Evaluation of local descriptors for action recognition in videos

Piotr Bilinski, François Bremond

Evaluation of local descriptors for action recognition in videos

Piotr Bilinski and Francois Bremond

INRIA Sophia Antipolis - PULSAR group
2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex, France
firstname.surname@inria.fr

Abstract. Recently, local descriptors have drawn a lot of attention as a representation method for action recognition. They are able to capture appearance and motion. They are robust to viewpoint and scale changes. They are easy to implement and quick to calculate. Moreover, they have shown to obtain good performance for action classification in videos. Over the last years, many different local spatio-temporal descriptors have been proposed. They are usually tested on different datasets and using different experimental methods. Moreover, experiments are done making assumptions that do not allow to fully evaluate descriptors. In this paper, we present a full evaluation of local spatio-temporal descriptors for action recognition in videos. Four widely used in state-of-the-art approaches descriptors and four video datasets were chosen. HOG, HOF, HOG-HOF and HOG3D were tested under a framework based on the bag-of-words model and Support Vector Machines.

1 Introduction

In last years, many researchers have been working on developing effective descriptors to recognize objects, scenes and human actions. Many suggested descriptors have proven to establish very good performance for action classification in videos. They are able to capture appearance and motion. They are robust to viewpoint and scale changes. Moreover, they are easy to implement and quick to calculate. For example [15] proposed Scale-Invariant Feature Transform (SIFT) descriptor, [1] proposed Speeded Up Robust Features (SURF), [3] proposed Histogram of Oriented Gradients (HOG) descriptor. [8] proposed PCA-SIFT, [4] proposed Cuboid descriptor, [13] proposed Histogram of Oriented Gradients (HOG) and Histogram of Optical Flow (HOF) descriptors computed on spatio-temporal grids. [9] proposed a Spatio-Temporal Descriptor based on 3D Gradients (HOG3D). Although much has been done, it is not clear which descriptors are better than the others. They are usually evaluated on different datasets and using different experimental methods. Moreover, existing comparisons usually involve smaller or bigger restrictions. For example, [21] to limit the complexity choose a subset of $100,000$ selected training features what is around 21% of all HOG-HOF descriptors and around 40% of all HOG3D descriptors computed for

the KTH Dataset. Moreover, the authors make all the experiments using only one codebook size (4000). Also in [19] the authors make an assumption about the codebook size (one codebook size) and evaluate descriptors on one dataset.

In this paper, we present an evaluation of local spatio-temporal features for action recognition in videos. Four widely used in the state-of-the-art approaches descriptors were chosen: HOG, HOF, HOG-HOF and HOG3D. All these descriptors are tested on the same datasets with the same split of training and testing data, and using the same identical classification method. Our evaluation framework is based on the bag-of-words approach, an approach that is very often used together with local features. Computed descriptors are quantized into visual words and videos are represented as histograms of occurrences of visual words. For action classification, non-linear Support Vector Machines (SVM) together with leave-one-out cross-validation technique are used. Our experiments are performed on several public datasets containing both low and high resolution videos recorded using static and moving camera (KTH Dataset, Weizmann Action Dataset, ADL Dataset and Keck Dataset). In contrast to other evaluations, we test all the computed descriptors, we perform evaluation on several differing in difficulty datasets and perform evaluation on several codebook sizes. We demonstrate that accuracy of evaluated descriptors depends on the codebook size and a dataset.

The paper is organized as follows. In section 2, we briefly present the main idea of our evaluation framework. Section 3, presents our experiments and obtained results. Finally, in section 4, we present our conclusion.

2 Evaluation Framework

Our evaluation framework is as follows. In the first step, for each video, local space-time detector is applied. For each obtained point, local space-time descriptor is computed (Section 2.1). In the second step, the bag-of-words model is used to represent actions (Section 2.2). For each video, four different codebooks and four different video representations are computed. Finally, to evaluated descriptors, the leave-one-person-out cross-validation technique and non-linear multi-class Support Vector Machine are applied (Section 2.3 and 2.4). To speed-up the evaluation process, clusters of computers are used.

