ADVANCES ON ACTION RECOGNITION IN VIDEOS USING AN INTEREST POINT DETECTOR BASED ON MULTIBAND SPATIO-TEMPORAL ENERGIES

Kevis Maninis, Petros Koutras and Petros Maragos

School of E.C.E., National Technical University of Athens, Greece

Email: {pkoutras, maragos}@cs.ntua.gr

ABSTRACT

This paper proposes a new visual framework for action recognition in videos, that consists of an energy detector coupled with a carefully designed multiband energy based filterbank. The tracking of video energy is performed using perceptually inspired 3D Gabor filters combined with ideas from Dominant Energy Analysis. Within this framework, we utilize different alternatives such as non-linear energy operators where actions are implicitly considered as manifestations of spatio-temporal oscillations in the dynamic visual stream. Texture and motion decomposition of actions through multiband filtering is the basis of our approach. This new energy-based saliency measure of action videos leads to the extraction of local spatio-temporal interest points that give promising results for the task of action recognition. Such interest points are processed further in order to formulate a robust representation of an action in a video. Theoretical formulation is supported by evaluation in two popular action databases, in which our method seems to outperform the state of the art.

Index Terms— Human action recognition, spatio-temporal interest point detectors, multiband Gabor filtering, dominant energy analysis, energy tracking in videos.

1. INTRODUCTION

The task of human action recognition through local space-time features has recently become very popular in computer vision. Local features give a more compact representation of the video, aiming at keeping only the useful information for action recognition. Video representations in terms of such features exhibit efficiency in distinguishing among action classes, while bypassing the need for precise background subtraction or tracking. Local image and video features have been successfully used for many tasks such as object and scene recognition [23] as well as human action recognition [9, 12, 24, 32]. Local spatio-temporal features are able to capture characteristic shape and motion in video. They can focus on specific events independently of their shifts and scales, as well as background clutter and multiple motions in the scene. The feature extraction process usually includes two discrete steps. The first deals with spatio-temporal interest point detection, that is usually performed directly on the videos, using a detector which maximizes a specific saliency function as in [9, 17, 37]. The second step includes the computation of local descriptors which employ video measurements such as gradients, optical flow and energy to encode local appearance and motion information in a neighbourhood of the detected points.

Laptev and Lindenberg in [22] introduce the Harris3D detector, an extension of the Harris edge detector [13] in 3 dimensions. Dollár et al. [9] argued against detection criteria such as spatiotemporal edges and suggested a detector relying on Gaussian smoothing in the spatial dimensions and a pair of Gabor filters in quadrature [1, 20] applied to the temporal dimension. Georgakis et al. [12] introduced the DCA3D detector, which is based on multichannel filtering via Gabor filters and Dominant Component Analysis. First, the filtering process takes place in the 2 spatial dimensions and then the dominant component volume is filtered by temporal filters to lead to the final energy volume and the selection of interest points.Wang et al. [35] introduced dense trajectories and motion boundary histograms to deal with the task of action recognition. They further extended their ideas by estimating and cancelling the camera motion from optical flow computations leading to higher performances.

Many different local descriptors have been proposed in the past few years [4,19,23,24,33,37]. Descriptors are used to represent each feature by measurements in their neighbourhood area. An overview of detectors and descriptors can be found in Wang et al. [36].

The above methods, followed by the *Bag-of-Features (BoF)* approach [24, 29, 32, 34] and classification with Support Vector Machines (SVMs), have lead to high accuracy results in many applications. Bettadapura et al. [2] proposed a variation of BoF that takes advantage of the temporal information given in an action video, named Augmented Bag-of-Words.

Several different methods for the task of action recognition have been proposed in [3, 30, 31, 38, 39]. Evaluations in several action datasets with different experimental setups have been performed using different approaches through the literature.

