types such as articulated motions or temporal deformable entities.[1]

Keywords: entities of interest, motion activity, probabilistic models, Markovian segmentation..

1 Introduction and related work

The extraction of entities of interest within images w.r.t. motion information is of key interest in applications such as video surveillance [11], obstacle detection for vehicle navigation, or content-based video indexing and retrieval [5]. Proposed approaches can be roughly classified into motion detection schemes [12, 14, 16] and motion-based segmentation algorithms [1, 2, 15]. The first ones only determine a binary partition into regions conforming or not to the dominant image motion supposed to be due to the camera motion. The latter ones aim at defining a complete motion-based partition into homogeneous regions usually in terms of 2D parametric motion models.

[1]Work done while the first author was at IRISA, Campus universitaire de Beaulieu 35042 Rennes Cedex, France.

Motion segmentation methods are known to be computationally expensive. Furthermore, they are likely to divide articulated or deformable objects, or group of objects, into several subregions, since they rely on motion-based homogeneity criteria w.r.t. 2D parametric motion models. Yet, in the context of dynamic scene analysis, the key issue is to extract semantically meaningful entities. Motion detection schemes are usually far less CPU time consuming. However, they cannot hold one group of objects as a whole. In many situations such as ones involving "fluid" or "temporal texture" configurations (e.g., fluttering leaves, sea waves, torrents,...), or comprising groups of motion entities (e.g., players in sport videos) (see Figure 1), regions of interest do not consist of one single object. To handle these kinds of dynamic contents, it is not relevant to exploit 3D models or 2D parametric motion models. Furthermore, using an intensity-based or contour-based approach is also questionable, since explicitly defining boundaries for such entities is not easy.

a b c

Figure 1: Examples of dynamic scenes to be dealt with: (a) tracking of a given hockey player; (b) focus on a group of players on a rugby playing field; (c) wind blown trees.

In this paper, we are addressing the extraction of meaningful moving regions within image sequences without any *a priori* information, on

motion types nor on object appearance (texture, shape). We also aim at handling a wide variety of situations from rigid motion to non-rigid motion and temporal texture examples. To this end, we investigate the use of more general characterization of dynamic contents in terms of motion activity [4, 8]. We propose a general statistical framework able to extract entities of interest based on non-parametric motion activity characterization. This simultaneously provides us with self-learned motion models for further analysis in terms of motion recognition or classification. The remainder of this paper is organized as follows. Section 2 presents the local motion-related measurements considered for non-parametric motion activity modeling. In Section 3, the statistical modeling of motion information and the estimation of these models are addressed. Section 4 is concerned with image segmentation w.r.t motion activity in order to automatically extract and characterize entities of interest. Experiments carried out on real image sequences involving different situations are reported in Section 5, and Section 6 contains concluding remarks.

2 Local motion-related measurements

2.1 Dominant motion estimation

Since our goal is to characterize the actual dynamic content of the scene, we have first to cancel camera motion. To this end, we estimate the dominant image motion between successive images, and we assume that it is due to camera motion. We then warp the images of the processed sequence to a reference frame.

To model the global transformation between two successive images, we consider a 2D affine motion model (a 2D quadratic model could also be considered). The velocity $\mathbf{w}_\Theta(p)$, at pixel p, related to the affine motion model parameterized by Θ is given by:

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

with $p = (x, y)$ and $\Theta = [a_1\ a_2\ a_3\ a_4\ a_5\ a_6]$. The six affine motion parameters are computed with the robust gradient-based incremental estimation method described in [13]. An important feature of this robust approach is to supply a map of weights ω. At each point p in the image, the weight value ω_p, comprised in $[0, 1]$ indicates whether the point is likely or not to belong to the part of the image undergoing the dominant image motion. The closer ω_p to 1, the more the point p conforms to the dominant image motion.