2.1 Space-Time Local Features

Local spatio-temporal features are extracted for each video. As a local feature detector, the Harris3D algorithm is applied. Then, for each detected feature, four types of descriptors are computed (HOG, HOF, HOG-HOF and HOG3D). The detector and descriptors were selected based on their use in the literature and availability of the original implementation[1][2]. For each algorithm, the default values of parameters were used.

[1] `http://www.irisa.fr/vista/Equipe/People/Laptev/download.html`
[2] `http://lear.inrialpes.fr/people/klaeser/software_3d_video_descriptor/`

Harris3D [11] detector - it is proposed by Laptev and Linderberg extension of the Harris corner detector [6]. The authors propose to extend the notion of spatial interest and detect local structures in space-time where the image values have significant local variations in both space and time. The authors use independent spatial and temporal scale values, a separable Gaussian smoothing function, and space-time gradients.

Histogram of Oriented Gradients (HOG) [13] - it is a 72-bins descriptor describing the local appearance. The authors propose to define a grid $n_x \times n_y \times n_t$ (default settings: $3 \times 3 \times 2$) in the surrounding space-time area and compute for each cell of the grid 4-bins histogram of oriented gradients.

Histogram of Optical Flow (HOF) [13] - it is a 90-bins descriptor describing the local motion. The authors propose to define a grid $n_x \times n_y \times n_t$ (default settings: $3 \times 3 \times 2$) around the encompassing space-time area and compute for each cell of the grid 5-bins histogram of optical flow.

HOG-HOF descriptor - it is a 162-bin descriptor combining both Histogram of Oriented Gradients and Histogram of Oriented Flow descriptors.

Spatio-Temporal Descriptor based on 3D Gradients (HOG3D) - it is a 300-bins descriptor proposed by Klaser et al. [9]. It is based on orientation histograms of 3D gradients. The authors propose to define a grid $n_x \times n_y \times n_t$ (default settings: $2 \times 2 \times 5$) in the surrounding space-time area and compute for each cell of the grid 3D gradients orientations.

2.2 Bag-of-words Model

To represent videos using local features we apply common bag-of-words model. All computed descriptors for all Harris3D detected points are used in the quantization process. First of all, the k-means clustering algorithm with the Euclidean distance is used to create a codebook. Then, each video is represented as a histogram of occurrences of the codebook elements. In our experiments we use four different sizes of codebooks (1000, 2000, 3000 and 4000).

2.3 Classification

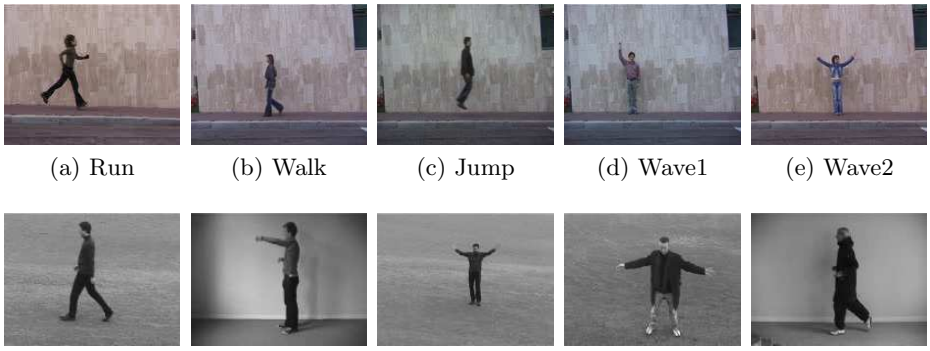
In order to perform classification, we use a multi-class non-linear Support Vector Machines using radial basis function defined by:

$$K(H_a, H_b) = exp(-\gamma D(H_a, H_b)) \tag{1}$$

where both $H_a = \{h_{a1}, ..., h_{an}\}$ and $H_b = \{h_{b1}, ..., h_{bn}\}$ are n-bins histograms. Function D is a χ^2 distance function defined by:

$$D(H_a, H_b) = \sum_{i=1}^{n} \frac{(h_{ai} - h_{bi})^2}{h_{ai} + h_{bi}} \tag{2}$$

Such defined kernel requires two parameters: (a) trade-off between training error and margin, and (b) parameter gamma in the rbf kernel. To evaluate



(a) Run (b) Walk (c) Jump (d) Wave1 (e) Wave2

(f) Walking (g) Boxing (h) H.-waving (i) H.-clapping (j) Jogging

Fig. 1: A few sample frames from video sequences from Weizmann (the first row) and KTH (the second row) datasets.

selected descriptors, we test: (a) all values 2^x where x is in range -5 to 16 (22 different values) for the trade-off between training error and margin, and (b) all values 2^y where y is in range 3 to -15 (19 different values) for the parameter gamma in the rbf kernel. As a Support Vector Machine, the SVM multi-class [20] is used (multi-class variant of SVM light [7]). To speed-up the evaluation process, clusters of computers are used.

