A Tensor Motion Descriptor Based on Multiple Gradient Estimators

Dhiego Sad*, Virgínia F. Mota [†], Luiz M. Maciel*, Marcelo B. Vieira *, Arnaldo de A. Araújo[†]
*Universidade Federal de Juiz de Fora
ICE/DCC, Juiz de Fora, Brazil
[†]Universidade Federal de Minas Gerais
ICEx/DCC, Belo Horizonte, Brazil

***Abstract*—This work presents a novel approach for motion description in videos using multiple band-pass filters which act as first order derivative estimators. The filters response on each frame are coded into individual histograms of gradients to reduce their dimensionality. They are combined using orientation tensors. No local features are extracted and no learning is performed, i.e., the descriptor depends uniquely on the input video. Motion description can be enhanced even using multiple filters with similar or overlapping frequency response. For the problem of human action recognition using the KTH database, our descriptor achieved the recognition rate of 93.3% using three Daubechies filters, one extra filter designed to correlate them, two-fold protocol and a SVM classifier. It is superior to most global descriptor approaches and fairly comparable to the state-of-the-art methods.**

***Keywords*-Multifilter analysis; Motion descriptor; Orientation tensor; Human action recognition.**

I. Introduction

Several works focused on the problem of recognition of human actions in videos in recent years. Many of them were dedicated to extract the most useful data from videos. Among several possibilities, motion is the main feature that represents semantic information in videos. Detect and track objects or persons are of great interest in many applications.

In this work, motion is assumed to be detectable through band-pass filters applied in the frames bidimensional support and in time. As such, multiple filters can be used to extract different spectra, relating the original video frequencies differently by their frequency response. The key point is that each filter correlates the original spectrum in a distinguished manner, and this is useful to capture motion nuances. This was motivated by the fact that even a simple derivative filter like the Sobel operator, applied after a Gaussian filter, can drive 92.01% of recognition rate in KTH classification [1].

The motion information extracted from a video should be represented in a compact form. Another problem is how the motion information detected per frame are combined into a whole video descriptor. We use histograms of gradients for dimension reduction [1] and orientation tensors to accumulate information [2].

Our main contribution is a new method to compute a global motion descriptor based on the application of multiple filters into the video. These filters are derivative operators whose band-pass frequency responses capture different motion nuances. Another contribution is a filter design approach to correlate them in order to obtain a better performance. Using a SVM classifier, our descriptor achieves recognition rates comparable to the state-of-the-art on KTH dataset [3] and superior to the most global descriptors in the literature.

A. Related Works

In [1], the motion information is extracted using the Sobel filter, achieving 92.01% of recognition rate in the KTH dataset. This filter is not always suitable for motion detection and we argue that applying multiple filters is better to extract subtle motion characteristics in each video.

Laptev et al [4] present a combination of HOG with histogram of optic flow (HOF) to characterize local motion and shape. Histograms of spatial gradient and optical flow are computed and accumulated in space-time neighborhoods of detected interest points. Similarly to the SIFT descriptor, normalized histograms are concatenated to HOG and HOF vectors. The final descriptor is computed through a bag-of-features technique.

In [5], HOG, HOF, MBH (motion boundary histograms) and trajectory are combined in order to create a better motion descriptor. A standard bag-of-features approach is used constructing a codebook for each descriptor (trajectory, HOG, HOF, MBH) separately. A SVM classifier is then used in the context of action classification for the KTH, Hollywood2, UCF11 and UCF sports datasets.

In [6] is proposed a method for creating descriptors based on wavelet transform. The first step of the method is to detect interest points and extract cuboids around these points. To create the descriptor, Daubechies wavelets are applied inside the cuboid to obtain information within them. A bag-of-features technique is also used in this work. Finally, in the classification step, it is used a support vector machine (SVM) with radial basis function kernel (RBF).

Minhas et al [7] present a combination of spatio-temporal features and local static features. Complex wavelet coefficients in different sub-bands are represented by lower dimensional vectors obtained using two-dimensional PCA. Dual-tree complex wavelet transform (DT-CWT) is constructed by designing an appropriate pair of orthogonal or biorthogonal filter banks that work in parallel. To determine local static features, affine SIFT descriptors are computed.