2.2 Local motion-related measurements

We compute the local motion-related measurements in the image sequence I^* generated by compensating for the estimated dominant image motion. More precisely, we consider a weighted local average of normal flows given by [4, 14]:

$$v_{obs}(p) = \frac{\sum_{s \in \mathcal{F}(p)} \|\nabla I^*(s)\| \cdot |I_t^*(s)|}{\max\left(\eta^2, \sum_{s \in \mathcal{F}(p)} \|\nabla I^*(s)\|^2\right)} \qquad (2)$$

where $\mathcal{F}(p)$ is a 3×3 window centered on p, η^2 a predetermined constant related to the noise level (typically, $\eta = 5$). $\nabla I^*(p)$ and $I_t^*(p)$ are respectively the spatial gradient and the temporal derivative of the intensity function I^*. The considered motion-related measurement $v_{obs}(p)$ forms a more reliable quantity than normal flow, while still simply computed from intensity derivatives. It has already been successively used for motion detection [14], and for motion-based video indexing and retrieval [7, 9].

3 Statistical motion activity modeling

3.1 Temporal Gibbs models of motion activity

To model and characterize motion activity, we exploit the probabilistic framework we previously presented in [8, 9]. We briefly outline it hereafter. It can be viewed as an extension of texture modeling for grey level images with local motion quantities playing a role similar to grey levels for texture analysis. Since we exploit cooccurrence statistics, we have to deal with motion-related quantities defined over a finite set. Hence, we have to perform a quantization of the continuous motion measurements within a predefined bounded interval. It will also allow us to compare

motion activity models associated to different entities for recognition issues. Another motivation to fix an upper bound is to reject spurious motion measurements.

Let Λ be the discretized range of values for $v_{obs}(p)$, k the time instant of the current image to be processed and $\{x_{k,x_{k+1}}\}$ the pair of maps of quantized motion-related quantities for the processed video sequence computer over successive images at time instants $(k-1,k)$, and $(k,k+1)$. Let $\mathcal{R}$ denote the spatial region of interest in the image k and $\{x_k^{\mathcal{R}}, x_{k+1}^{\mathcal{R}}\}$ the restriction of the pair $\{x_{k+1}, x_k\}$ over region R. We assume that the pair $x^{\mathcal{R}} = \{x_k^{\mathcal{R}}, x_{k+1}^{\mathcal{R}}\}$ is the realization of a first-order Markov chain:

$$P_{\mathcal{M}}(x^{\mathcal{R}}) = P_{\mathcal{M}}(x_k^{\mathcal{R}}) \prod_{p\in\mathcal{R}} P_{\mathcal{M}}(x_{k+1}^{\mathcal{R}}(p)|x_k^{\mathcal{R}}(p)) \quad (3)$$

where $\mathcal{M}$ refers to the motion activity model. $P_{\mathcal{M}}(x_k^{\mathcal{R}})$ designates the a priori distribution of $x_k^{\mathcal{R}}$. We will consider in practice a uniform law for $P_{\mathcal{M}}(x_k^{\mathcal{R}})$.

In this causal modeling framework, we evaluate only temporal interactions, i.e., cooccurrence of two given motion quantities at the same grid point for two successive instants. First, it means that we focus on the temporal evolution of motion content and we can handle certain kinds of temporal non-stationarity. Second, it makes feasible an exact computation of the conditional likelihood $P_{\mathcal{M}}(x^{\mathcal{R}})$. This is crucial since it will enable to achieve model estimation in an easy way and to define an appropriate similarity measure between motion activity models based on the Kullback-Leibler divergence [3].

In addition, we consider an equivalent Gibbsian formulation of $P_{\mathcal{M}}(x_{k+1}^{\mathcal{R}}(p)|x_k^{\mathcal{R}}(p))$. It implies the introduction of potentials $\Psi_{\mathcal{M}}\left(x_{k+1}^{\mathcal{R}}(p), x_k^{\mathcal{R}}(p)\right)$ specifying the motion activity model $\mathcal{M}$ as follows:

$$\begin{array}{c} P_{\mathcal{M}}\left(x_{k+1}^{\mathcal{R}}(p)|x_k^{\mathcal{R}}(p)\right) = \\ \exp\left[\Psi_{\mathcal{M}}\left(x_{k+1}^{\mathcal{R}}(p), x_k^{\mathcal{R}}(p)\right)\right] \end{array} \quad (4)$$