2.4 Evaluation

To evaluate selected descriptors, we use leave-one-person-out cross-validation technique (unless specified for a dataset), where videos of one actor are used as the validation data and videos from the remaining actors as the training data. This is repeated in such a way that videos from one person are used exactly once as the validation data. In our experiments we use all the descriptors calculated for all detected points to comprehensively try out the effectiveness of used local feature descriptors.

3 Experiments

Our experiments are performed on four different datasets: Weizmann Action Recognition Dataset (Section 3.1), KTH Dataset (Section 3.2), ADL Dataset (Section 3.3) and Keck Dataset (Section 3.4). A few sample frames from these video datasets can be found in Figure 1 and Figure 2. These datasets contain various types of videos: low and high resolution videos, recorded using static and moving camera, and containing one and many people. Information about these databases are summarized in Table 1.

	Weizmann	KTH	ADL	Keck
resolution	180×144	160×120	640×360	640×480
#videos	93	599	150	98
#frames	6,108	289,715	72,729	25,457
#Harris3D points	13,259	473,908	718,440	227,310
#HOG-HOF descriptors	13,259	473,908	718,440	227,310
#HOG3D descriptors	9,116	252,014	690,907	207,655
#frames/#videos	65.68	483.66	484.86	259.77
#points/#videos	142.57	791.17	4789.60	2319.49
#HOG-HOF/#videos	142.57	791.17	4789.60	2319.49
#HOG3D/#videos	98.02	505.04	4606.05	2118.93
#points/#frames	2.17	1.64	9.88	8.93
#HOG-HOF/#frames	2.17	1.64	9.88	8.93
#HOG3D/#frames	1.49	0.87	9.50	8.16
#HOG-HOF/#points	1.00	1.00	1.00	1.00
#HOG3D/#points	0.69	0.53	0.96	0.91

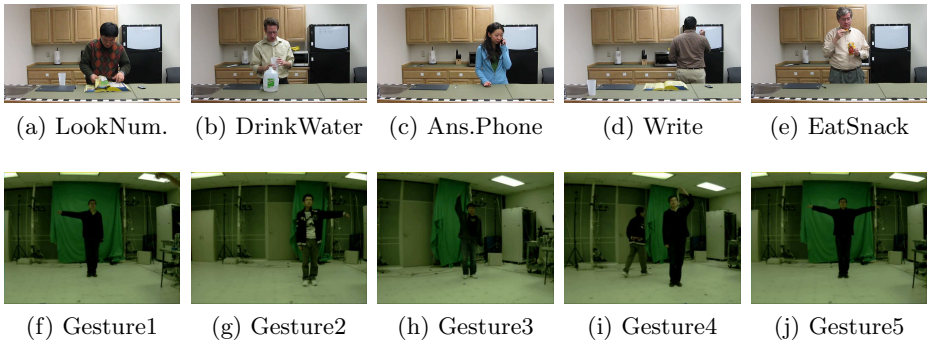
Table 1: Statistics for the Weizmann, KTH, ADL and Keck datasets: number of videos, frames, detected points, HOG-HOF (HOG, HOF) descriptors, HOG3D descriptors, frames per video, detected points per video, HOG-HOF descriptors per video, HOG3D descriptors per video, points per frame, HOG-HOF descriptors per frame, HOG3D descriptors per frame, ratio of number of HOG-HOF descriptors to number of points, and ratio of number of HOG3D descriptors to number of points.

