International Journal of Innovation and Applied Studies
ISSN 2028-9324 Vol. 9 No. 4 Dec. 2014, pp. 1708-1717

http://www.ijias.issr-journals.org/

Evaluation of Local Space-time Descriptors based on Cuboid Detector in Human Action Recognition

Haiam A. Abdul-Azim[1], Elsayed E. Hemayed[2], and Magda B. Fayek[2]

[1]Physics Department,
Faculty of Women for Arts, Science and Education, Ain Shams University,
Cairo, Egypt

[2]Computer Engineering Department,
Faculty of Engineering, Cairo University,
Giza, Egypt

ABSTRACT: Human action recognition remains a challenging problem for researchers. Several action representation approaches have been proposed to improve the action recognition performance. Recently, local space-time features have become a popular representation approach for human actions in video sequences. Many different space-time detectors and descriptors have been proposed. They are evaluated on different datasets using different experimental conditions. In this paper, the performance of Cuboid detector is evaluated with four space-time description methods; namely, Gradient, HOG, HOF and HOG-HOF. All descriptors were tested on two datasets (KTH and Weizmann) using the bag-of-words model and Support Vector Machine.

KEYWORDS: Space-time features; Cuboid detector; space-time feature descriptors; bag-of-words; human action recognition.

1 INTRODUCTION

With the wide use of digital cameras, video sequences have become an important source of information in our life. More and more videos are generated for many applications, e.g. internet sharing, traffic monitoring, health care, security surveillance of public places, etc. Manually extracting and analyzing this information are exhausting and time consuming. Recently, computer vision researchers have shown interest in automatically recognizing human actions in video sequences. In Aggarwal and Ryoo [1], the term "action" refers to a simple human activity that has been carried out for a period of time in video, such as walking, running and waving. Much research in human action recognition has been done; however, it is still a challenging problem. The challenge is how to recognize human actions under cluttered backgrounds, illumination changes, different physiques of humans, variety of clothing, camera motion, partial occlusions, viewpoint changes, scale variation of video screen, etc.

In the survey paper by Weinland et al. [2], human action representation approaches are classified into two main approaches: global and local representations. Two vision tasks are required for each representation approach: First, detection of features (local or global) from the video; second, description of motion using the detected features. Global representation approaches focus on detecting the whole body of the person by using background subtraction or tracking. Silhouettes, contours or optical flow are usually used for representing the localized person. These representations are more sensitive to viewpoint changes, personal appearance variations and partial occlusions.

In local representation approaches, videos are represented as a collection of small independent patches (named cuboids). These patches involve the regions of high variations in both the spatial and the temporal domain in the video [3]. Centres of the cuboids are called space-time interest points. After detecting the interest points, the cuboids are described and a model

based on independent features (Bag of Words) is built. Due to its invariant to viewpoint changes, the person's appearance variations and partial occlusions, space-time local features have become a popular representation approach for human action recognition. Different space-time interest point detectors have been proposed such as Harris3D detector [3], Cuboid detector [4], Hessian detector [5]. They differ in the type and sparsity of selected points. Also, several descriptors have been proposed to capture appearance and/or motion cues from the cuboids. For example [4] proposed Gradient descriptor, [6] proposed Histogram of Oriented Gradients (HOG) and Histogram of Optical Flow (HOF) descriptors, [7] proposed 3D Scale-Invariant Feature Transform (3D SIFT), [8] proposed a Spatio-Temporal Descriptor based on 3D Gradients descriptor (HOG3D) and [5] proposed the extended Speeded Up Robust Features descriptor (ESURF).

Several evaluations of local space-time detectors and descriptors were reported [4], [8]–[11] to find the best combinations that can be used in the human action recognition framework. Dollár et al. [4] evaluated the Cuboid detector with brightness, gradient and optical flow descriptors. Wang et al.[9] evaluated the detectors HOG/HOF, HOG3D and ESURF with the descriptors Harris3D, Cuboid, Hessian and Dense sampling. All of these evaluations were carried out on different datasets and using different experimental methods.

