ACTIVITY-BASED TEMPORAL SEGMENTATION FOR VIDEOS OF INTERACTING OBJECTS USING INVARIANT TRAJECTORY FEATURES

A. Hervieu[1]*, P. Bouthemy*[1] *and J-P. Le Cadre*[2]

[1]INRIA, Centre Rennes - Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes, France
[2]IRISA / CNRS, Campus de Beaulieu, 35042 Rennes, France

ABSTRACT

This paper presents a content-based approach for temporal segmentation of videos. Tracked objects are characterized by their 2D trajectories which are used in a meaningful way to model visual semantics, *i.e.*, the observed single video object activities and their interactions. To this end, hierarchical Semi-Markov Chains (SMCs) are computed in order to take into account the temporal causalities of object motions. Object movements are characterized using local invariant features computed from the curvature and velocity values while interactions are represented by the temporal evolution of the distance between objects. We have evaluated our method on squash video sequences, and have favorably compared with other methods including Hidden Markov Models (HMMs).

Index Terms— Video signal processing, Hidden Markov models, Motion analysis, Pattern classification.

1. INTRODUCTION

Understanding activities and behaviors in videos is of increasing interest in a number of applications such that video surveillance, sports video exploitation, video on demand ... Object detection and tracking now provide reliable information (*i.e.*, mobile object's trajectories) that may be helpful for semantical analysis of videos.

The typical structure for content-based video analysis relies on the "frame-based" approach, including first a shot boundary detection and, in a second stage, shot classification and characterization by keyframes [5, 10, 3]. These methods are well-suited for broadcast applications but do not focus on the available object-based information embedded in the videos. However, when considering the problem of video-surveillance and sports video analysis, this traditional video analysis structure is not adapted since such actions are often filmed using a single camera (for example, crossroads or parking surveillance scene are often continuously filmed using only one single fixed camera). Considering a shot analysis (segmentation and classification) approach is then unadapted since the whole sequence would be identified as a single shot. Considering the high-level information provided by the video object detection and tracking techniques may then be of crucial interest.

Several works tackle the issue of using video object for semantical analysis. Günsel et al. [6] developed an object-based indexing of video filmed by a single camera, dealing with the motion and shape properties of the viewed objects and considering the camera motion and the object trajectories and interactions. A work by Hervieu et al. [7, 8] proposed a HMM-based shot classification method and rare event detection using the mobile object trajectories that may be used after a shot segmentation and a tracking processing. However, it did not account for interactions between objects. A system that efficiently models interactions between moving entities in a video surveillance context and relying on Coupled Hidden Markov Models [1] was also proposed by Oliver et al. [11]. Hongeng et al. [9] described a complex event recognition method based on the definition of scenarios and relying on the use of multi-agent Semi-Markov Chains (SMCs) to analyze object trajectories.

In this paper a method is proposed for recognizing actions in videos and, thus, allowing for temporal segmentation of videos filmed by a single camera. Invariant features (to translation, rotation and scale transformation) are computed on the object trajectories. In contrast to previous works, these invariant features are adapted to learning and processing the same activities filmed by different cameras (one single camera for each considered video) in a justified way. To this end, a hierarchical SMC-based modeling is proposed where each considered SMC state corresponds to a semantic phase of the viewed activity, providing an efficient modeling to detect and segment phases in video surveillance and sports videos.

In Section 2, we introduced the translation, rotation and scale invariant features. In Section 3, an original SMC-based method for temporal segmentation and activity phases recognition is proposed. In Section 4, the data set used to test the method is presented, results are then described and analyzed.

2. INVARIANT ACTIVITY FEATURES

To process different video shootings, a model of activity should be invariant to irrelevant transformation of the data. In the video context, invariance to 2D translation, 2D rotation and 2D is often a desirable component.

2.1. Kernel approximation

A video object VO_k is characterized by a trajectory T_k, which is composed of a set of n_k points corresponding to the temporal successive positions of the moving object in the image plane, *i.e.*, $T_k = \{(x_{1,k}, y_{1,k}), .., (x_{n_k,k}, y_{n_k,k})\}$.
Relying on the works of Hervieu et al. [7, 8] we reliably compute the local differential trajectory features (*i.e.*, $\dot{u}_{t,k}$, $\dot{v}_{t,k}$, $\ddot{u}_{t,k}$ and $\ddot{v}_{t,k}$, u and v being defined below) from a continuous representation of a curve approximating the trajectory T_k defined by $\{(u_{t,k}, v_{t,k})\}_{t\in[1;n_k]}$ with:

$$u_{t,k} = \frac{\sum_{j=1}^{n_k} e^{-(\frac{t-j}{h})^2} x_{j,k}}{\sum_{j=1}^{n_k} e^{-(\frac{t-j}{h})^2}},\ v_{t,k} = \frac{\sum_{j=1}^{n_k} e^{-(\frac{t-j}{h})^2} y_{j,k}}{\sum_{j=1}^{n_k} e^{-(\frac{t-j}{h})^2}}.$$