In this paper we propose a new visual framework for action recognition. We employ physiologically inspired 3D Gabor filters that cover both the spatial and the temporal frequency domain in order to detect spatio-temporal energies. In this approach we use the 3D extension of non-linear Teager-Kaiser Energy Operator (TKEO) [18,25] together with ideas from Dominant Energy Analysis in order to select the energy value for each voxel of the videos. This novel spatio-temporal interest point detector tracks energy of both texture and motion through biologically plausible Gabor filters, which are applied simultaneously in space and time. The resulting multiband energies are processed further to formulate dominant energy representations that lead to local features which successfully deal with the task of action recognition.

The remainder of this paper is organized as follows. In Section 2, we present the spatio-temporal energy detector. In Section 3 we describe the process of interest point extraction and classification of the videos. The experimental framework is provided in Section 4 and finally Section 5 concludes this paper.

2. SPATIO-TEMPORAL ENERGY DETECTOR: THE GABOR3D ALGORITHM

Our energy-based model for spatio-temporal interest point detection for action classification uses biologically plausible spatio-temporal

This research work was supported by the project "COGNIMUSE" which is implemented under the "ARISTEIA" Action of the Operational Program Education and Lifelong Learning and is co-funded by the European Social Fund and Greek National Resources.

filters, like oriented 3D Gabor filters, in order to extract visual features which are considered to describe actions efficiently. The overall processing for interest point detection is shown in Fig. 1. In a first phase the initial RGB video volume is transformed into grayscale. Then follows the main process step, called *Spatio-Temporal Energy Analysis (STEA)*, which is applied directly to the grayscale video volume. The last stage includes the interest point extraction process, in which the 3D local maxima of the resulting energy volume are considered as the interest points of the input video.

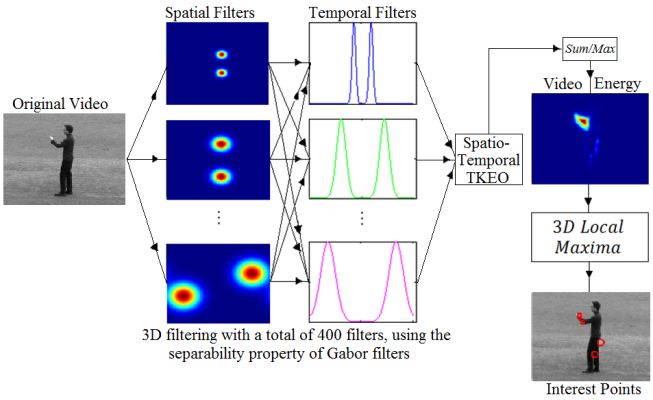


Fig. 1: Spatio-temporal interest point detection with the Gabor3D algorithm of a sample video of *Boxing* action of KTH Action Dataset. The input video is processed by the Gabor filterbank and the video energy is computed. The interest points are selected as the thresholded local maxima of the 3D energy volume.

Spatio-Temporal 3D Gabor Filtering

The first step of STEA is the filtering process of the video volume. Among the filtering approaches that have been proposed based on psychophysical experiments, the two with the widest acceptance are the Gabor filters and the Gaussian Derivatives (GD). We choose to use oriented Gabor filters in a spatio-temporal version, due to their biological plausibility and their uncertainty-based optimality [6, 11]. In addition, for high order derivatives the GD filters are approximations of the Gabor filters [21].

GD filters combined with their Hilbert transform (quadrature pair) are widely used in many spatial and spatio-temporal tasks [7, 8, 28], mainly because they can be implemented in an efficient way since they are steerable [10]. Gabor filters, on the other hand, are not strictly steerable mathematically, but as Heeger [15, 16] showed they can become separable, which means that a high dimensional Gabor filter can be built from 1D Gabor impulses responses.