As in [10, 17], a relation can be established between this Gibbsian setting and cooccurrence distributions. The conditional likelihood $P_{\mathcal{M}}(x^{\mathcal{R}})$ can be expressed according to an exponential formulation involving $\Psi_{\mathcal{M}} \bullet \Gamma^{\mathcal{R}}$, the dot product between the cooccurrence distribution $\Gamma^{\mathcal{R}}$, $(\Gamma^{\mathcal{R}}(\nu,\nu'))_{(\nu,\nu')\in\Lambda^2}$, and the potentials $\{\Psi_{\mathcal{M}}(\nu,\nu')\}_{(\nu,\nu')\in\Lambda^2}$, defined by:

$$\begin{cases} \Gamma^{\mathcal{R}}(\nu,\nu') = \sum_{p\in\mathcal{R}} \delta(\nu - x_{k+1}^{\mathcal{R}}(p))\delta(\nu' - x_k^{\mathcal{R}}(p)) \\ \Psi_{\mathcal{M}} \bullet \Gamma^{\mathcal{R}} = \sum_{(\nu,\nu')\in\Lambda^2} \Psi_{\mathcal{M}}(\nu,\nu') \cdot \Gamma^{\mathcal{R}}(\nu,\nu') \end{cases} \quad (5)$$

where δ is the Kronecker symbol. We finally get:

$$P_{\mathcal{M}}(x^{\mathcal{R}}) = P_{\mathcal{M}}(x_k^{\mathcal{R}}) \cdot \exp\left[\Psi_{\mathcal{M}} \bullet \Gamma^{\mathcal{R}}\right] \quad (6)$$

This modeling scheme can be stated as non-parametric in two ways. On one hand, the statistical model $\mathcal{M}$ does not refer to a 2D parametric motion model, but expresses in a broad sense the notion of motion activity. On the other hand, $\{\exp \Psi_{\mathcal{M}}(\nu,\nu')\}_{(\nu,\nu')\in\Lambda^2}$ is not assumed to follow a known parametric law (Gaussian,...).

The availability of the exponential formulation (6) presents several interests. First, it makes the computation of the conditional likelihood $P_{\mathcal{M}}(x^{\mathcal{R}})$, for any sequence x and region $\mathcal{R}$, feasible and simple. Second, for classification or recognition issues, the storage of the motion-related quantities $x^{\mathcal{R}}$ is not required. In fact, all the motion information exploited in our approach is captured by the cooccurrence distribution $\Gamma^{\mathcal{R}}$.

3.2 Maximum likelihood estimation

Given a set of motion-related quantities $x^{\mathcal{R}}$ within a region $\mathcal{R}$, we have to identify the model $\widehat{\mathcal{M}}^{\mathcal{R}}$ specified by its potentials $\{\Psi_{\widehat{\mathcal{M}}^{\mathcal{R}}}(\nu,\nu')\}_{(\nu,\nu')\in\Lambda^2}$ which best fits $x^{\mathcal{R}}$. We adopt the Maximum Likelihood (ML) criterion:

$$\begin{array}{c} \widehat{\mathcal{M}}^{\mathcal{R}} = \arg\max_{\mathcal{M}} LL_{\mathcal{M}}(x^{\mathcal{R}}) \\ \text{with } LL_{\mathcal{M}}(x) = \ln P_{\mathcal{M}}(x^{\mathcal{R}}) \end{array} \quad (7)$$

The considered temporal Gibbsian model reduces to a product of $|\mathcal{R}|$ independent conditional transitions from instant k to instant $k+1$, and is specified by the transition matrix $\{P_{\widehat{\mathcal{M}}^{\mathcal{R}}}(\nu|\nu')\}_{(\nu,\nu')\in\Lambda^2}$. Therefore, the ML model estimate $\widehat{\mathcal{M}}^{\mathcal{R}}$ is readily determined from the empirical estimation of the transition probability $P_{\widehat{\mathcal{M}}^{\mathcal{R}}}(x_{k+1}^{\mathcal{R}}(r)|x_k^{\mathcal{R}}(r))$ as follows:

$$\Psi_{\widehat{\mathcal{M}}^{\mathcal{R}}}(\nu,\nu') = \ln\left(\Gamma^{\mathcal{R}}(\nu,\nu') / \sum_{\vartheta\in\Lambda} \Gamma^{\mathcal{R}}(\vartheta,\nu')\right) \quad (8)$$

4 Motion activity segmentation

We now describe how we exploit the non-parametric statistical motion modeling described above to extract meaningful entities, comprising pertinent motion activity, in frame k of the processed image sequences. Motion activity properties are learned in an unsupervised way at the level of a group of pixels (elementary blocks or subset of blocks). We initially consider a block-based partition of the image (typically, 32×32 blocks).