2.2. Invariant feature for individual video object activity characterization

In this subsection, the chosen feature providing an invariant representation of the activity embedded in a single moving video object is presented. To have the desired invariant representation of a video object VO_k, a relevant feature was considered, defined by:

$$\dot{\gamma}_{t,k} = \frac{\ddot{v}_{t,k}\dot{u}_{t,k} - \ddot{u}_{t,k}\dot{v}_{t,k}}{\dot{u}_{t,k}^2 + \dot{v}_{t,k}^2} = \kappa_{t,k}.w_{t,k}$$

where $\gamma_{t,k} = \arctan(\frac{\dot{v}_{t,k}}{\dot{u}_{t,k}})$ corresponds to the local orientation of the trajectory T_k, $\kappa_{t,k} = \frac{\ddot{v}_{t,k}\dot{u}_{t,k} - \ddot{u}_{t,k}\dot{v}_{t,k}}{(\dot{u}_{t,k}^2 + \dot{v}_{t,k}^2)^{\frac{3}{2}}}$ is the curvature of the trajectory T_k and $w_{t,k} = (\dot{u}_{t,k}^2 + \dot{v}_{t,k}^2)^{\frac{1}{2}}$ is the velocity of point $(u_{t,k}, v_{t,k})$.

It can be shown [7] that $\dot{\gamma}_{t,k}$ is invariant to translation, rotation and scale in the frame. The considered feature vector used to characterize a given activity of a video object VO_k is the vector containing the successive values of $\dot{\gamma}_{t,k}$:

$$V_k = [\dot{\gamma}_{1,k}, \dot{\gamma}_{2,k}, ..., \dot{\gamma}_{n_k-1,k}, \dot{\gamma}_{n_k,k}].$$

2.3. Invariant feature for interaction characterization

Taking into account the interaction between two video objects is of crucial interest to have a representation of complex activities in videos. A way to characterize these interactions is to consider the spatial distance. At each successive time i, this distance between two video objects VO_k and VO_l represented by two trajectories T_k and T_l is defined by:

$$d_i = \sqrt{(u_{i,k} - u_{i,l})^2 + (v_{i,k} - v_{i,l})^2}.$$

More specifically, the normalized distance is computed, *i.e.*:

$$\tilde{d}_i = d_i / d_{norm}.$$

The distance d_i is trivially a translation and rotation invariant feature that may help characterizing interactions between video objects. To also have a scale invariant feature, a contextual normalizing factor d_{norm} has to be known and computed in the considered videos (in the processed squash videos, the distance between the two sides of the court has been considered). The feature vector D used to characterize a given interaction between two video objects VO_k and VO_l is the vector containing the successive values of $\tilde{d}_i$:

$$D = [\tilde{d}_1, \tilde{d}_2, ..., \tilde{d}_{n_k-1}, \tilde{d}_{n_k}].$$

Hence, considering the V_k and D feature vectors helps characterizing invariantly (to translation, rotation and scale transformations) both the single video objects activities and the interactions between video objects.

3. SUPERVISED MODELING OF ACTIVITY USING HIERARCHICAL SMC

The use of SMCs to model activities is based upon a specific modeling of the feature (*i.e.*, $\dot{\gamma}$ and $\tilde{d}$) used to characterize the SMCs states. Each of these features is modeled, in a first layer, using a HMM-based approach proposed in [7, 8]. In a second stage, these HMM-based modelings will be used in a hierarchical way to model activities using SMCs (see Fig. 1).

3.1. Feature modelings using HMMs

Tackling with the first layer modeling of the features (*i.e.*, $\dot{\gamma}_i$ and $\tilde{d}$), the HMM modeling proposed in [7] has been used to build a probabilistic modeling of the spatio-temporal behavior of the V_k and D feature vectors.

To model activities involving two distinct video objects, a common HMM modeling will be used both for the $\dot{\gamma}_i$ and $\tilde{d}$ features. In the following, we present this HMM used to model the $\dot{\gamma}$ of the two video objects. The $\tilde{d}$ features are further modeled using the same HMM-framework.