The use of local features for human action recognition is more exploited, as they provide higher recognition rates. Hence, there are few references about global descriptors which do not rely on locally based features. Global approaches, however, are much simpler to compute and can achieve fast and fairly high recognition rates.

A global descriptor based on the histogram of oriented gradients is presented in [8], using the Weizmann database. The descriptor is computed using several time scales. From each scale, the gradient for each pixel is computed, resulting in a HOG for each video. The histograms are compared to classify the database.

Solmaz et al [9] present a global descriptor based on bank of 68 Gabor filters. For each video, they extract a fixed number of clips and compute the 3-D Discrete Fourier Transform. Applying each filter of the 3-D filter bank separately to the frequency spectrum, the output is quantized in fixed subvolumes. They concatenate the outputs and perform dimension reduction using PCA and classification by a SVM.

II. Proposed Method

A. Motion extraction with multiple filters

The first step of our method is to apply a 5×5 Gaussian low-pass filter in all video frames. This is necessary to smooth the noisy highest frequencies. The noise reduction performed by the Gaussian filter showed to be relevant. The high frequency attenuation, however, should not be strong in order to preserve significant motion information. Indeed, other filter sizes were used but resulting in lower performances. The subsequent gradient estimators are, thus, affected by this preprocessing.

In this work, a unidimensional filter is defined by a pair of impulse responses $(\mathrm{H}_a, \mathrm{G}_a)$, where $a \in \{1, 2, \cdots, f\}$ is the filter index, f is the number of filters for motion detection, G_a has high-pass frequency response, and H_a has low-pass response. Their multidimensional filter version is separable, having H_a and G_a as factors.

To capture motion information, G_a is usually a derivative estimator with frequency response $\widetilde{\mathrm{G}}_a$. For multidimensional signals, H_a attenuates the noise on orthogonal directions. In this work, its frequency response $\widetilde{\mathrm{H}}_a$ is assumed to have some degree of complementarity in relation to $\widetilde{\mathrm{G}}_a$, in order to attenuate undesired correlated noise among the main axes.

The partial derivatives, or gradient, resulted from the application of a filter a on the j-th video frame I_j, at point p, is defined as the vector

$$\vec{g} = [dx_p^a \; dy_p^a \; dt_p^a]^T = \left[\frac{\partial I_j(p)}{\partial x} \; \frac{\partial I_j(p)}{\partial y} \; \frac{\partial I_j(p)}{\partial t}\right]^T,$$

or, equivalently, in spherical coordinates $\left[\rho_p^a \;\; \theta_p^a \;\; \psi_p^a\right]$ with $\theta_p^a \in [0,\ \pi]$, $\psi_p^a \in [0,\ 2\pi)$ and $\rho_p^a = ||\vec{g}||$. It indicates brightness variation that might be the result of local motion. The dx_p^a component is computed by firstly filtering the video in orthogonal directions Y and time using H_a, and afterwards in the main direction X using G_a. The same logic is used to obtain dy_p^a and dt_p^a.

We chose to apply wavelets as derivative estimators because of their widespread use. Note that the Gaussian low-pass filtering in image space followed by the application of a high-pass filter results in a band-pass frequency response.

B. Computing the HOG3D for each frame

The filtered output of a frame I_j, with n points p, can be compactly represented by a tridimensional histogram of gradients $\vec{h}_j^a = \{h_{k,l}^a\}$, $k \in [1, nb_\theta]$ and $l \in [1, nb_\psi]$, where nb_θ and nb_ψ are the number of cells for θ and ψ coordinates respectively. There are several methods for computing the HOG3D and we chose, for simplicity, an uniform subdivision of the angle intervals to populate the $nb_\theta \cdot nb_\psi$ bins:

$$h_{k,l}^a = \sum_p \rho_p^a \cdot w(dist_{k,l}^{q,r}),$$

where $dist_{k,l}^{q,r}$ is the Euclidean distance between the integer bin indices (k, l) and the mapped real coordinates $(q, r) = (1 + \frac{nb_\theta \cdot \theta_p^a}{\pi}, 1 + \frac{nb_\psi \cdot \psi_p^a}{2\pi})$ of the gradient at point p, and $w(dist_{k,l}^{q,r})$ is a Gaussian weighting function [10] with $\alpha = 1.0$. The whole gradient field of the j-th frame is then represented by the vector $\vec{h}_j^a$ with $nb_\theta \cdot nb_\psi$ elements. All results in this work were computed using $nb_\theta = 8$, $nb_\psi = 16$ [1]. Note that there is one HOG3D per frame and per filter.

To empirically reduce interframe brightness unbalance, the histogram of gradients $\vec{h}_j^a \in \mathbb{R}^{nb_\theta \cdot nb_\psi}$ might have all of its elements $h_{k,l}^a$ optionally adjusted to ${h_{k,l}^a}^\gamma$, with $\gamma = 0.5$. This is called power normalization and reduces the relative differences between gradient bins. This is applied only for the correlation filters to improve their performance.

1) Orientation tensor: coding HOG3D coefficients: An orientation tensor is a $m \times m$ real symmetric matrix, for m-dimensional signals. Given a vector $\vec{v}$ with m elements, it can be represented by the tensor $T = \vec{v}\vec{v}^T$. Note that the well known structure tensor is a specific case of orientation tensor [11] and it has been used for image [12] and video classification [1], [2]. The tensor of the frame I_j using the filter a is:

$$T_j^a = \vec{h}_j^a {\vec{h}_j^a}^T,$$

that carries the information of the gradient distribution of the j-th frame, computed using filter a. Individually, this tensor has the same information of $\vec{h}_j^a$. Since T_j^a is a symmetric matrix, it can be stored with $\frac{m(m+1)}{2}$ elements.

C. Filter based tensor descriptor: series of frame tensors

For a given derivative filter a, we have to express the motion average of consecutive frames using a series of tensors. The average motion of a video can be given by $T^a = \sum_j T_j^a$ using all of its frames or an interval of interest. By normalizing T^a with a L_2 norm, we are able to compare different video clips or snapshots regardless their length or image resolution.

If the accumulation series diverges, we obtain an isotropic tensor which does not hold useful information extracted using the derivative estimator pair a. But, if the series converge as an anisotropic tensor, it carries meaningful average motion information of the frame sequence. The conditions of divergence and convergence need further studies.

1) Frame subdivision: When a gradient histogram is computed using the whole image, the cells are populated with vectors regardless their position in the image. This implies in a loss of the correlation between the gradient vectors and their neighbors. As observed in several works [10], the subdivision of the video into cubes of frames enhances the recognition rate.

Suppose the video frame j is uniformly subdivided in $\vec{x}$ and $\vec{y}$ directions by a grid with n_x and n_y non-overlapping blocks. Each block can be viewed as a distinct video varying in time. The smaller images result in gradient histograms $\vec{h}^a_j(c,r)$, $c \in [1, n_x]$ and $r \in [1, n_y]$, with better position correlation. The tensor for frame j, using derivative filter a, is then computed as the addition of all block tensors:

$$T^a_j(c,r) = \sum_{c,r} \vec{h}^a_j(c,r)\ \vec{h}^a_j(c,r)^T ,$$

which captures the uncertainty of the direction of the m-dimensional histogram vectors $\vec{h}^a_j(c,r)$ for the frame j. The tensor series become:

$$T^a = \sum_{j,c,r} \frac{T^a_j(c,r)}{||T^a_j(c,r)||},$$

where a is the derivative filter used, j is the frame index, and $(c,r) \in [1, n_x] \times [1, n_y]$ are the subimage coordinates.

The final video tensor descriptor for the derivative filter a is then $T^a/||T^a||$. It has the same number of elements of the version without image subdivision.