We believe that the loss of accuracy induced by considering small blocks is not a real shortcoming here. First, for cases involving few mobile objects, the interest is mainly to identify pertinent entities and not necessarily to accurately determine their boundaries. Second, this formulation allows us to handle a wider range of motion types compared to other techniques. Besides, this approach will prove far less time consuming than motion-based region segmentation or level-set algorithms.

Let us consider a partition of the image defined by the set of N_{bl} blocks $\{\mathcal{B}_i\}_{i\in\{0,\ldots,N_{bl}\}}$. We further assume that motion activity within each block $\mathcal{B}_i$ is characterized by means of the associated statistical model $\mathcal{M}^{\mathcal{B}_i}$ derived from the cooccurrence distribution $\Gamma^{\mathcal{B}_i}$ as explained in Section 3.

4.1 Statistical similarity measure related to motion activity

Considering two regions $\mathcal{R}$ and $\mathcal{R}'$ (regions $\mathcal{R}$ and $\mathcal{R}'$ can be elementary blocks among $\{\mathcal{B}_i\}_{i\in\{0,\ldots,N_{bl}\}}$ or groups of blocks), we have to evaluate the degree of similarity between their respective motion activity contents. Let us note $\mathcal{M}^{\mathcal{R}}$ and $\mathcal{M}^{\mathcal{R}'}$ the statistical models of motion activity attached to regions $\mathcal{R}$ and $\mathcal{R}'$ resp. corresponding to the pair of motion-related measurements. $x^{\mathcal{R}} = \{x_k^{\mathcal{R}}, x_{k+1}^{\mathcal{R}}\}$ and $x^{\mathcal{R}'} = \{x_k^{\mathcal{R}'}, x_{k+1}^{\mathcal{R}'}\}$, and to the associated cooccurrence distributions $\Gamma^{\mathcal{R}}$, $\Gamma^{\mathcal{R}'}$.

We have designed a similarity measure relying on the Kullback-Leibler (KL) divergence [3]The KL divergence evaluates the distance between two probability distributions as the expectation of their log-ratio. The motion activity similarity measure $D(\mathcal{M}^{\mathcal{R}}, \mathcal{M}^{\mathcal{R}'})$ is a symmetrical version of the KL divergence:

$$D(\mathcal{M}^{\mathcal{R}}, \mathcal{M}^{\mathcal{R}'}) = \frac{\left[KL(\mathcal{M}^{\mathcal{R}}\|\mathcal{M}^{\mathcal{R}'}) + KL(\mathcal{M}^{\mathcal{R}'}\|\mathcal{M}^{\mathcal{R}})\right]}{2} \quad (9)$$

where $KL(\mathcal{M}^{\mathcal{R}}\|\mathcal{M}^{\mathcal{R}'})$ denotes the KL divergence. It is approximated as [9]:

$$KL(\mathcal{M}^{\mathcal{R}}\|\mathcal{M}^{\mathcal{R}'}) \approx \frac{1}{|\mathcal{R}|} \ln\left(\frac{P_{\mathcal{M}^{\mathcal{R}}}(x^{\mathcal{R}})}{P_{\mathcal{M}^{\mathcal{R}'}}(x^{\mathcal{R}})}\right) \quad (10)$$

Since $\mathcal{M}^{\mathcal{R}}$ is the ML model estimate associated with the cooccurrence distribution $\Gamma^{\mathcal{R}}$, $KL(\mathcal{M}^{\mathcal{R}}\|\mathcal{M}^{\mathcal{R}'})$ is positive and equals 0 if the two statistical distributions are identical. In fact, this ratio quantifies the loss of information occurring when substituting $\mathcal{M}^{\mathcal{R}'}$ for $\mathcal{M}^{\mathcal{R}}$ to account for motion activity within area $\mathcal{R}$.