The considered HMM framework is based upon a proper quantization of $\dot{\gamma}$. An interval $[-I, I]$ containing a given percentage P_v of all the computed $\dot{\gamma}$ (for any video object) is defined. A quantization is performed on $[-I, I]$ into a fixed number N of bins (defined as the states of the HMMs, [7]).

The HMM modeling the video object VO_k is then characterized by:

- the state transition matrix $A_k = \{a_{ij,k}\}$ with

$$a_{ij,k} = P[\ q_{t+1,k} = S_j \mid q_{t,k} = S_i\ ],\ 1 \leq i, j \leq N,$$

where $q_{t,k}$ is the state variable at instant t and S_i is its value (corresponding to the ith bin of the quantized histogram);

- the initial state distribution $\pi_k = \{\pi_{i,k}\}$, with $\pi_{i,k} = P[\ q_{1,k} = S_i\ ],\ 1 \leq i \leq N$;

- the conditional observation probabilities $B = \{b_i(\dot{\gamma}_t)\}$, where $b_i(\dot{\gamma}_t) = P[\dot{\gamma}_t \mid q_t = S_i]$, since the computed $\dot{\gamma}_t$ are the observed values.

The conditional observation probabilities are independent for any video object, and defined as Gaussian distributions of mean μ_i (corresponding to the median value of the histogram bin S_i). Their standard deviations σ are specified so that the interval $[\mu_i - \sigma, \mu_i + \sigma]$ corresponds to the bin width [7].

Empirical estimations of A and π are given by (see [4]):

$$a_{ij,k} = \frac{\sum_{t=1}^{n_k-1} H_{t,k}^{(i)} H_{t+1,k}^{(j)}}{\sum_{t=1}^{n_k-1} H_{t,k}^{(i)}} \text{ and } \pi_i = \frac{\sum_{t=1}^{n_k} H_{t,k}^{i}}{n_k}$$

where $H_{t,k}^{(i)} = P(\dot{\gamma}_t | q_t = i)$. Training videos are used to find the parameters of the HMMs modeling ϕ, *i.e.*, the B, A and π parameters for each of the defined HMMs.

3.2. Activity modeling by hierarchical SMC

As showed on Fig. 1, the HMM-based modelings defined in the previous subsection are used to characterize activities in a higher sense. The states S'_i of the SMC modeling defines activity phases (such that "rally" and "passive" phases in a squash game, for example) and SMCs are used to model their respective state duration sd_i. Each of these SMC states is respectively hierarchically characterized by two HMMs describing the activities of two video objects (using the $\dot{\gamma}$ feature of each of the two video objects), and one other HMM describing the interactions (*i.e*, describing the $\tilde{d}$ feature). In addition to these three HMMs, the states durations are also modeled, for each SMC state, by Gaussian Mixture Models (GMMs) using forward-backward procedures (initializations of these procedures being computed using K-means).

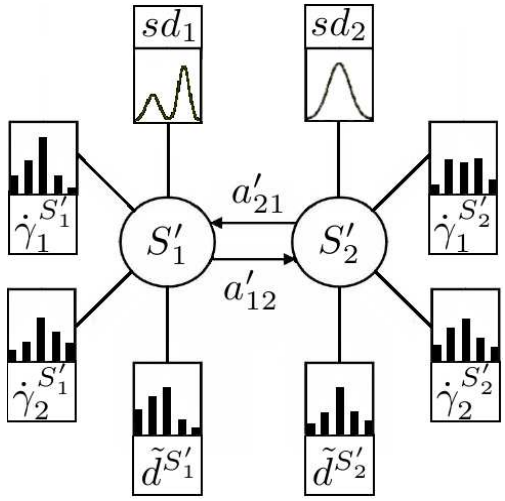

Fig. 1. Hierarchical SMC modeling with 2 states corresponding to different activity phases S'_1 and S'_2 with 2 video objects VO_1 and VO_2. Each of these states is characterized by three HMMs (modeling $\dot{\gamma}_1^{S'_i}$, $\dot{\gamma}_1^{S'_i}$ and $\tilde{d}^{S'_i}$ for the considered state S'_i) and by a GMM modeling sd_i, the state duration density in the SMC state S'_i.

As well as with ϕ, training videos are used to find ψ which are the state duration modeling parameters.