In this paper, we evaluate the performance of Cuboid detector with four widely used space-time descriptors: Gradient, HOG, HOF, and HOG-HOF. All experiments are performed on two datasets: KTH, and Weizmann. The experiments are based on the bag-of-words approach to find the correlations between cuboids and the use of non-linear Support Vector Machine (SVM) for classification.

The paper is organized as follows. In section two, the feature detection and feature description methods used in evaluation are summarized, respectively. In section three, the evaluation framework is illustrated. In section four, the experiments and their results are discussed in detail. Finally, section five presents the conclusion.

2 LOCAL SPACE-TIME FEATURES

This section describes the applied space-time interest point detector (cuboid detector) and the four selected descriptors (Gradient, HOG, HOF and HOG-HOF). The selection of descriptors is based on their previous evaluation in Wang et al.[9]. All detection and description methods are implemented in the Matlab environment.

2.1 CUBOID DETECTOR

Dollár et al.[4] proposed the Cuboid detector based on applying a set of separable linear filters (Gaussian filter applied on the spatial domain and 1-D Gabor filter applied temporally). The response function for a stack of images denoted by *I(x,y,t)* is given by

$$\boldsymbol{R} = (\boldsymbol{I} * \mathbf{g} * \boldsymbol{h_{even}})^2 + (\boldsymbol{I} * \mathbf{g} * \boldsymbol{h_{odd}})^2 \qquad \textbf{(1)}$$

where *g(x,y;σ)* is the 2D spatial Gaussian smoothing kernel, and $\boldsymbol{h_{even}}$ and $\boldsymbol{h_{odd}}$ are a quadrature pair of 1D Gabor filters which are applied temporally. They are defined as

$$\boldsymbol{h_{even}}(\boldsymbol{t};\boldsymbol{\tau},\boldsymbol{\omega}) = -\mathbf{cos}(\mathbf{2}\boldsymbol{\pi t\omega})\boldsymbol{e}^{-\boldsymbol{t}^2/\boldsymbol{\tau}^2}$$

and $$\boldsymbol{h_{odd}}(\boldsymbol{t};\boldsymbol{\tau},\boldsymbol{\omega}) = -\mathbf{sin}(\mathbf{2}\boldsymbol{\pi t\omega})\boldsymbol{e}^{-\boldsymbol{t}^2/\boldsymbol{\tau}^2} \qquad \textbf{(2)}$$

with $\omega = 4/\tau$. The two parameters σ and τ of the response function R correspond roughly to the spatial and temporal scale of the detector. The local maxima of the response function are the interest points. These interest points correspond to the local regions where the complex motion patterns are occurred. Examples of local space-time features detected by Cuboid detector for different human actions are shown in figure 1. In our experiments, the Cuboid detector parameters $1 \le \sigma \le 4$ and $1 \le \tau \le 4$ were tested to select the best for each dataset according to Dollár et al.[4].

Fig. 1. Sample sequences with detected interest points by Cuboid detector for the KTH dataset. (a) boxing action. (b) running action.

2.2 SPACE-TIME DESCRIPTORS

Describing the patches around the detected interest points is an important step in the human action recognition framework in which the appearance and/or motion information are captured. The spatial and temporal sizes of a patch are functions of σ and τ respectively. Examples of extracted cuboids at interest points are shown in figure 2. In our experiments, the number of cuboids is fixed to 100 for all video sequences based on Shao et al.[10] and the patch size provided by the authors is used.

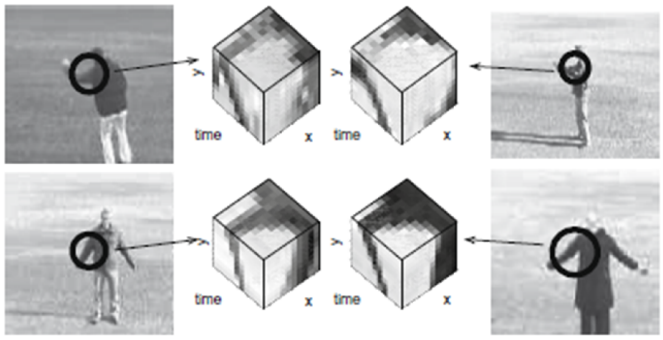

Fig. 2. Extraction of space–time cuboids at interest points from similar actions performed by different persons [12].