D. Global video descriptor: concatenating multiple filter based tensors

We propose the concatenation of the individual tensors, computed for all the f filter pairs, to form the final descriptor for the input video:

$$T = \{T^1, T^2, \cdots, T^f\}.$$

Despite other combination methods are possible, concatenation preserves the motion information extracted by each filter. The drawback is that the number of coefficients in the descriptor is multiplied by the number of filters f. In this work, the HOG3D has 128 bins yielding tensors with 8256 elements for a single filter. A video descriptor using four derivative filters has 33024 elements, slowing down SVM classification.

III. Results

Validation set. To validate our descriptor, we use the KTH dataset [3] since a major number of works in literature provide clearly reproducible results for it.

Classification protocol. We run a two-fold strategy on a non-linear SVM classifier. All results were computed using $nb_\theta = 8$, $nb_\psi = 16$ [1], leading to HOG3D with 128 bins per frame. The tensor for one filter has then 8256 elements.

Derivative estimators. We used the decomposition filter pairs of several wavelet families. Results with a single pair are presented in Table I. The recognition rates differ slightly among them. The frame subdivision with 8×8 subimages enhanced almost all results. The Daubechies wavelet family is adequate to separate the spectrum of dyadic bands, with an easy way to control the null moments. Due this and based on the best results achieved, we restricted the subsequent experiments to the Daubechies wavelet family with a 8×8 grid.

Filter	Grid 1x1	Grid 8x8
db1	87.8%	90.9%
bio1.3	86.0%	90.6%
sym	83.6%	89.9%
db2	83.9%	88.8%
cf1	82.8%	87.5%
db4	82.8%	83.6%
db5	81.1%	82.9%
db6	78.6%	82.9%
db3	80.0%	82.7%
cf2	83.9%	82.0%
db7	79.3%	81.7%
db8	79.3%	81.1%

TABLE I
Recognition rates using decomposition filters of several wavelet families, without power normalization.

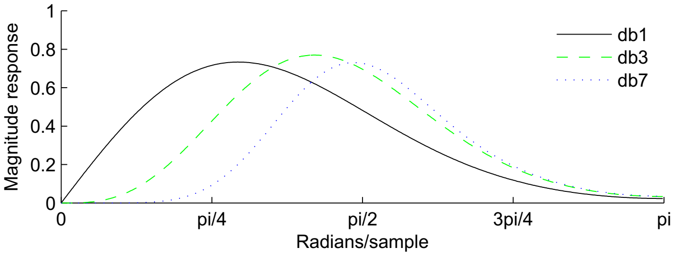

Fig. 1. Transfer functions of db1, db3 and db7, modulated by a Gaussian filter.

The best result using a single filter is achieved with the db1 pair, reaching 90.9% of recognition rate. Regarding the frequency response of db1, combined with the low-pass Gaussian filter (Fig. 1), one may note that it preserves better the low frequencies in the first quarter of the spectrum, compared with db3 and db7 responses. The db3 and db7 filters, however, give much lower recognition rates: 82.7% and 81.7%, respectively. It seems that some mid or high frequencies are noise, while some low frequencies are suitable for classification with the KTH dataset. Finding a proper combination of the response of several filters can lead to a better performance.

Table II shows the results for several filter combinations. The db1 filter is a good choice for all sets because of its isolated performance. Using db2 and db3, for example, gave only 87.5% of recognition rate, against 92.1% for db1 and db2, and 91.5% for db1 and db3. For two filters, the best combination found was db1 and db7 with 92.6% of recognition rate.

Based on the results with one and two filters, we tested some combinations with three filters. The best results are also presented in Table II. The db1, db3 and db7 filters, combined in a descriptor with 24768 elements, achieved 93.2% of recognition rate.

Filter pairs	Rate (%)
db1 db2	92.1
db1 db3	91.5
db1 db4	91.2
db1 db6	92.2
db1 db7	92.6

Filter pairs	Rate (%)
db1 db8	91.9
db2 db3	87.5
db1 db3 db7	93.2
db1 db3 db8	92.5
db1 db3 db10	92.7

TABLE II
Recognition rates for multiple filter combinations.