Suppose a state sequence s that has R segments, and let q_r be the time index of the end-points of the r^{th} segment, such that the data points in the r^{th} segment are $y_{(q_{r-1}+1,q_r]} = y_{qr-1+1}, \ldots, y_{q_r}$ and $s'_{q_{r-1}+1} = \ldots = s'_{q_r}$ (s' being the SMC state sequence). A' is the SMC state transition probability matrix at $\{q_i\}$, so that in the proposed modeling with only two SMC states (Fig. 1), $a'_{21} = a'_{12} = 1$ and $a'_{11} = a'_{22} = 0$.

Hence, after training, the whole modeling parameter set $\theta = \{A', \phi, \psi\}$ is available. Thus, to retrieve the temporal phases of the activity and, hence, to process a temporal segmentation of the video, a Viterbi decoding is processed.

The Viterbi algorithm find the SMC state sequence that maximizes the likelihood. This likelihood $P(y, s'|\theta)$ is defined such that, for an observation sequence y, and the corresponding SMC state sequence s':

$$\begin{aligned} P(y, s'|\theta) &= \prod_{r=1}^{R} P(s'_r|s'_{r-1}) \\ &\times \prod_{r=1}^{R} P(sd_i = q_r - q_{r-1}|\psi; S'_i = s'_{q_r}) \\ &\times \prod_{r=1}^{R} P(y_{q_{r-1}}|\phi; S'_i = s'_{q_r}). \end{aligned}$$

4. EXPERIMENTS

To test the proposed temporal segmentation modeling, sports videos have been treated, and more specifically squash videos. The data were taken from the "CVBASE'06" sports video database [2] which gives the squash videos (see Fig. 2 that presents one frame of the squash video) as well as the squash players respective coordinates in the images (*i.e.*, the video trajectories) and the game phases ("rally" and "passive" phases) to be used as ground truth for the results evaluation.

Fig. 2. A frame of a squash video (this whole squash video contains 15508 frames).

The first half of the squash video (about six minutes long) was used for training a hierarchical SMC with two states S'_1 and S'_2 corresponding to the two activity phases "rally" and "passive" (Fig. 4 shows the training result when fitting a GMM on the duration state distribution of the SMC "rally" state). The second half were used to test the proposed modelings. Fig. 3 presents the squash players trajectories of one squash video respectively used for training and testing.

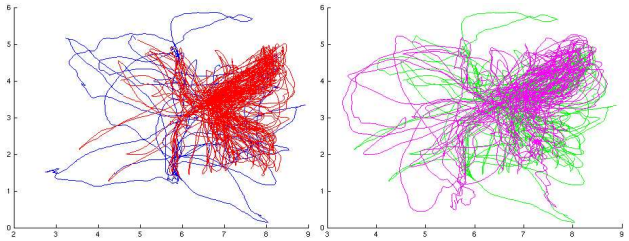

Fig. 3. Left: training trajectories corresponding to the trajectories of 2 squash players in the video plane (in blue and red) during the first half (about 5 minutes) of a squash video. Right: test trajectories corresponding to the trajectories of 2 squash players (in green and magenta) during the second half of the squash video.

The considered results were obtained using a P_v parameter value (as defined in Subsection 3.1) equal to 95%, a h parameter value (as defined in Subsection 2.1) fixed to the constant 3. Presented results corresponds to the best obtained ones when testing the method with a large range of N values.

These experiments are of great interest since it is hard, when only considering the players movements, to visually determine if the two squash players are in a "rally" phase or in a "passive" phase. Indeed, the players movements are often very reduced both in the "rally" phase (where the placement of the player is more important than its mobility) and in the "passive" phase. Furthermore, inside the "rally" phases, there are periods were the players are almost static, so that it looks like a "passive" phase. When trying to visually proceed a temporal segmentation, squash phases characterization can be done by using the relative distance evolution (*i.e.*, the evolution of the $\tilde{d}$ feature).

Hence, very satisfying results were obtained since precisions of about 88% of good phase segmentations were reached using the SMC modeling for the processed squash videos (see Fig. 5), retrieving the exact number of activity phases (*e.g.*, the number of played points) with little lags. Using a HMM having the same structure as the SMC (but with state durations not modeled by GMMs, *i.e.* state duration in state i follows a geometric law in regards of a'_{ii}) gave little less accurate segmentations (about 85% of good phase segmentation). Results below 70% of good phase segmentation were obtained when not considering the $\tilde{d}$ feature, highlighting the inherent information of the temporal trends of the distance (*i.e.*, of the interactions). Additional experiments (that can not be developped by lack of space) are also carried out to further assess the performances of the method with more video objects and activity phases, for example on basket-ball videos.

