Human activity recognition in the semantic simplex of elementary actions

Beaudry Cyrille
cyrille.beaudry@univ-lr.fr
Péteri Renaud
renaud.peteri@univ-lr.fr
Mascarilla Laurent
laurent.mascarilla@univ-lr.fr

MIA lab.
University of La Rochelle
La Rochelle, France

Analyzing and recognizing human actions in videos has received considerable attention for many years in the computer vision community. Works on this topic are motivated by several potential applications (video monitoring, automatic video indexing, crowd analysis, human-machine interaction, etc). The wide variability of human actions makes it difficult to design generic methods. Many proposed approaches are based on discriminative supervised models [3, 4, 5, 6]. Other studies are focused on generative probabilistic models, and are based on LDA [2] or semi-supervised LDA [7]. However, generative models fail to match already known actions occurring in videos and it is moreover difficult to semantically analyze discovered topics.

In this paper, we present an original approach for human activities recognition in videos. It relies on a semantic representation of videos rather than a Bag of visual features approach, allowing better generalization. We characterize activities as temporal sequences of elementary actions by estimating their probabilities over time. Elementary actions are not discovered as in generative probabilistic models but learned via a robust action recognition method developed in our previous works [1]. Video sequences are characterized by critical points of optical flow and by their temporal trajectories. These features are computed at different spatio-temporal scales, using a dyadic subdivision of the sequence. A robust and generic representation of the elementary actions is then extracted using a cross-dataset learning process. A complex action is then decomposed as a sequence of elementary action proportions by transforming the decision boundaries of the classifier into probabilities. The classifier takes into account the case when there is no movement in the sequence.

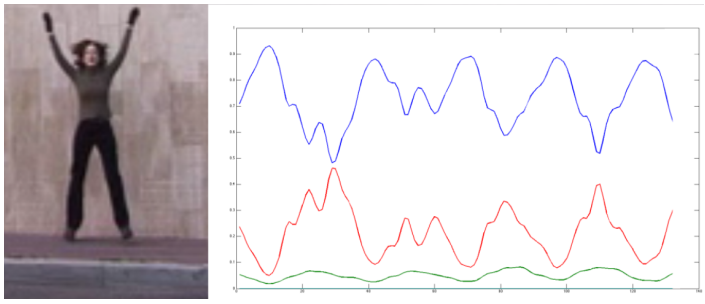

Figure 1: `Jack` action with its representation as a sequence of elementary action probabilities. The `Jack` action is composed of `Jump` and `Handwave` elementary actions. The graph represents the evolution of elementary action probabilities over time. Red curve is for `Handwave` action, blue curve for `Jump` action. The periodicity and alternation between the two elementary actions is well noticeable on the graph.

Once a frame is characterized by its L elementary action probabilities, its feature vector lies in a simplex $\mathcal{P}_L$ defined such as:
$\mathcal{P}_L = \{\pi \in R^{L+1} | \sum_{i=1}^{L+1} \pi_i = 1, \pi > 0\}$. $\mathcal{P}_L$ endowed with the Fisher information metric is a submanifold of R^{L+1}.
Figure 2 shows the global scheme for projecting activities in $\mathcal{P}_L$.

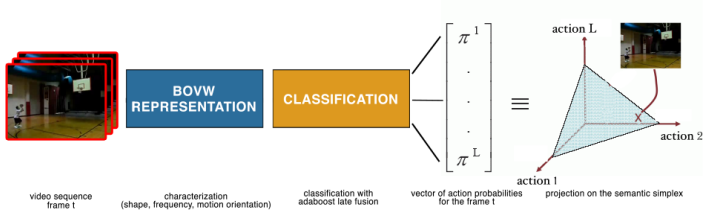

Figure 2: Global scheme for characterizing activities in the semantic simplex.

Activities are then represented as trajectories on the semantic simplex $\mathcal{P}_L$. By using a diffeomorphism F between P_L and the L-hypersphere $\mathcal{S}_L^+$, geodesic distances in P_L can be computed as shortest curves on $\mathcal{S}_L^+$. The geodesic distance between two points (π_{k1}, π_{k2}) of P_L is simply the great circle arc linking $(F(\pi_{k1}), F(\pi_{k2}))$ on S_L^+.
A robust descriptor of the trajectory shape in the simplex is obtained using the Fourier transform coefficients of the cumulative curvature in angular coordinates.
To test the discriminative performance of the proposed method, three complex actions from `UCF11`, `UCF50` datasets and `Olympic Sport` dataset are considered: `High-Jump`, `Basket-ball` and `Base-ball`. Figure 3 shows some trajectories in $\mathcal{S}_L^+$ for each activity class.

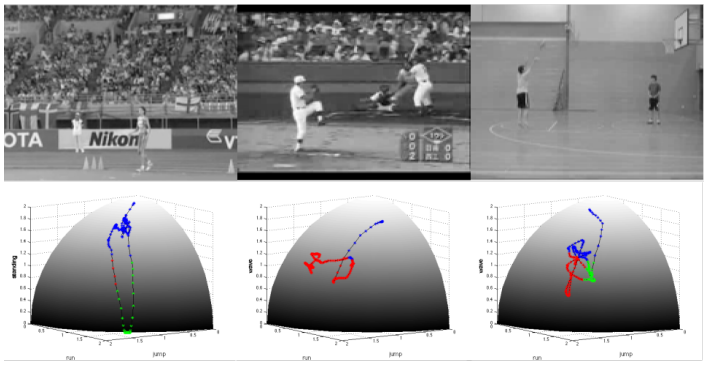

Figure 3: Examples of activity trajectories in $\mathcal{S}_L^+$: `High-Jump` activity, `Base-ball` activity, and `Basket-ball` activity.

A Leave-One-Out cross-validation test using a SVM with a RBF kernel gives a recognition rate of 96.6%, to be compared with the recognition rate of 86.6% using the STIP method [4]. The semantic aspect of our method also allows a better generalization of human activities.
Considering human actions as trajectories on the semantic manifold opens the way to different applications, such as video summary by computing a mean shape on the manifold.

[1] Cyrille Beaudry, Renaud Péteri, and Laurent Mascarilla. Action recognition in videos using frequency analysis of critical point trajectories. In *Proc. IEEE International Conference on Image Processing*, pages 1445–1449, Paris, France, 2014.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65 – 72, October 2005.

[4] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. Conf. Comp. Vision Pattern Rec.*, pages 1 –8, June 2008.

[5] M. Vrigkas, V. Karavasilis, C. Nikou, and A. Kakadiaris. Matching mixtures of curves for human action recognition. *Computer Vision and Image Understanding*, 119(0):27 – 40, 2014.

[6] H. Wang, A. Klaser, C. Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Proc. Conf. Comp. Vision Pattern Rec.*, pages 3169 –3176, June 2011.

[7] Y. Wang, P. Sabzmeydani, and G. Mori. Semi-latent dirichlet allocation: A hierarchical model for human action recognition. In *Proceedings of the 2nd Conference on Human Motion*, pages 240–254, Berlin, Heidelberg, 2007. Springer-Verlag.