Cuboid descriptor proposed by Dollár et al. [4]. The cuboid is firstly smoothed before computing the image gradients. Then, the gradients are computed for each pixel in the patch along x, y and t directions. The computed gradients are concatenated into a single vector. A principal component analysis (PCA) is then used to project the feature vector into a lower dimensional space. The spatial size for the descriptor is given by 2 * ceil (3σ) + 1 and the temporal size by 2* ceil (3τ) + 1. After PCA projection, the Gradient feature vector length is reduced to 100.

Histogram of Oriented Gradients (HOG) descriptor was introduced by Laptev et al. [6] which describes the local appearance in each cuboid by computing the histograms of spatial gradient. The authors subdivide each cuboid into a grid $n_x \times n_y \times n_t$. For each cell, 4-bins histogram of gradient orientations (HOG) is computed. The default grid parameters (3 × 3 × 2) which yield a feature vector of length 72 is used. The spatial size for the descriptor is given by (18σ) and the temporal size by (8τ). In the implementation, we follow the author's descriptor scales [9] for the Cuboid detector which set $\sigma^2=4$ and $\tau^2=2$.

Histogram of Optical Flow (HOF) descriptor is another descriptor proposed by Laptev et al. [6] with the same idea of HOG descriptor. It describes the local motion in each cuboid by computing 5-bins histogram of optical flow (HOF) for its cells. The HOF descriptor vector length is 90.

HOG-HOF descriptor is a combination of HOF and HOG descriptors provided by Laptev et al. [6] that captures both local appearance and motion information to enhance human action recognition performance. After combining both descriptors, a feature vector of length 162 is obtained.

3 EVALUATION FRAMEWORK

Final representation of videos is done by applying a common bag-of-words approach on the described space-time features [13]. The visual words (or codebook) are built from training data by using k-means clustering algorithm. Then, each video is represented as the frequency histogram of the codebook elements. The dimension of each video descriptor is equal to the size of the codebook. In our experiments, six codebook sizes (250, 500, 750, 1000, 1250 and 1500) are used. Due to the random initialization of k-means clustering algorithm, we report the best result over 10 runs.

For classification, we use a non-linear Support Vector Machine (SVM) with radial basis function (RBF) kernel;

$$\boldsymbol{K(H_i, H_j) = exp(-\gamma \sum_{n=1}^{V} \left\| h_{in} - h_{jn} \right\|^2). \quad \gamma > 0} \qquad \textbf{(2)}$$

Where, $H_i = h_{in}$, $H_j = h_{jn}$ are the histograms of feature, V is the codebook size and γ parameter defines how far the distance of a single training example reaches, with low values meaning 'far' and high values meaning 'close'. The library libSVM [14] was used for multi-class classification and the performance is reported as the average accuracy over all classes. The best classification parameters *C* and γ are selected by a grid search on all values of 2^x where *x* is in the range −5 to 16 and 2^y where *y* is in the range 3 to −15, respectively as provided by P. Bilinski et al. [11]. The C parameter trades off misclassification of training examples against simplicity of the decision surface. A low C makes the decision surface smooth, while a high C aims to classify all training examples correctly.

4 EXPERIMENTS AND RESULTS

In this section, the datasets and scenarios in which experiments were performed are described, and the results are explored.

4.1 DATASETS

Our experiments are carried out on two commonly used datasets for human action recognition: the KTH and Weizmann datasets. **KTH actions dataset** contains six actions: walking, jogging, running, boxing, hand waving and hand clapping (Figure 3) [13][1]. Each action is performed by 25 different actors in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. In total, the dataset consists of 600 video sequences. The video resolution is 160×120 pixel. All videos were taken over homogeneous backgrounds with a static camera with 25 fps. We follow the original authors setup dividing the dataset into testing set (9 subjects: 2, 3, 5, 6, 7, 8, 9, 10, and 22) and training set (the remaining 16 subjects). The length of every video sequence is limited to the first 300 frames, as it has been done in [10]. The Cuboid detector is run on this dataset with parameters σ=3 and τ=1.5, which gave better results in our evaluations. To limit the complexity, 25% of training data is randomly selected to construct the visual words. The 5-fold cross-validation technique is used on the training set to choose the best classification parameters.