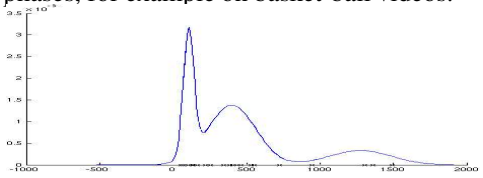

Fig. 4. Duration state density modeling using GMM for one considered SMC state (*i.e.* phase "rally"). The x-axis corresponds to the observed state durations.

5. CONCLUSION

This paper presents a SMC-based method for recognizing activities involving several video objects. Single video object behaviors as well as interacting processes are taken into account in the same framework. Feature are extracted and defined so that they are invariant to translation, rotation and scale transformation, hence providing an activity representation that may be independent of the considered video. The developed approach has been tested on large squash videos, with two interacting video objects and a two phases activity modeling, providing promising results. Extensions to more interacting video objects with a larger activity phases number are currently being investigated using the same framework.

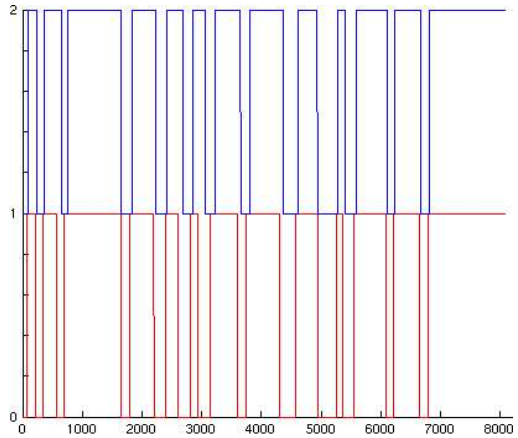

Fig. 5. Up: Result of a processed temporal segmentation results plotted in blue. The "1" and "2" values respectively correspond to the "passive" and "rally" phases. Down: Ground truth plotted in red, the "0" and "1" values here respectively correspond to the "passive" and "rally" phases. The x-axis corresponds to the frame index.

6. REFERENCES

[1] M. Brand, and N. Oliver. Coupled hidden Markov models for complex action recognition. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'96*, San Francisco, US, pages 994-999, Jun. 1996.

[2] http://vision.fe.uni-lj.si/cvbase06/downloads.html

[3] H. Denman, N. Rea, A. Kokaram. Content-based analysis for video from snooker broadcasts. *Computer Vision and Image Understanding*, 92(2-3):176-195, Dec. 2003.

[4] J. Ford and J. Moore. Adaptive estimation of HMM transition probabilities. *IEEE Trans. on Signal Processing*, 46(5):1374-1385, May 1998.

[5] B. Günsel, A. M. Tekalp, and P. J.L. van Beek. Temporal video segmentation using unsupervised clustering and semantic object tracking. *Journal of Electronic Imaging*, Special Issue, 7(3):592-604, July 1998.

[6] B. Günsel, A. M. Tekalp, and P. J.L. van Beek. Content-based access to video objects: temporal segmentation, visual summarization, and feature extraction. *Signal Processing*, Special Issue, 66(2):261-280, April 1998.

[7] A. Hervieu, P. Bouthemy, and J-P. Le Cadre. A HMM-based method for recognizing dynamic video contents from trajectories. *Proc. of the IEEE Int. Conf. on Image Processing, ICIP'07*, San Antonio, US, Sept. 2007.

[8] A. Hervieu, P. Bouthemy, and J-P. Le Cadre. Video event classification and detection using 2D trajectories. *Proc. of the Int. Conf. on Computer Vision Theory and Applications, VISAPP'08*, Madeira, Portugal, Jan. 2008.

[9] S. Hongeng, R. Nevatia, and F. Bremond. Video-based event recognition: Activity representation and probabilistic recognition methods. *Computer Vision and Image Understanding*, 96(2):129-162, 2003.

[10] A. Kokaram, N. Rea, R. Dahyot, M. Tekalp, P. Bouthemy, P. Gros, and I. Sezan. Browsing sports video (Trends in sports-related indexing and retrieval work). *IEEE Signal Processing Magazine*, 23(2):47-58, Mar. 2006.

[11] N. M. Oliver, B. Rosario, and A. P. Pentland. A Bayesian computer vision system for modeling human interactions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):831-843, Aug. 2000.