4.2 Region-level graph labeling

We now present the labeling stage of the block-based partition of the image, $\{\mathcal{B}_i\}_{i\in\{0,\ldots,N_{bl}\}}$. It will be exploited to extract regions of interest comprising significant motion activity as described in the next section. It relies on a Markovian region-level labeling framework applied to the adjacency graph $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ where $\mathcal{N}$ is the set of nodes of graph $\mathcal{G}$ and $\mathcal{A}$ the set of its arcs. Each node $n_i \in \mathcal{N}$ holds for block $\mathcal{B}_i$ with $i \in \{0, \ldots, N_{bl}\}$, and $\mathcal{A}$ represents the set of arcs between graph nodes corresponding to connected blocks (in practice, we consider a four-connectivity block neighborhood). Over this graph structure $\mathcal{G}$, we define a region-level Markov random field model the sites of which are the nodes of graph $\mathcal{G}$. In addition, two-site clique neighborhood system is deduced from the set of arcs $\mathcal{A}$.

Let us assume that a set of labels $\mathcal{L}$ relative to different motion activity models has been specified. We further consider that to each label $l \in \mathcal{L}$ is attached the cooccurrence distribution Γ^l and the associated motion activity model $\mathcal{M}^l$. We will more precisely explain in the next subsection how these motion activity models are defined and estimated.

Let us note $e = \{e_{n_i}\}_{i\in\{0,\ldots,N_{bl}\}}$ the label field with e_{n_i} taking value in $\mathcal{L}$, and $o =$

$\{o_{n_i}\}_{i\in\{0,\ldots,N_{bl}\}}$ the observation field. In our case, at each node n_i, o_{n_i} refers to the motion activity characterization attached to block $\mathcal{B}_i$, i.e. both cooccurrence distribution $\Gamma^{\mathcal{B}_i}$ and motion activity model $\mathcal{M}^{\mathcal{B}_i}$. Adopting the Maximum A Posteriori (MAP) criterion, the Markovian labeling scheme comes to solve for:

$$\hat{e} = \arg \min_{e\in\mathcal{L}^{N_{bl}}} U(e,o) \tag{11}$$

where $U(e,o) = U^a(e,o) + U^b(e)$, with U^a the data-driven energy term, and U^b the regularization term. Both are split in the sum of local potentials and they are defined by:

$$\begin{cases} U^a(e,o) = \sum\limits_{n_i\in\mathcal{N}} V^a(e_{n_i}, o_{n_i}) \\ U^b(e) = \sum\limits_{(n_i,n_j)\in\mathcal{A}} \beta\cdot\delta(e_{n_i} - e_{n_j}) \end{cases} \tag{12}$$

with β a parameter tuning the relative importance of the regularization term and δ the Kronecker function. The potential $V^a(e_{n_i}, o_{n_i})$ quantifies at node n_i how relevant the description of observation o_{n_i} by label e_{n_i} is. It involves the similarity measure D, defined in relation (9), and is given by:

$$V^a(e_{n_i}, o_{n_i}) = \exp\left[-D\left(\mathcal{M}^{e_{n_i}}, \mathcal{M}^{\mathcal{B}_i}\right)\right] \tag{13}$$

We introduce an exponential form to get values of potential V^a within the range $[0,1]$, which enables to more easily set the regularization parameter β. The minimization issue (11) is tackled using the HCF (Highest Confidence First) algorithm [6], since the number of nodes of the considered graph is relatively small.

4.3 Separation of entities of interest from static background

We want to separate regions of interest from background. Since we consider image sequences compensated for the estimated dominant image motion assumed to be due to the camera motion, this issue reduces to extract regions which do not conform to the dominant image motion.