A. *Correlation filter design*

We combine the best filters of Tables I and II in a single linear filter to get even better results. Our proposal is to derive a pair $(\mathrm{H}_{f+1}, \mathrm{G}_{f+1})$ such that

$$|\widetilde{\mathrm{H}}_{f+1}(\omega)| = \sum_{a=1}^{f} |\widetilde{\mathrm{H}}_a(\omega)| \quad \text{and} \quad |\widetilde{\mathrm{G}}_{f+1}(\omega)| = \sum_{a=1}^{f} |\widetilde{\mathrm{G}}_a(\omega)|,$$

i.e., the magnitude response is the same as the sum of the magnitude of the $f > 1$ filters. Using db1, db3 and db7, for example, gives the db$_{1,3,7}$ filter whose normalized high pass magnitude response is depicted in Figure 2. It alone gives 85.5% of recognition: slightly better than the average 85.1% of recognition of its components (Table I).

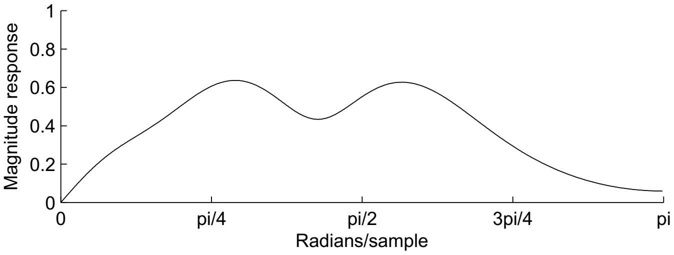

Fig. 2. Transfer function of the correlation filter db$_{1,3,7}$, modulated by a Gaussian filter.

Filters	Recognition rate (%)
db1 db3 db7 db$_{1,3,7}$	**93.3**
db1 db3 db8 db$_{1,3}$	92.9
db1 db3 db8 db$_{1,3,10}$	93.1

TABLE III
Recognition rates using correlation filters.

Table III shows recognition rates using four filters, one of which is a correlation filter derived as above. Note that the db$_{1,3,7}$ and db$_{1,3}$ augmented the rates compared to Table II. Our best result is 93.3% using db1, db3, db7 and their correlation filter db$_{1,3,7}$. The corresponding confusion matrix is shown in Table IV.

	Box	HClap	HWav	Jog	Run	Walk
Box	95.8	2.8	0.0	0.0	0.0	1.4
HClap	2.1	96.5	1.4	0.0	0.0	0.0
HWav	6.2	0.0	93.8	0.0	0.0	0.0
Jog	0.7	0.0	0.0	90.3	6.2	2.8
Run	0.0	0.0	0.0	12.5	86.8	0.7
Walk	0.0	0.0	0.0	3.5	0.0	96.5

TABLE IV
Confusion matrix for the best result, 93.3% using four filters: db1, db3, db7, db$_{1,3,7}$.

Global Methods	Recognition rate (%)
HOG pyramids [8]	72.00
Bank of Gabor filters [9]	92.00
HOG3D + Tensor [1]	92.01
Our Method (4 filters)	**93.30**

Local Methods	Recognition rate
Harris3D + HOG/HOF [4]	91.80
Interest Points + Wavelets [6]	93.89
HOG+HOF+MBH+Trajectory [5]	94.20

TABLE V
Comparison with other methods for the KTH dataset.

A comparison with the state-of-the-art methods is presented in Table V. The proposed method achieves a competitive accuracy with a much simpler global approach, using only the information from derivative estimators, without any bag-of-features strategy [4], [6], [5]. Moreover, some authors use leave-one-out protocol, like Minhas et al [7] who achieves 94.83% of recognition rate. Using this protocol, our method results in 95.5% of recognition rate using the db1, db3 and db10 filters.