Fig. 3. A few sample frames from video sequences of KTH dataset. From (a) to (e), the actions are walking, boxing, waving, clapping and running.

[1] *http://www.nada.kth.se/cvap/actions/*

Weizmann actions dataset contains 10 actions: walk, run, jump, gallop sideways, bend, one-hand wave, two-hands wave, jump in place, jumping jack and skip (Figure 4) [15][2]. Each action is performed by 9 persons. The Weizmann dataset contains 93 video sequences with spatial resolution 180 × 144 pixel. All videos were taken over homogeneous backgrounds with a static camera with 50 fps. In our experiments, the best Cuboid detector parameters for Weizmann dataset are σ=2, τ=1.5. The Evaluation on this dataset is done using leave-one-person-out cross-validation technique.

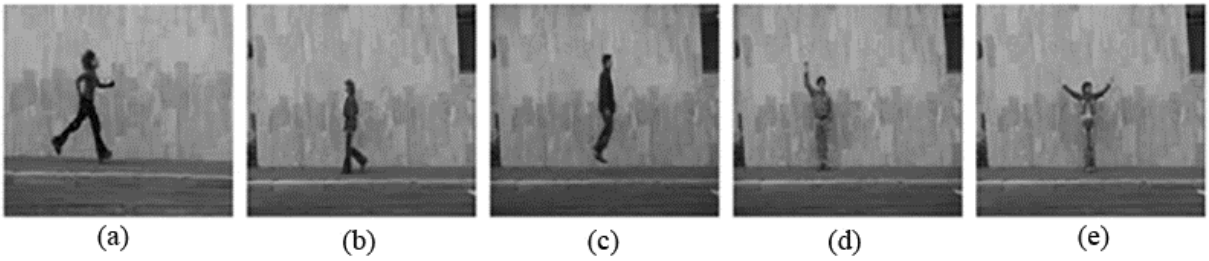

Fig. 4. A few sample frames from video sequences of Weizmann dataset. From (a) to (e), the actions are run, walk, jump, wave1 and wave2.

4.2 RESULTS

For KTH dataset, the results are presented in table 1. The best result of 92.12% is obtained with the HOG-HOF descriptor. Both Gradient and HOG-HOF descriptors preform best for codebook 750 and HOG for codebook 1250 and HOF for codebook 1500. According to the results, we can rank the descriptors as: HOG-HOF > HOF > Gradient > HOG. Wang et al. [9] obtained 89.1% for the Gradient descriptor, 82.3% accuracy for the HOG descriptor, 88.2% for the HOF descriptor and 88.7% for the HOG-HOF. They tested the Cuboid detector and the descriptors by choosing a subset of 100, 000 randomly selected training features and using codebook size 4000. Shao et al. [10], report around 91% for Gradient descriptor , however, we reached only 89.81%. This could be due to the subset of the training data used in our evaluation for codebook construction. It should also be noted that many evaluations of HOG, HOF, HOG-HOF descriptors [8]–[11] have been proposed using Harris3D detector. For example, P. Bilinski et al. [11] obtained 83.33% accuracy for the HOG descriptor, 95.37% for the HOF descriptor and 94.44% for the HOG-HOF. The use of Cuboid detector achieved up to 2.78% better results for HOG descriptor.

Table 1. Action recognition accuracy for various Codebook's size /descriptor on the KTH dataset

	GRADIENT	HOG	HOF	HOG-HOF
250	85.18%	79.16%	86.57%	89.35%
500	87.5%	81.48%	87.5%	90.27%
750	**89.81 %**	82.40%	89.35%	**92.12%**
1000	88.88%	85.18%	88.42%	91.20%
1250	89.35%	**86.11%**	89.81%	90.74%
1500	88.42%	84.25%	**91.66%**	90.74%

In tables 2, 3, 4, 5, we show the confusion matrices of the best classification result for each descriptor on the KTH dataset. As the tables show, there is some confusion among the actions “walk”, “run” and “jog”. Similarly, the actions “box”, “wave” and “clap” always confuse each other. This is expected because these actions include some similar features in their representation. Among all descriptors, the worst error ratio of “run” which is recognized as “jog” 25% is obtained by using the HOF descriptor. This ratio is decreased to 22.22% for HOG-HOF descriptor, 19.44% for HOG descriptor and reached its

[2] *http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html*

lowest value 13.89% when the Gradient descriptor is used. In contrast to all descriptors, HOG descriptor reports a high error ratio of "clap" which is recognized as "box" 22.22%. This confusion may explain the low performance of HOG descriptor on KTH dataset. The performance of all tested descriptors for each action is shown in table 6.