So, we apply quadrature pairs of 3D (spatio-temporal) Gabor filters with identical central frequencies and bandwidth. These filters can arise from 1D Gabor filters [11] in a similar way as Daugman proposed 2D Oriented Gabor Filters [5]. An 1D complex Gabor filter consists of a complex sine wave modulated by a Gaussian window. Its impulse response with unity norm has the form:

$$g(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{t^2}{2\sigma^2}\right) \exp(j\omega_{t_0} t) = g_c(t) + j g_s(t) \qquad (1)$$

The above complex filter can be split into one odd(sin)-phase ($g_s(t)$) and one even(cos)-phase ($g_c(t)$) filters, which form a quadrature pair filter. Almost all Gabor filters are bandpass filters whose center frequency coincides with their modulating frequency ω_{t_0}; the only exception where they become lowpass filters is when $\omega_{t_0} = 0$, which makes them Gaussians. Thus, we can cover the whole spatio-temporal 3D spectral domain with Gabor filters whose frequency responses are centered around specific frequencies.

The 3D Gabor extension (as for example used for optical flow in [16]) yields an even (cos) 3D Gabor filter whose impulse response is:

$$g_c(x,y,t) = \frac{1}{(2\pi)^{3/2}\sigma_x\sigma_y\sigma_t} \exp\left[-\left(\frac{x^2}{2\sigma_x^2} + \frac{y^2}{2\sigma_y^2} + \frac{t^2}{2\sigma_t^2}\right)\right] \cdot \cos(\omega_{x_0} x + \omega_{y_0} y + \omega_{t_0} t) \qquad (2)$$

where $\omega_{x_0}, \omega_{y_0}, \omega_{t_0}$ are the spatial and temporal angular center frequencies and $\sigma_x, \sigma_y, \sigma_t$ are the standard deviations of the 3D Gaussian envelope. Similarly for the impulse response of odd (sin) filter which we denote by $g_s(x,y,t)$.

The frequency response of an even (cos) Gabor filter consists of two Gaussian ellipsoids symmetrically placed at frequencies $(\omega_{x_0}, \omega_{y_0}, \omega_{t_0})$ and $(-\omega_{x_0}, -\omega_{y_0}, -\omega_{t_0})$. Figure 2 shows isosurfaces of the 3D spatio-temporal filterbank as well as a top view of a filterbank slice designed at some temporal frequency ω_{t_0}. Note that the symmetric lobes of each filter appear at the plane defined by the temporal frequency $-\omega_{t_0}$ in contrast with the 2D case. So, if we want to cover the spatial frequency plane at each temporal frequency we must include in our filterbank both positive and negative temporal frequencies. Further, the bandwidth of each filter varies with the spatial scale and temporal frequency.

For the spatio-temporal filterbank we used $N = 400$ Gabor filters (isotropic in the spatial components) which are arranged in five spatial scales, eight spatial orientations and ten temporal frequencies. The spatial scales and orientations are selected to cover a squared 2D frequency plane in a similar way to the design by Havlicek et al. [14]. Then both center frequencies and Gaussian bandwidths are divided by the spatial sampling frequencies in order to get discrete filters with normalized frequency parameters that can be directly applied at every image size. We note that this process can lead to anisotropic spatial Gabor filters for non-square images, although the original design includes isotropic filters.

We use ten temporal Gabor filters, five at positive and five at negative center frequencies due to the 3D spectrum symmetries. These are linearly spaced to span the normalized frequency axis and each filter's half-peak octave bandwidth is 0.75 octaves. Figure 2 shows spatio-temporal views of our design. Note that including both positive and negative frequencies does not increase the filtering complexity because, due to Gabor filters' separability, no additional convolutions are needed but only changing the signs at (3)-(4).

Reducing the complexity of the 3D filtering process

The 3D filtering is a time consuming process due to the complexity of all required 3D convolutions. However, Gabor filters are separable [15, 16], which means that we can filter each dimension separately using an impulse response having the form (1). In this way, we apply only 1D convolutions instead of 3D, which increases the efficiency of the computations. Then the 3D output can be easily composed from 1D filtering outputs by using simple trigonometric properties in two steps (first 2D and then 3D). First, we compose the 2D spatial output from the impulse responses $g_c(x), g_s(x), g_c(y), g_s(y)$ for both the even- and odd-phase filter:

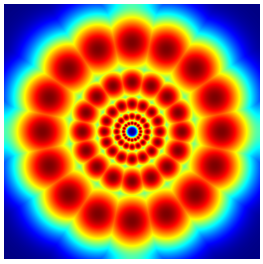
$$y_c^{2D}(x,y,t) = (V(x,y,t) * g_c(x)) * g_c(y) - (V(x,y,t) * g_s(x)) * g_s(y)$$
$$y_s^{2D}(x,y,t) = (V(x,y,t) * g_s(x)) * g_c(y) + (V(x,y,t) * g_c(x)) * g_s(y)$$

where V is the grayscale video volume. Then the 3D output corresponding to spatio-temporal filtering can be obtained by convolving the above 2D output with the 1D temporal impulse responses:

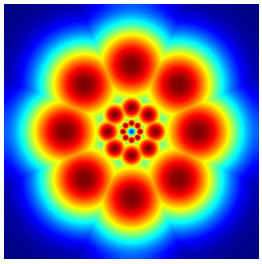
$$y_c^{3D}(x,y,t) = y_c^{2D}(x,y,t) * g_c(t) - y_s^{2D}(x,y,t) * g_s(t) \qquad (3)$$
$$y_s^{3D}(x,y,t) = y_c^{2D}(x,y,t) * g_s(t) + y_s^{2D}(x,y,t) * g_c(t) \qquad (4)$$

For an image of size $n \times n \times n$ and a convolution kernel of $m \times m \times m$ the complexity is reduced from $\mathcal{O}(n^3 \cdot m^3)$ that is required for 3D convolutions to $\mathcal{O}(3n^3 \cdot m)$ that is required for three separable 1D convolutions. Color is also a factor that can be ignored without serious loss of information for the action recognition task. The conversion of a colored video to grayscale is a process that reduces its size to $1/3$ of its original size.

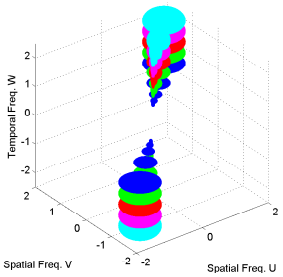
The number of Gabor filters used for extracting the energy maps can be adjusted to reduce further the time consumed by the filtering process. Experimental results presented in Section 4 show that using a reduced version of filterbank in space (Fig. 2b) instead of the filterbank proposed by Havlicek et al. [14] (Fig. 2a) results in approximately the same recognition accuracy.



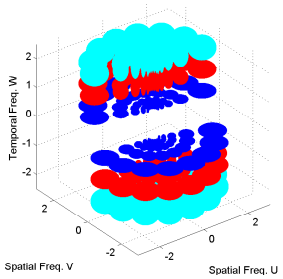
(a) Full filterbank at ω_{t_0} and $-\omega_{t_0}$ (top view).



(b) Reduced filterbank at ω_{t_0} and $-\omega_{t_0}$ (top view).



(c) Spatio-Temporal Filterbank at 5 different spatial scales, 1 of 8 orientation and 5 temporal frequencies of the full filterbank.



(d) Spatio-Temporal Filterbank at 5 different spatial scales, 8 spatial orientations and 3 of 5 temporal frequencies of the full filterbank.

Fig. 2: Isosurfaces of the 3D Spatio-Temporal Filterbank and a top view of a filterbank slice designed at temporal frequency ω_{t_0}, for the full and the reduced version. Isosurfaces correspond at 70%-peak bandwidth magnitude while different colors are used for different temporal frequencies. We can see that the symmetric lobe of each filter appeared at the plane defined by the temporal frequency $-\omega_{t_0}$ in contrast with the 2D case.

Parameters of the Gabor3D detector for Action Classification

While designing the Gabor3D detector, we had to select among different alternatives of the following parameters: the type of energy to be used, the type of Gabor filters and the way of handling the output energy volumes of all filters.