3.1 Weizmann Action Recognition Dataset

The Weizmann Action Recognition Dataset [2, 5][3] is a low-resolution (180×144 pixel resolution, 50 fps) dataset of natural human actions. The dataset contains 93 video sequences showing 9 different people. The dataset contains 10 actions. The full list of actions is: run, walk, skip, jumping-jack (shortly jack), jump-forward-on-two-legs (shortly jump), jump-in-place-on-two-legs (shortly pjump), gallop-sideways (shortly side), wave-two-hands (shortly wave2), wave-one-hand (shortly wave1), and bend. Statistics about this dataset are available in table 1. Evaluation is done using leave-one-person-out cross-validation technique.

Results are presented in table 2. As we can observe, all the descriptors obtain the same accuracy for codebook 2000 and 4000. In this case, the codebook of size 2000 is preferred (faster codebook computation and faster SVM classification). The HOG descriptor performs the best for codebook 2000, the HOF descriptor for codebook 3000, HOG-HOF for codebook 2000 and HOG3D descriptor for codebook 3000. According to the results, the HOG-HOF is the best descriptor for the Weizmann dataset and the HOG descriptor is the worst. Ranking is: HOG-HOF > HOF = HOG3D > HOG. The HOF descriptor obtains the same classification accuracy as HOG3D descriptor but HOF descriptor is smaller in

[3] `http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html`



(a) LookNum. (b) DrinkWater (c) Ans.Phone (d) Write (e) EatSnack

(f) Gesture1 (g) Gesture2 (h) Gesture3 (i) Gesture4 (j) Gesture5

Fig. 2: A few sample frames from video sequences from ADL (the first row) and Keck (the second row) datasets.

size (90-bins descriptor instead of 300-bins). It takes less time to compute codebook and perform classification for the HOF descriptor. Kläser [10], employing random sampling on training features for codebook generation (codebook size 4000), obtained 75.3% accuracy for the HOG descriptor, 88.8% for the HOF descriptor, 85.6% for the HOG-HOF and 90.7% for the HOG3D descriptor. This shows that the codebook selection method has significant importance to the effectiveness of the BOW method (we obtained up to 10.72% better results).

	HOG	HOF	HOG-HOF	HOG3D
codebook size 1000	83.87%	88.17%	91.40%	89.25%
codebook size 2000	**86.02%**	90.32%	**92.47%**	90.32%
codebook size 3000	86.02%	**91.40%**	91.40%	**91.40%**
codebook size 4000	86.02%	90.32%	92.47%	90.32%

Table 2: Action recognition accuracy for the Weizmann dataset.

3.2 KTH Dataset

The KTH dataset [18][4] contains six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed several times by 25 subjects in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. The dataset contains 599 videos. All videos were taken over homogeneous backgrounds with a static camera with 25 fps. The sequences were down-sampled by the authors to the spatial resolution of 160×120 pixels. For this dataset, as it is recommended by the authors, we

[4] `http://www.nada.kth.se/cvap/actions/`

divide the dataset into testing part (person 2, 3, 5, 6, 7, 8, 9, 10 and 22) and training part (other video sequences). Statistics about this dataset are available in the table 1.

Results are presented in table 3. Both HOG and HOF descriptors perform the best for codebook 1000 and both HOG-HOF and HOG3D descriptors for codebook 3000. According to the results, the HOF descriptor is superior descriptor for the KTH dataset and again the HOG descriptor is inferior quality. Ranking is: HOF > HOG-HOF > HOG3D > HOG. Wang et al. [21], choosing a subset of 100,000 selected training features and using codebook size of 4000, obtained 80.9% accuracy for the HOG descriptor, 92.1% for the HOF descriptor, 91.8% for the HOG-HOF and 89% for the HOG3D descriptor. We obtain up to 4.52% better results on this dataset. Selecting only a subset of descriptors can cause loss of some important information.

	HOG	HOF	HOG-HOF	HOG3D
codebook size 1000	**83.33%**	**95.37%**	93.06%	91.66%
codebook size 2000	83.33%	94.44%	93.98%	92.13%
codebook size 3000	83.33%	94.91%	**94.44%**	**93.52%**
codebook size 4000	82.41%	94.91%	93.98%	93.06%

Table 3: Action recognition accuracy for the KTH dataset.