In a first step, we determine a binary labeling of the initial block partition of the image in terms of blocks conforming or not to the dominant image motion. As the image segmentation proceeds from motion activity characterization, we have to establish a model corresponding to the static background. Even if we could merely infer an activity model corresponding to zero-value motion measurements, we prefer to explicitly estimate the background model from actual motion-related quantities at points attached to the extracted background, since camera motion cannot be perfectly cancelled. To achieve this, we exploit the map of weights ω issued from the robust multiresolution estimation of the affine motion model accounting for the dominant image motion (see Section 2). By thresholding map ω, we can roughly determine the support associated to the estimated dominant image motion. It is formed by points p satisfying $\omega_p > \mu$. Using our statistical motion activity modeling framework, we can estimate the model $\mathcal{M}^{\mathcal{S}_d}$ attached to the background $\mathcal{S}_d$. If $\overline{\mathcal{S}_d}$ designates the complementary set of $\mathcal{S}_d$ (corresponding to the outlier map), we can evaluate in the same way the associated motion activity model $\mathcal{M}^{\overline{\mathcal{S}}_d}$.

At this stage, we achieve afirst Markovian block-based labeling, as described in subsection 4.2, considering only two labels referring to statistical models attached to regions $\mathcal{S}_d$ and $\overline{\mathcal{S}_d}$ (i.e. label set $\mathcal{L}$ contains only two labels in that case). The obtained binary segmentation allows us to update the support of regions $\mathcal{S}_d$ and $\overline{\mathcal{S}_d}$, and consequently their associated models $\mathcal{M}^{\mathcal{S}_d}$ and $\mathcal{M}^{\overline{\mathcal{S}}_d}$ can be updated too. Since $\overline{\mathcal{S}_d}$ includes the regions of interest we are seeking, we determine its connected components. Let us denote $\{\mathcal{R}_i\}_{i\in\{1,\ldots,N_{reg}\}}$ the N_{reg} resulting regions. For each region $\mathcal{R}_i$, we estimate its motion activity model $\mathcal{M}^{\mathcal{R}_i}$. We then perform a second region-level labeling stage applied to the original block-based partition as explained in subsection 4.2, with now $|\mathcal{L}| = N_{reg} + 1$. We consider $N_{reg}+1$ different labels $\{l_0, .., l_{N_{reg}}\}$ corresponding to the updated model $\mathcal{M}^{\mathcal{S}_d}$ (label l_0) and to models $\{\mathcal{M}^{\mathcal{R}_i}\}_{i\in\{1,\ldots,N_{reg}\}}$. Once convergence is reached, regions $\mathcal{Q}_i$ formed by blocks with labels $l_i \neq l_0$ are regarded as the entities of interest for the processed image sequence. Moreover, the dynamic content within each extracted entity $\mathcal{Q}_i$ is characterized by its motion activity model $\mathcal{M}^{\mathcal{Q}_i}$, which is updated at the final step of our scheme.

a b c

Figure 2: Illustration of the successive stages of our method to motion activity segmentation: (a) one image of the processed sequence; (b) support (in white) of the dominant image motion estimated between the considered image (a) and the preceding frame in the sequence; (c) result of the motion activity segmentation (parameter setting is given in text in subsection 5.1). The black area holds for the region regarded as the static background and the two extracted regions of interest are displayed and delimited by a white border.

5 Results

5.1 Illustration of the successive method stages

We first illustrate the different stages of our method in Fig.2. The parameters involved are set as follows: the motion-related measurements have been quantized within $[0, 4]$ on 16 levels; to determine the support of the dominant image motion, we take $\mu = 0.2$; estimator is set to 0.5; the Markovian region-level labeling has been performed with $\beta = 0.5$ and 32×32 blocks (i.e., about 20 square blocks for images of size 120×160). Let us stress that the same parameter setting is used in all experiments reported in this paper.

As illustration, we have considered a wide-angle shot of a basketball game with one of the processed images displayed in Fig.2.a. This sequence involves a slight panning and zooming of the camera. Fig.2.b contains the support of the dominant image motion estimated between the considered frame and the preceding one in the sequence. Extracted regions of interest are shown in Fig.2.c. These two entities indeed comprise the interesting parts of the scene content. It should be pointed out that one player does not appear in any of these two areas since he is static. Our method also appears efficient in terms of computational complexity. It requires about 0.2 second of CPU time to process images of size 120×160 comprising 20 blocks of size 32×32 on a Sun Creator workstation 360MHZ.