Table 2. Confusion matrix for the best classification result of the Gradient descriptor on KTH dataset.

(%)	BOX	CLAP	WAVE	JOG	RUN	WALK
BOX	100	0	0	0	0	0
CLAP	0	91.67	0	0	0	8.33
WAVE	11.11	0	83.33	2.78	0	2.78
JOG	0	0	0	83.33	11.11	5.56
RUN	0	0	0	13.89	83.33	2.78
WALK	0	0	0	2.78	0	97.22

Table 3. Confusion matrix for the best classification result of the HOG descriptor on KTH dataset.

(%)	BOX	CLAP	WAVE	JOG	RUN	WALK
BOX	88.89	2.78	0	5.56	2.78	0
CLAP	22.22	72.22	5.56	0	0	0
WAVE	2.78	2.78	91.67	0	2.78	0
JOG	0	0	0	86.11	8.33	5.55
RUN	0	0	0	19.44	80.65	0
WALK	0	0	0	2.78	0	97.22

Table 4. Confusion matrix for the best classification result of the HOF descriptor on KTH dataset.

(%)	BOX	CLAPP	WAVE	JOG	RUN	WALK
BOX	100	0	0	0	0	0
CLAP	0	100	0	0	0	0
WAVE	0	5.56	91.67	0	0	2.78
JOG	0	0	0	86.11	11.11	2.78
RUN	0	0	0	25	75	0
WALK	0	0	0	2.78	0	97.22

Table 5. Confusion matrix for the best classification result of the HOG-HOF descriptor on KTH dataset.

(%)	BOX	CLAP	WAVE	JOG	RUN	WALK
BOX	100	0	0	0	0	0
CLAP	0	97.22	2.78	0	0	0
WAVE	0	8.33	88.89	2.78	0	0
JOG	0	0	0	91.67	5.56	2.78
RUN	0	0	0	22.22	77.77	0
WALK	0	0	0	2.78	0	97.22

Table 6. Comparison of recognition accuracy for each action.

DESCRIPTION METHOD	BOX	CLAP	WAVE	JOG	RUN	WALK
GRADIENT	100	91.67	83.33	83.33	83.33	97.22
HOG	88.89	72.22	91.67	86.11	80.65	97.22
HOF	100	100	91.67	86.11	75	97.22
HOG-HOF	100	97.22	88.89	91.67	77.78	97.22

Weizmann dataset evaluation results are presented in table 7. As shown, the HOF is the best descriptor for the Weizmann dataset as it obtained 92.22% accuracy. Furthermore, all descriptors (except the HOG descriptor) perform the best for codebook size 750. This may be explained by the little variations in this dataset that the videos were captured in one scenario unlike KTH dataset (which has four scenarios). The HOG descriptor performs best for codebook size 1000. P. Bilinski et al. [11] used the Harris3D detector in their evaluation and obtained 86.02% accuracy for the HOG descriptor, 91.40% for the HOF descriptor and 92.74% for the HOG-HOF. The Cuboid detector obtained up to 1.75% better results for HOG descriptor and 0.82% better results for HOF descriptor.

Table 7. Action recognition accuracy for various Codebook's size /descriptor on the Weizmann dataset.

	GRADIENT	HOG	HOF	HOG-HOF
250	87.77%	84.44%	88.88%	88.88%
500	87.77%	84.44%	91.11%	90%
750	**88.88%**	85.55%	**92.22%**	**91.11%**
1000	87.77%	**87.77%**	91.11%	91.11%
1250	86.66%	85.55%	90%	91.11%
1500	84.44%	86.66%	90%	90%

Table 8. Confusion matrix for the best classification result of the Gradient descriptor on Weizmann dataset.