First of all, for the type of energy we had to choose between the simple square energy and the Teager-Kaiser energy. Square energy $E_S(\cdot)$ is defined as: $E_S(f(x,y,t)) = f^2(x,y,t)$ where $f(x,y,t)$ is the output of the filtering process at a pixel (x,y,t). The Teager-Kaiser Energy Operator (TKEO) [18] $\Psi[s(t)] \equiv [s'(t)]^2 - s(t)s''(t)$ has facilitated the energy representation of signals modeled by non-stationary sinusoids, with amplitude and frequency modulation (AM-FM) of the form $s(t) = a(t)cos(\phi(t))$. The estimation of the Teager-Kaiser energy of a signal presupposes the fact that it is narrowband [26], which in our case is implemented through band-pass Gabor filtering. Maragos and Bovik [25] extended TKEO to signals of higher dimensions. In our approach we make use of the 3D TKEO $\Phi(\cdot)$ which is defined as:

$$\Phi(f) \equiv \|\nabla f\|^2 - f \cdot \nabla^2 f = f_x^2 + f_y^2 + f_t^2 - f \cdot (f_{xx} + f_{yy} + f_{tt})$$

where $f = f(x,y,t)$. The discrete Teager-Kaiser energy used in our experiments is an outcome of the discretization of the spatial and temporal derivatives and is defined as:

$$\begin{aligned}\Phi_d[f[x,y,t]] = 3f^2[x,y,t] - f[x-1,y,t] \cdot f[x+1,y,t] \\ - f[x,y-1,t] \cdot f[x,y+1,t] - f[x,y,t-1] \cdot f[x,y,t+1]\end{aligned}$$

The motivation for using the TKEO is its ability to track spatio-temporal energy oscillations and separate them into their amplitude and frequency components with excellent spatio-temporal resolution and very small complexity.

The second choice that has to be made concerns the type of Gabor filters. We have to choose between simple cosine filters and quadrature filters, supported by [20] and [1]. If $f_c(x,y,t)$ and $f_s(x,y,t)$ are the outputs of the video filtered by the cosine and the sine Gabor filters, the energies resulting from a simple cosine Gabor filter and a pair of Gabor filters in quadrature respectively are defined as:

$$E_{cos}(x,y,t) = E(f_c(x,y,t)) \quad (5)$$

$$E_{quad}(x,y,t) = E(f_c(x,y,t)) + E(f_s(x,y,t)) \quad (6)$$

where $E(\cdot)$ is defined as the general energy operator. In our case, (5) and (6) can be defined for both Square and Teager-Kaiser energy by substituting $E(\cdot)$ with $E_S(\cdot)$ and $\Phi(\cdot)$, respectively. Filters in quadrature (or filters in 90° phace difference) are used to detect different type of edges. In the case of Gabor functions, this is easy to implement by simply using the sine and cosine versions of the same filter and squaring the outputs of those two as described in (6).

As presented in Fig. 1, the energy outputs of all 400 filters are handled by some operator in order to obtain the final energy map of each video. We used some ideas from Dominant Energy Analysis (DEA), as in [12], where the energy of the most dominant channel is considered as the energy value in each voxel:

$$E_{max}(x,y,t) = \max_{1 \le k \le N} E_k(x,y,t) \quad (7)$$

where $E_k(x,y,t)$ is the energy output of the k-th filter (or filter pair in case of quadrature filters) and N is the total number of the filters. We compared the performance of DEA with an other approach where we compute the average value of all filters as the final energy representation, which is equivalent to the superposition of all outputs divided by the number of filters:

$$E_{ave}(x,y,t) = \sum_{k=1}^{N} E_k(x,y,t)/N \quad (8)$$

We evaluated those parameters to select the combination that leads to the most suitable interest points for our approach regarding the task of action classification. Specifically, the evaluation was performed on the KTH Action Dataset [32] due to its simplicity, and the combination that lead to the highest classification accuracy was selected for the Gabor3D interest point detector. Finally, we used Teager-Kaiser Energy, Gabor filters in quadrature and Dominant Energy Analysis since they achieved the highest recognition accuracy.