5.2 Experimental evaluation of the extraction of regions of interest

We report in Fig.3 five examples of extraction of regions of interest in image sequences. We have considered four kinds of dynamic scenes: sequences involving a single or only a few moving elements, rigid motion situations, wide-angle sport shots where regions of interest are formed by several players, and sequences involving temporal texture samples. For all situations, we have succeeded in identifying meaningful entities. As far as sport sequences are concerned, it should be pointed out that superimposed logos or score captions are also extracted as illustrated in Fig.3. This is due to the fact they are associated to important residual motion displacements in case of a moving camera, since we perform motion activity analysis in images compensated for the estimated dominant image motion.

The last two examples shown in Fig.3 are particularly interesting. The first one is a wide-angle shot of a rugby match with the camera tracking the action. We successfully recover an area comprising all players except one. The last example appears to be a particularly difficult situation. It involves a camera tracking a windsurfer with the background formed by a static area (sky) and wavy sea. Furthermore, the boundary of the windsurf board is not well-defined due to the presence of waves with foam. In spite of these difficulties, we identify a relevant area comprising almost all the windsurfer and his board.

As shown in Figure 4, our approach for scene activity segmentation could also be exploited for tracking entities of interest within image sequences. For “Stefan” sequence, we succeed in extracting from image to image the region of interest in the scene, i.e., the tennis player.

6 Conclusion

We have presented an original approach for the extraction of regions of interest in an image sequence upon a broad notion of pertinent dynamic content. Motion information is expressed as non-

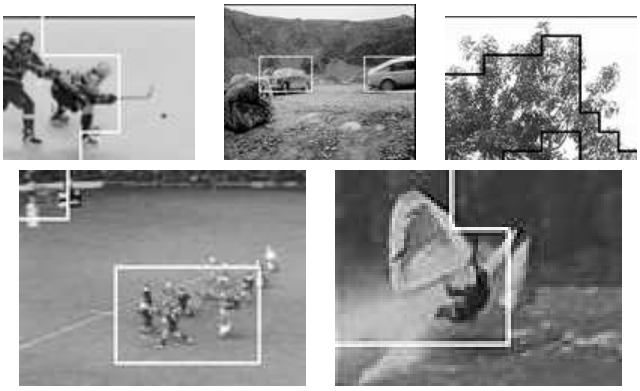
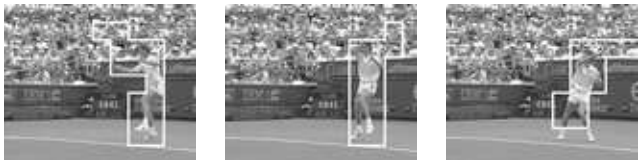


Figure 3: Extraction of regions of interest according to motion activity in image sequences. The extracted areas are delimited in white or in black depending on the examples.



frame 105 frame 110 frame 115

Figure 4: Extraction of regions of interest over successive images within "Stefan" sequence. Results for frames 100, 105 and 110 of "Stefan" sequence are reported. Entities of interest are delimited in white.

parametric statistical motion models. These models are directly estimated from the data (image sequences). They supply a characterization in terms of motion activity and are far more flexible compared to 2D parametric motion models or 3D motion models. In particular, it allows us to cope with a wide variety of dynamic content such as rigid motion, articulated motion, deformable and fluid motion (temporal texture), and group of moving entities. The extraction of regions of interest according to motion activity is formulated as a block-based labeling issue embedded in a properly formalized statistical framework. Moreover, our method directly discriminates regions comprising relevant motion activity from the non-pertinent background. We have provided a convincing validation of our method on real image sequences representative of various dynamic scenes (car sequences, sports videos, temporal textures). We have in particular demonstrated its capability to extract meaningful regions of interest comprising several moving objects, as in wide-angle shots of sport sequences, or to handle temporal textures such as wind blown trees ow wavy sea..