(%)	BEND	JACK	JUMP	P-JUMP	RUN	SIDE	SKIP	WALK	WAVE1	WAVE2
BEND	100	0	0	0	0	0	0	0	0	0
JACK	0	100	0	0	0	0	0	0	0	0
JUMP	0	0	77.77	0	11.11	0	11.11	0	0	0
P-JUMP	0	0	0	100	0	0	0	0	0	0
RUN	0	0	22.22	0	66.66	0	11.11	0	0	0
SIDE	0	0	0	0	0	88.88	11.11	0	0	0
SKIP	0	0	33.33	0	11.11	0	55.55	0	0	0
WALK	0	0	0	0	0	0	0	100	0	0
WAVE1	0	0	0	0	0	0	0	0	100	0
WAVE2	0	0	0	0	0	0	0	0	0	100

Table 9. Confusion matrix for the best classification result of the HOG descriptor on Weizmann dataset.

(%)	BEND	JACK	JUMP	P-JUMP	RUN	SIDE	SKIP	WALK	WAVE1	WAVE2
BEND	100	0	0	0	0	0	0	0	0	0
JACK	0	100	0	0	0	0	0	0	0	0
JUMP	0	0	77.77	0	11.11	0	11.11	0	0	0
P-JUMP	0	0	0	100	0	0	0	0	0	0
RUN	0	0	11.11	0	77.77	0	11.11	0	0	0
SIDE	0	0	0	0	11.11	88.88	0	0	0	0
SKIP	0	0	33.33	0	11.11	0	44.44	11.11	0	0
WALK	0	0	0	0	0	0	0	100	0	0
WAVE1	0	0	0	0	0	0	0	0	100	0
WAVE2	0	0	0	0	0	0	0	0	11.11	88.88

The confusion matrices of the best classification result for each descriptor on the Weizmann dataset are presented in tables 8, 9, 10, 11. For all descriptors, the recognition accuracies of actions “bend,” “jack,” “p-jump,” “walk” and “wave1” reach 100%. The confusion also occurs between the actions “jump,” “run” and “skip”. The 33.33% of “skip” are recognized as “jump” when we use Gradient, HOG and HOG-HOF descriptors. Only the HOF descriptor can decrease this confusion to 22.22%. Another high confusion is obtained by using Gradient descriptor where 22.22% of “run” samples are recognized as “jump”. This confusion is enhanced by using HOG descriptor and it disappeared with HOF and HOG-HOF descriptors. Table 12 summarizes the performance of all tested descriptors for each action.

Table 10. Confusion matrix for the best classification result of the HOF descriptor on Weizmann dataset.

(%)	BEND	JACK	JUMP	P-JUMP	RUN	SIDE	SKIP	WALK	WAVE1	WAVE2
BEND	100	0	0	0	0	0	0	0	0	0
JACK	0	100	0	0	0	0	0	0	0	0
JUMP	0	0	77.77	0	11.11	0	11.11	0	0	0
P-JUMP	0	0	0	100	0	0	0	0	0	0
RUN	0	0	0	0	100	0	0	0	0	0
SIDE	0	0	0	0	11.11	88.88	0	0	0	0
SKIP	0	0	22.22	0	11.11	0	66.66	0	0	0
WALK	0	0	0	0	0	0	0	100	0	0
WAVE1	0	0	0	0	0	0	0	0	100	0
WAVE2	0	0	0	0	0	0	0	0	11.11	88.88

Table 11. Confusion matrix for the best classification result of the HOG-HOF descriptor on Weizmann dataset.