3.3 ADL Dataset

The University of Rochester Activities of Daily Living (ADL) dataset [16][5] is a high-resolution (1280 × 720 pixel resolution, 30 fps) video dataset of activities of daily living. The dataset contains 150 video sequences showing five different people. The dataset contains ten activities. The full list of activities is: answering a phone, dialling a phone, looking up a phone number in a telephone directory, writing a phone number on a whiteboard, drinking a glass of water, eating snack chips, peeling a banana, eating a banana, chopping a banana, and eating food with silverware. These activities were selected to be difficult to separate on the basis of single source of information (e.g. eating banana and eating snack or answering a phone and dialling a phone). These activities were each performed three times by five people differing in shapes, sizes, genders, and ethnicity. The videos were down-sampled to the spatial resolution of 640×360 pixels. Statistics about this dataset are available in table 1. Evaluation is done using leave-one-person-out cross-validation technique.

Results are presented in table 4. As we can observe, apart from the HOG descriptor, all the other descriptors perform the best for codebook 1000. The HOG descriptor performs the best for codebook 2000. According to the results, the

[5] `http://www.cs.rochester.edu/~rmessing/uradl/`

HOG-HOF is the best descriptor for the ADL dataset and the HOG descriptor is again the worst. Ranking is: HOG-HOF > HOG3D > HOF > HOG.

	HOG	HOF	HOG-HOF	HOG3D
codebook size 1000	85.33%	**90.00%**	**94.67%**	**92.00%**
codebook size 2000	**88.67%**	90.00%	92.67%	91.33%
codebook size 3000	83.33%	89.33%	94.00%	90.67%
codebook size 4000	86.67%	89.33%	94.00%	85.00%

Table 4: Action recognition accuracy for the ADL dataset.

3.4 Keck Dataset

The Keck gesture dataset [14][6] consists of 14 different gesture classes, which are a subset of military signals. The full list of activities is: Turn left, Turn right, Attention left, Attention right, Attention both, Stop left, Stop right, Stop both, Flap, Start, Go back, Close distance, Speed up, Come near. The dataset is collected using a color camera with 640×480 resolution. Each gesture is performed by 3 people. In each sequence, the same gesture is repeated 3 times by each person. Hence there are $3 \times 3 \times 14 = 126$ video sequences for training which are captured using a fixed camera with the person viewed against a simple, static background. There are $4 \times 3 \times 14 = 168$ video sequences for testing which are captured from a moving camera and in the presence of background clutter and other moving objects. Statistics about this dataset are available in table 1.

Results are presented in table 5. The HOG descriptor performs the best for codebook 3000, the HOF descriptor for codebook 4000, the HOG-HOF for codebook 2000 and the HOG3D for codebook 3000. The HOG3D is the best descriptor for the Keck dataset and the HOF descriptor is the worst. Ranking is: HOG3D > HOG-HOF > HOG > HOF.

	HOG	HOF	HOG-HOF	HOG3D
codebook size 1000	42.86%	30.36%	37.50%	50.00%
codebook size 2000	39.29%	33.93%	**46.43%**	50.00%
codebook size 3000	**44.64%**	37.50%	39.29%	**53.57%**
codebook size 4000	41.07%	**42.86%**	44.64%	44.64%

Table 5: Action recognition accuracy for the Keck dataset.

[6] `http://www.umiacs.umd.edu/~zhuolin/Keckgesturedataset.html`

According to the obtained results, we observe that accuracy of descriptors depends on the codebook size (12.5% difference on the Keck dataset for the HOF descriptor, 7% difference on the ADL dataset for the HOG3D descriptor), codebook selection method (up to 10.72% better results comparing to [10] on the Weizmann dataset) and dataset (HOF descriptor obtains 95.37% on the KTH dataset but only 42.86% on the Keck dataset). Also, we observe that smaller codebook sizes (1000, 2000, 3000) are found to lead to consistently good performance across the different datasets. Due to random initialization of k-means used for codebook generation, we observe no linear relationship accuracy of codebook size.

Our experiments show that the HOG-HOF, combination of gradient and optical flow based descriptors, seems to be a good descriptor. For the Weizmann and ADL datasets, the HOG-HOF descriptor performs best and takes the second place for the KTH and Keck datasets. The HOG descriptor usually perform the worst. The accuracy of the HOF and HOG3D descriptors depends on a dataset. Also, we observe that regardless of the dataset, the HOG-HOF and HOG3D descriptors always work better than the HOG descriptor.