3. ACTION CLASSIFICATION WITH INTEREST POINTS

So far we have described the Gabor3D detector until the point where the energy map is computed. In our approach, the next step is to extract spatio-temporal interest points. Interest points are a set of voxels, each defined by (x,y,t) in a neighbourhood of which we make some measurements. Those interest points are extracted from the energy map, as the local 3D maxima of the final energy volume, as shown in Fig. 3. To reduce the false alarms we use non-maxima suppression. Specifically, a threshold is put to the values of detected maxima, and values lower than the threshold are ignored.

The interest points resulting from Gabor3D algorithm are used for the classification of the actions. We have extracted a number of voxels, that constitute the local regions of interest in a video. To complete the recognition process we cascade the detected points with post processing that has three parts: computation of a local descriptor in a neighbourhood around each interest point, construction of the Bag-of-Features histograms and classification with SVMs. The block diagram of the entire process can be found in Fig. 4.

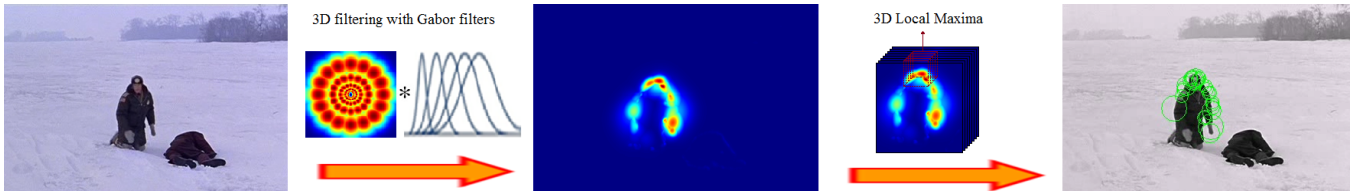


Fig. 3: Spatio-temporal interest point detection in a Hollywood2 video. Interest points are defined as the thresholded 3D local maxima of the energy volume.

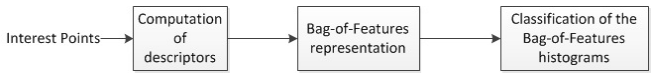


Fig. 4: For each interest point a descriptor is computed. Each descriptor is considered as a visual feature so descriptors resulting from all the interest points represent a video as the bag of visual features histogram (BoF). At last, the BoF histogram is classified by an SVM classifier.

4. EXPERIMENTAL RESULTS

For the experimental evaluation of our methods, we conducted action classification experiments in two databases, the *KTH* and the *Hollywood2* datasets. We compare our work to popular methods based on interest points such as Harris3D [24] and Cuboids [9].

The *KTH Action Dataset* [32] contains six types of different human action classes: walking, jogging, running, boxing, waving and clapping. Each action is performed several times by 25 annotators. The sequences were recorded in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. The actions are simple and performed without background noise. In total, the dataset consists of 2391 video sequences. We follow the experimental setup of Schüldt et al. [32] to be able to compare the results to the results reported in [36].

Interest points are detected with Gabor3D algorithm while we experiment with HOG/HOF [24] and HOG3D [19] descriptors and classification with χ^2-kernel SVMs is performed to the BoF histograms. Our accuracy of 93.5% outperforms the 91.8 % published in [24] (90.38 % in our experiments) as shown in Table 1. Table 2 shows the confusion matrix of our best result. As expected, the biggest confusion occurs between Jogging and Running due to the similarity of the actions. No confusion occurs between hand based and foot based actions. We should mention that the accuracy obtained by using the reduced version of the filterbank shown in Fig. 2b is almost the same as the one with the original filterbank of 400 filters. The reduced version employs a smaller number of 120 Gabor filters and yields in about 4 times faster execution of the algorithm.

	Method Accuracy				
Descriptor	DCA3D	Cuboids	Harris3D	Gabor3D (Full)	Gabor3D (Reduced)
HOG/HOF	78.8%	88.7 %	91.8%	91.2%	-
HOG3D	-	90.0 %	89.0%	**93.5%**	**93.4%**

Table 1: Mean accuracy of various methods for the KTH Action Dataset. The reduced Gabor3D algorithm seems to perform as well as the full version with 400 filters.