(%)	BEND	JACK	JUMP	P-JUMP	RUN	SIDE	SKIP	WALK	WAVE1	WAVE2
BEND	100	0	0	0	0	0	0	0	0	0
JACK	0	100	0	0	0	0	0	0	0	0
JUMP	0	0	88.88	0	11.11	0	0	0	0	0
P-JUMP	0	0	0	100	0	0	0	0	0	0
RUN	0	0	0	0	100	0	0	0	0	0
SIDE	0	0	0	0	11.11	77.77	11.11	0	0	0
SKIP	0	0	33.33	0	22.22	0	44.44	0	0	0
WALK	0	0	0	0	0	0	0	100	0	0
WAVE1	0	0	0	0	0	0	0	0	100	0
WAVE2	0	0	0	0	0	0	0	0	0	100

Finally, the computational time of the tested descriptors are compared as shown in table 13. The time is calculated in seconds for one cuboid on a PC with Core 2 Duo 2.13GHz processor and 4GB RAM. The comparison is performed on a video sequence from KTH dataset and the average over 10 runs is reported. As the results show, the Gradient descriptor is the fastest one. HOG descriptor is also quite fast. They are based on gradient computations but slightly differ in the following description step. HOF and HOG-HOF descriptors are quite similar in time and they are much slower than other descriptors. This can be explained by the optical flow computations which take more time than the gradient computations.

Table 12. Computational Time in seconds for one cuboid's descriptors.

DESCRIPTION METHOD	COMP TIME
GRADIENT	0.0122
HOG	0.0189
HOF	0.5645
HOG-HOF	0.5995

5 CONCLUSIONS

In this paper, an evaluation of four widely used local space-time descriptors for human action recognition is presented. In contrast to other existing evaluations, a Cuboid detector is chosen to extract the space-time interest points in videos and the evaluation is performed on several codebook sizes. The experiments are carried out on two popular datasets (KTH Action Dataset, Weizmann Action Dataset) under the framework based on the bag-of-words approach and non-linear Support Vector Machine. The objective is to find the best descriptor with Cuboid detector for future use in more realistic and challenging scenarios.

In the tests, the optical flow based descriptors (HOF, HOF-HOG) seem to be good descriptors for the space-time features that were detected using Cuboid detector. For the KTH dataset, the HOF-HOG descriptor achieved the best performance (92.12%) and the HOF descriptor took the second place (91.66%). The best performance on the Weizmann database (92.22%) has been achieved using the HOF descriptor and the HOG-HOF descriptor ranked the second (91.11%). The experiments also showed that the HOG descriptor reports the lowest accuracy for all tested datasets; however, the Gradient based descriptors (Gradient, HOG) are approximately 30 times faster than the optical flow based descriptors.

REFERENCES

[1] J. K. Aggarwal and M. S. Ryoo, "Human Activity Analysis: A Review," *Acm Comput. Surv.*, vol. 43, no. 3, pp. 1–43, 2011.

[2] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer Vision and Image Understanding*, vol. 115, no. 2. pp. 224–241, 2011.

[3] I. Laptev, "On Space-Time Interest Points," *International Journal of Computer Vision*, vol. 64, no. 2–3. pp. 107–123, 2005.

[4] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," *2005 IEEE Int. Work. Vis. Surveill. Perform. Eval. Track. Surveill.*, 2005.

[5] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," *Comput. Vision–ECCV 2008*, 2008.

[6] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," *2008 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008.

[7] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *15th international conference on Multimedia*, 2007, pp. 357–360.

[8] A. Kl, C. Schmid, and I. Grenoble, "A Spatio-Temporal Descriptor Based on 3D-Gradients," in *British Machine Vision Conference*, 2008.

[9] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," *Procedings Br. Mach. Vis. Conf. 2009*, p. 11, 2009.

[10] L. Shao and R. Mattivi, "Feature detector and descriptor evaluation in human action recognition," in *Proceedings of the ACM International Conference on Image and Video Retrieval - CIVR '10*, 2010, p. 477.

[11] P. Bilinski and F. Bremond, "Evaluation of local descriptors for action recognition in videos," *Comput. Vis. Syst.*, 2011.

[12] I. Laptev, B. Caputo, C. Schüldt, and T. Lindeberg, "Local velocity-adapted motion events for spatio-temporal recognition," *Comput. Vis. Image Underst.*, vol. 108, no. 3, pp. 207–229, 2007.

[13] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," *Proc. 17th Int. Conf. Pattern Recognition, 2004. ICPR 2004.*, vol. 3, 2004.

[14] C.-C. Chang and C.-J. Lin, "LIBSVM: a Library for Support Vector Machines (Version 2.31)," *Science*. pp. 1–22, 2001.

[15] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, 2007.