4 Conclusions

In this paper, we present a full evaluation of local spatio-temporal descriptors for action recognition in videos. Four widely used in state-of-the-art approaches descriptors (HOG, HOF, HOG-HOF and HOG3D) were chosen and evaluated under the framework based on the bag-of-words approach, non-linear Support Vector Machine and leave-one-out cross-validation technique. Our experiments are performed on four public datasets (KTH Action Dataset, Weizmann Action Dataset, ADL Dataset and Keck Dataset) containing low and high resolution videos recorded by static and moving cameras. In contrast to other existing evaluations, we test all the computed descriptors, perform evaluation on several differing in difficulty datasets and perform evaluation on several codebook sizes.

Acknowledgements. This work was supported by the Région Provence-Alpes-Côte d'Azur and partly by the Sweet-Home, Video-Id, ViCoMo, Vanaheim, and Support projects. However, the views and opinions expressed herein do not necessarily reflect those of the financing institutions.

References

1. Bay H., Ess A., Tuytelaars T., Gool L.V.: SURF: Speeded Up Robust Features. In: Computer Vision and Image Understanding, 2008.
2. Blank M., Gorelick L., Shechtman E., Irani M., Basri R.: Actions as Space-Time Shapes. In: International Conference on Computer Vision, 2005.
3. Dalal N., Triggs B.: Histograms of Oriented Gradients for Human Detection. In: IEEE Conference on Computer Vision and Pattern Recognition, 2005.

4. Dollar P., Rabaud V., Cottrell G., Belongie S.: Behavior recognition via sparse spatio-temporal features. In: Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, in conjunction with ICCV, 2005.
5. Gorelick L., Blank M., Shechtman E., Irani M., Basri R.: Actions as Space-Time Shapes. In: Transactions on Pattern Analysis and Machine Intelligence, 2007.
6. Harris C., Stephens M.: A Combined Corner and Edge Detector. In: Alvey Vision Conference, 1988.
7. Joachims T.: Making Large-Scale SVM Learning Practical. In: Advances in Kernel Methods - Support Vector Learning, 1999.
8. Ke Y., Sukthankar R.: PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. In: IEEE Conference on Computer Vision and Pattern Recognition, 2004.
9. Kläser A., Marszałek M., Schmid C.: A Spatio-Temporal Descriptor Based on 3D-Gradients. In: British Machine Vision Conference, 2008.
10. Kläser A.:Learning human actions in video. In: PhD thesis, Université de Grenoble, 2010.
11. Laptev I., Lindeberg T.: Space-Time Interest Points. In: International Conference on Computer Vision, 2003.
12. Laptev I.: On Space-Time Interest Points. In: International Journal of Computer Vision, 2005.
13. Laptev I., Marszałek M., Schmid C., Rozenfeld B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008.
14. Lin Z., Jiang Z., Davis L.S.: Recognizing Actions by Shape-Motion Prototype Trees. In: International Conference on Computer Vision, 2009.
15. Lowe D.G.: Distinctive Image Features from Scale-Invariant Keypoints. In: International Journal of Computer Vision, 2004.
16. Messing R., Pal C., Kautz H.: Activity recognition using the velocity histories of tracked keypoints. In: International Conference on Computer Vision, 2009.
17. Rosten E., Drummond T.: Machine learning for high-speed corner detection. In: European Conference on Computer Vision, 2006.
18. Schuldt C., Laptev I., Caputo B.: Recognizing Human Actions: A Local SVM Approach. In: International Conference on Pattern Recognition, 2004.
19. Stöttinger J., Goras B.T., Pönitz T., Sebe N., Hanbury A., Gevers T.: Systematic Evaluation of Spatio-temporal Features on Comparative Video Challenges. In: International Workshop on Video Event Categorization, Tagging and Retrieval, in conjunction with ACCV, 2010.
20. Tsochantaridis I., Joachims T., Hofmann T., Altun Y.: Large Margin Methods for Structured and Interdependent Output Variables. In: Journal of Machine Learning Research, 2005.
21. Wang H., Ullah M.M., Kläser A., Laptev I., Schmid C.: Evaluation of local spatio-temporal features for action recognition. In: British Machine Vision Conference, 2009.
22. Willems G., Tuytelaars T., Gool L.V.: An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector. In: European Conference on Computer Vision, 2008.