Action	Walking	Jogging	Running	Boxing	Waving	Clapping
Walking	1.00	0	0	0	0	0
Jogging	0.03	0.88	0.09	0	0	0
Running	0	0.20	0.80	0	0	0
Boxing	0	0	0	1.00	0	0
Waving	0	0	0	0	0.95	0.05
Clapping	0	0	0	0.02	0	0.98

Table 2: Confusion Matrix of classification experiments on KTH Action database with Gabor3D algorithm. Mean accuracy is 93.5 %.

In Table 3 we show the average accuracy of our method while changing the kernel of the SVMs, the type of energy and the way of handling the output of the filters. The combination of χ^2-SVMs, Teager-Kaiser Energy and Dominant Energy Analysis leads to the highest accuracy.

	Linear SVMs		χ^2 SVMs	
	Max	Sum	Max	Sum
Squared Energy	90.85%	89.34%	92.35%	90.61%
Teager-Kaiser Energy	91.31%	90.73%	**93.50%**	92.47%

Table 3: Mean Accuracy of the KTH Action Dataset Classification task while changing the parameters of the detection. χ^2 kernels perform better than Linear kernels, while Dominant Energy Analysis (Max) outperforms Energy Superposition (Sum) and Teager-Kaiser energy gives better results than Squared energy.

The *Hollywood2 Action Database* [27] contains 12 action classes collected from 69 different Hollywood movies: answering the phone, driving car, eating, fighting, getting out of the car, hand shaking, hugging, kissing, running, sitting down, sitting up, and standing up. In our experiments, we used the clean training dataset for the training process and we evaluated our method in the test set, as in [36]. In total, there are 1707 action samples divided into a training set (823 sequences) and a test set (884 sequences). Train and test sequences are obtained from different movies. The performance is evaluated as suggested in [27] by computing the average precision (AP) for each of the action classes and reporting the mean AP over all classes (mAP).

In Table 4 we compare our method to the results reported in [36]. Our method seems to give slightly better results than Cuboids [9] and Harris3D [23] detectors when using the same experimental setup. As local descriptor we chose HOG3D which showed the best performance for the KTH Action Dataset. It is necessary to say that the Hollywood2 database contains actions with noise, background motion and great variance in scale and the performances of each subject. So, the precision scores in every method is significantly lower than the KTH Action Dataset.

Method	Cuboids	Harris3D	Gabor3D
mAP	46.2%	45.2%	**47.7%**

Table 4: Mean average precision (mAP) of various methods for the Hollywood2 database. Comparison to the results reported in [36] with the same experimental setup. HOG3D is used as local descriptor in our experiments.

5. CONCLUSIONS

We proposed a new video energy tracking method that relies on detection of multiband spatiotemporal modulation components. The resulting energy volume is used as a basis for sparse space-time feature extraction in order to classify action videos. Experimental results show relatively higher results compared to other popular detectors in both KTH and Hollywood2 action databases. As future work, we would like to focus on improvement of action localization and extention of our experimental comparisons to more databases.

6. REFERENCES

[1] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Amer.*, 2(2):284–299, 1985.

[2] V. Bettadapura, G. Schindler, T. Plötz, and I. Essa. Augmenting bag-of-words: Data-driven discovery of temporal and structural information for activity recognition. In *Proc. IEEE Conf. CVPR*, 2013.

[3] M. Bregonzio, S. Gong, and T. Xiang. Recognising action as clouds of space-time interest points. In *Proc. IEEE Conf. CVPR*, 2009.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. CVPR*, 2005.

[5] J. Daugman. Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research*, 20(10):847–856, 1980.

[6] J. Daugman. Uncertainty Relation for Resolution in Space, Spatial Frequency and Orientation Optimized by Two-Dimensional Visual Cortical Filters. *J. Opt. Soc. Amer.*, 2(7):1160–1169, 1985.

