Evaluation of local spatio-temporal features for action recognition

Heng Wang[1,3]
hwang@nlpr.ia.ac.cn
Muhammad Muneeb Ullah[2]
Muhammad.Muneeb.Ullah@inria.fr
Alexander Kläser[1]
Alexander.Klaser@inria.fr
Ivan Laptev[2]
Ivan.Laptev@inria.fr
Cordelia Schmid[1]
Cordelia.Schmid@inria.fr

[1] LEAR, INRIA, LJK
Grenoble, France
[2] VISTA, INRIA
Rennes, France
[3] LIAMA, NLPR, CASIA
Beijing, China

Local space-time features have recently become a popular video representation for action recognition. Several methods for feature localization and description have been proposed in the literature, and promising recognition results were demonstrated for different action datasets. The comparison of those methods, however, is limited given the different experimental settings and various recognition methods used.

The purpose of this paper is first to define a common evaluation setup to compare local space-time detectors and descriptors. All experiments are reported for the same bag-of-features SVM recognition framework. Second, we provide a systematic evaluation of different spatio-temporal features. We evaluate the performance of several space-time interest point detectors and descriptors along with their combinations on datasets with varying degree of difficulty. We also include a comparison with dense features obtained by regular sampling of local space-time patches.

Feature detectors. In our experimental evaluation, we consider the following feature detectors.

(1) The *Harris3D* detector [3] extends the Harris detector for images to image sequences. At each video point, a spatio-temporal second-moment matrix μ is computed using a separable Gaussian smoothing function and space-time gradients. Interest points are located at local maxima of $H = \det(\mu) - k\,\mathrm{trace}^3(\mu)$.

(2) The *Cuboid* detector [1] is based on temporal Gabor filters. The response function has the form: $R = (I * g * h_{ev})^2 + (I * g * h_{od})^2$, where $g(x,y;\sigma)$ is the 2D Gaussian smoothing kernel, and h_{ev} and h_{od} are 1D Gabor filters. Interest points are detected at local maxima of R.

(3) The *Hessian* detector [6] is a spatio-temporal extension of the Hessian saliency measure. The determinant of the 3D Hessian matrix is used to measure saliency. The determinant of the Hessian is computed over several spatial and temporal scales. A non-maximum suppression algorithm selects extrema as interest points.

(4) *Dense sampling* extracts multi-scale video blocks at regular positions in space and time and for varying scales. In our experiments, we sample cuboids with 50% spatial and temporal overlap.

Feature descriptors. The following feature descriptors are investigated.

(1) For the *Cuboid* descriptor [1], gradients computed for each pixel in a cuboid region are concatenated into a single vector. PCA projects vectors to a lower dimensional space.

(2) The *HOG/HOF* descriptors [4] divide a cuboid region into a grid of cells. For each cell, 4-bin histograms of gradient orientations (*HOG*) and 5-bin histograms of optic flow (*HOF*) are computed. Normalized histograms are concatenated into HOG, HOF as well as HOG/HOF descriptor vectors.

(3) The *HOG3D* descriptor [2] is based on histograms of 3D gradient orientations. Gradients are computed via an integral video representations. Regular polyhedrons are used to uniformly quantize the orientation of spatio-temporal gradients. A given 3D volume is divided into a grid of cells. The corresponding descriptor concatenates gradient histograms of all cells.

(4) The *extended SURF (ESURF)* descriptor [6] extends the image SURF descriptor to videos. Again 3D cuboids are divided into a grid of cells. Each cell is represented by a weighted sum of uniformly sampled responses of Haar-wavelets aligned with the three axes.

Experimental Setup. We represent video sequences as a bag of local spatio-temporal features [5]. Spatio-temporal features are first quantized into visual words and videos are consequently represented as frequency histograms over the visual words. A non-linear support vector machine with χ^2-kernel [4] classifies the histograms.

	HOG3D	HOG/HOF	HOG	HOF	Cuboids	ESURF
Harris3D	89.0%	**91.8%**	80.9%	**92.1%**	–	–
Cuboids	**90.0%**	88.7%	82.3%	88.2%	89.1%	–
Hessian	84.6%	88.7%	7767%	88.6%	–	81.4%
Dense	85.3%	86.1%	79.0%	88.0%	–	–

Table 1: Average accuracy on the KTH actions dataset.

	HOG3D	HOG/HOF	HOG	HOF	Cuboids	ESURF
Harris3D	79.7%	78.1%	71.4%	75.4%	–	–
Cuboids	**82.9%**	77.7%	72.7%	76.7%	76.6%	–
Hessian	79.0%	79.3%	66.0%	75.3%	–	77.3%
Dense	**85.6%**	81.6%	77.4%	**82.6%**	–	–

Table 2: Average accuracy on the UCF sports dataset.

	HOG3D	HOG/HOF	HOG	HOF	Cuboids	ESURF
Harris3D	43.7%	45.2%	32.8%	43.3%	–	–
Cuboids	45.7%	**46.2%**	39.4%	42.9%	45.0%	–
Hessian	41.3%	**46.0%**	36.2%	43.0%	–	38.2%
Dense	45.3%	**47.4%**	39.4%	45.5%	–	–

Table 3: Mean average precision (mAP) on the Hollywood2 dataset.

Datasets that have been used in this evaluation are the KTH actions, the UCF sport actions, and the Hollywood2 actions dataset. The classification results for these datasets and different combinations of detectors and descriptors are presented in Tables 1-3. The best three combinations of feature detector and descriptor are highlighted.

Conclusions. Experimental results show that dense sampling consistently outperforms all tested interest point detectors in realistic settings. Note, however, that dense sampling also produces a very large number of features (usually 15-20 times more than feature detectors). This is more difficult to handle than the relatively sparse number of interest points. The performance of interest point detectors seems to be rather similar across datasets. Harris 3D performs better on the KTH dataset, while the Cuboid detector gives better results for UCF and Hollywood2 datasets.

Among tested descriptors, the combination of gradient-based and optical flow-based descriptors seems to be a good choice. The combination of dense sampling with the HOG/HOF descriptor provides best results for the most challenging Hollywood2 dataset. On the UCF dataset, the HOG3D descriptor performs best in combination with dense sampling.

[1] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, 2005.

[2] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In *BMVC*, 2008.

[3] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, 2003.

[4] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[5] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *ICPR*, 2004.

[6] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *ECCV*, 2008.