Another important feature of our segmentation method is that it simultaneously delivers motion activity characterization within extracted entities of interest, which could be straightforwardly exploited for motion classification or recognition purpose, or for video retrieval with partial query by example. In particular, in [7], we have shown the effectiveness of our approach to tackle motion-based video retrieval with partial query by example. To this end, we first build a database of entities of interest E extracted from video shot keyframe and characterized by a statistical motion activity model $\mathcal{M}^E$. Then, given an area in the image proposed as partial query $\mathcal{R}_q$, we compute the associated set of motion-related measurements $x^{\mathcal{R}_q}$ and temporal cooccurrence distribution $\Gamma^{\mathcal{R}_q}$. The retrieval operation is then stated as a Bayesian inference issue w.r.t. the MAP criterion and relies on the ranking of the conditional likelihood $P_{\mathcal{M}^E}(x^{\mathcal{R}_q})$ for all the elements E of the database of entities of interest.

Acknowledgments

The authors wish to acknowledge the support of INA, Département Innovation, Direction de la Recherche, for providing the MPEG-1 car sequences, which are excerpts of the INA/GDR-ISIS video corpus, and to National University of Singapore for supplying the sequences of temporal textures.

References

[1] Y. Altunbasak, P. Ehran Eren, and A. Murat Tekalp. Region-based parametric motion segmentation using color information. *Graphical Models and Image Processing*, 60(1):13–23, 1998.

[2] S. Ayer and H.S. Sawhney. Layered representation of motion video using robust maximum-likelihood estimation of mixture models and MDL encoding. In *Proc. of 5th IEEE Int. Conf. on Computer Vision,*

ICCV'95, pages 777–784, Boston, June 1995.

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

[4] P. Bouthemy and R. Fablet. Motion characterization from temporal cooccurrences of local motion-based measures for video indexing. In *Proc. of 14th Int. Conf. on Pattern Recognition, ICPR'98*, pages 905–908, Brisbane, Aug. 1998.

[5] R. Brunelli, O. Mich, and C.M. Modena. A survey on the automatic indexing of video data. *Jal of Vis. Comm. and Im. Repr.*, 10(2):78–112, 1999.

[6] P.B. Chou and C.M. Brown. The theory and practice of Bayesian image modeling. *Int. Jal of Comp. Vis.*, 4(3):185–210, 1990.

[7] R. Fablet and P. Bouthemy. Non-parametric motion activity analysis for statistical retrieval with partial query. *Journal of Mathematical Imaging and Vision*, 14(3):257–270, 2001.

[8] R. Fablet and P. Bouthemy. Non parametric motion recognition using temporal multiscale gibbs models. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'2001*, Kauai, Dec. 2001.

[9] R. Fablet, P. Bouthemy, and P. Pérez. Statistical motion-based video indexing and retrieval. In *Proc. of 6th Int. Conf. on Content-Based Multimedia Information Access, RIAO'2000*, pages 602–619, Paris, Apr. 2000.

[10] G.L. Gimel'Farb. Texture modeling by multiple pairwise pixel interactions. *IEEE Trans. on PAMI*, 18(11):1110–1114, 1996.

[11] I. Haritaoglu, D. Harwood, and L. S. Davis. W^4: Real-time surveillance of people and their activities. *IEEE Trans. on PAMI*, 22(8):809–830, 2000.

[12] M. Irani, B. Rousso, and S. Peleg. Detecting and tracking multiple moving objects using temporal integration. In *Proc. of 2nd Eur. Conf. on Computer Vision, ECCV'92*, pages 282–287, Santa Margherita, May 1992.

[13] J.M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Jal of Vis. Comm. and Im. Repr.*, 6(4):348–365, 1995.

[14] J.M. Odobez and P. Bouthemy. Separation of moving regions from background in an image sequence acquired with a mobile camera. In *Video Data Compression for Multimedia Computing*, chapter 8, pages 295–311. H. H. Li, S. Sun, and H. Derin, eds, Kluwer, 1997.

[15] J.M. Odobez and P. Bouthemy. Direct incremental model-based image motion segmentation for video analysis. *Signal Processing*, 6(2):143–155, 1998.

[16] N. Paragios and R. Deriche. Geodesic active contours and level sets for the detection and tracking of moving objects. *IEEE Trans. on PAMI*, 22(3):266–280, 2000.

[17] S.C. Zhu, T. Wu, and D. Mumford. Filters, random fields and maximum entropy (FRAME) : towards a unified theory for texture modeling. *Int. Jal of Comp. Vis.*, 27(2):107–126, 1998.