[7] K. G. Derpanis, M. Sizintsev, K. Cannons, and R. P. Wildes. Efficient action spotting based on a spacetime oriented structure representation. In *Proc. IEEE Conf. CVPR*, 2010.

[8] K. G. Derpanis, M. Sizintsev, K. Cannons, and R. P. Wildes. Action spotting and recognition based on a spatiotemporal orientation analysis. *IEEE Trans. PAMI*, 35(3):527–540, 2013.

[9] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proc. IEEE Int'l Workshop on VS-PETS*, 2005.

[10] W. T. Freeman and E. H. Adelson. The Design and Use of Steerable Filters. *IEEE Trans. PAMI*, 13(6):891–906, 1991.

[11] D. Gabor. Theory of Communication. *IEE Journal (London)*, 93:429–457, 1946.

[12] C. Georgakis, P. Maragos, G. Evangelopoulos, and D. Dimitriadis. Dominant spatio-temporal modulations and energy tracking in videos: Application to interest point detection for action recognition. In *Proc. ICIP*, 2012.

[13] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. Alvey Vision Conf.*, 1988.

[14] J. P. Havlicek, D. S. Harding, and A. C. Bovik. Multidimensional quasi-eigenfunction approximations and multicomponent am-fm models. *IEEE Trans. Image Processing*, 9(2):227–242, 2000.

[15] D. J. Heeger. Model for the extraction of image flow. *J. Opt. Soc. Amer.*, 4(8):1455–1471, 1987.

[16] D. J. Heeger. Optical flow using spatiotemporal filters. *Int'l J. Comp. Vision*, 1(4):279–302, 1988.

[17] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *Proc. ICCV*, 2007.

[18] J. F. Kaiser. On a simple algorithm to calculate the energy'of a signal. In *Proc. IEEE Conf. ICASSP*, 1990.

[19] A. Kläser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3D-Gradients. In *Proc. BMVC*, 2008.

[20] H. Knutsson and M. T. Andersson. What's so good about quadrature filters? In *Proc. ICIP*, 2003.

[21] J. J. Koenderink and A. van Doorn. Representation of local geometry in the visual system. *Biol. Cybern.*, 55:367–375, 1987.

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

[23] I. Laptev and T. Lindeberg. Local descriptors for spatio-temporal recognition. In *Proc. SCVMA*, 2004.

[24] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. IEEE Conf. CVPR*, 2008.

[25] P. Maragos and A. C. Bovik. Image demodulation using multidimensional energy separation. *J. Opt. Soc. Amer.*, 12(9):1867–1876, 1995.

[26] P. Maragos, J. F. Kaiser, and T. F. Quatieri. Energy separation in signal modulations with application to speech analysis. *IEEE Trans. Signal Processing*, 41(10):3024–3051, 1993.

[27] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *Proc. IEEE Conf. CVPR*, 2009.

[28] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Trans. PAMI*, 26(5):530–549, 2004.

[29] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *Int'l J. Comp. Vision*, 79(3):299–318, 2008.

[30] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010.

[31] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *Proc. IEEE Conf. CVPR*, 2012.

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

[33] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proc. Int'l Conf. on Multimedia*, 2007.

[34] C. Theriault, N. Thome, and M. Cord. Dynamic scene classification: Learning motion descriptors with slow features analysis. In *Proc. IEEE Conf. CVPR*, 2013.

[35] H. Wang, A. Kläser, C. Schmid, and C. Liu. Dense trajectories and motion boundary descriptors for action recognition. *Int'l J. Comp. Vision*, 103(1):60–79, 2013.

[36] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proc. BMVC*, 2009.

[37] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Computer Vision–ECCV 2008*, pages 650–663. Springer, 2008.

[38] Y. Yang, I. Saleemi, and M. Shah. Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions. *IEEE Trans. PAMI*, 35(7):1635–1648, 2013.

[39] Z. Zhang and D. Tao. Slow feature analysis for human action recognition. *IEEE Trans. PAMI*, 34(3):436–450, 2012.