Results for other datasets. Based on these results for KTH dataset, we tested the same Daubechies filter combination on more challenging datasets, the UCF11 [13] and the Hollywood2 [14]. Tables VI and VII shows the results for several combinations using db1, db3, db7 and their correlation filter db$_{1,3,7}$ for UCF11 and Hollywood2 datasets, respectively. All results were computed using $nb_\theta = 8$, $nb_\psi = 16$ and a subdivision of 8×8.

Filters	Recognition rate(%)
db1	70.1
db3	63.4
db7	51.2
db1 db3	**72.6**
db1 db7	71.9
db3 db7	64.4
db1 db3 db7	72.3
db1 db3 db7 db$_{1,3,7}$	72.5

TABLE VI
Recognition rates using several combinations of filters for UCF11 dataset.

Perez et al [1] did not present results for UCF11 dataset. Thus, we evaluated the descriptor using the same parameters as ours, $nb_\theta = 8$, $nb_\psi = 16$ and a subdivision of 8×8. The recognition rate was 67.5%. We can see in Table VI that using db1 instead of Sobel operator and its combination with db3 improved the result, achieving 72.6%. Therefore, the combination of multiple filters also improved the recognition for this dataset. However, the filter combination which achieves the best result for KTH dataset did not improve the rate. This shows that further studies are needed to discover the best filter combination and to design a correlation filter for UCF11 dataset.

Filters	**Recognition rate(%)**
db1	**41.9**
db3	30.7
db7	24.1
db1 db3	41.9
db1 db7	40.3
db3 db7	30.4
db1 db3 db7	40.4
db1 db3 db7 db$_{1,3,7}$	40.7

TABLE VII
RECOGNITION RATES USING SEVERAL COMBINATIONS OF FILTERS FOR HOLLYWOOD2 DATASET.

Results for Hollywood2 (Table VII) show that it is better to use db1 instead of Sobel operator. In [1] the best result reported is 34.0% achieved with a 4x4 grid and a HOG 8x16. The filter combinations used for KTH did not improve the rate for this dataset. However, it is important to notice that further investigations are needed to find out which bands are more representative of motion with this dataset.

IV. CONCLUSION

In this paper, we presented a novel global approach for motion description in videos using multiple filters that act as first order derivative estimators. Our main contribution is to provide a method for combining the response of multiple filters using tensors and by the derivation of a correlation filter. It is an effective approach reaching 93.3% of recognition rate with KTH, fairly comparable to the best local and learning-based methods, as depicted in V.

Our method indicate that it is possible to improve recognition rate for human action datasets using multiple filters. Each filter contributes with information corresponding to its transfer spectrum. Once we discover the best combinations, a band-pass filter can be derived in order to combine spectrum bands. Thus, the use of multiple filters is promising for the problem of motion description.

Further studies are needed to improve filter correlation design. It is important to note that an adequate bandpass filter could exist and be different for each action dataset. As a future work, it is important to check the existence of an optimal filter, or set of filters, that improves the detection of frequencies indicating motion in a given dataset.

ACKNOWLEDGMENT

The authors would like to thank to CAPES, CNPq and FAPEMIG for funding.

REFERENCES

[1] E. A. Perez, V. F. Mota, L. M. Maciel, D. Sad, and M. B. Vieira, "Combining gradient histograms using orientation tensors for human action recognition," in *International Conference on Pattern Recognition*, 2012, pp. 3460–3463.

[2] V. F. Mota, E. A. Perez, M. B. Vieira, L. M. Maciel, F. Precioso, and P.-H. Gosselin, "A tensor based on optical flow for global description of motion in videos," in *SIBGRAPI 2012 (XXV Conference on Graphics, Patterns and Images)*, Ouro Preto, MG, Brazil, august 2012, pp. 298–301.

[3] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *In Proc. ICPR*, 2004, pp. 32–36.

[4] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision & Pattern Recognition*, jun 2008. [Online]. Available: http://lear.inrialpes.fr/pubs/2008/LMSR08

[5] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin, "Action Recognition by Dense Trajectories," in *IEEE Conference on Computer Vision & Pattern Recognition*, Colorado Springs, United States, Jun. 2011, pp. 3169–3176. [Online]. Available: http://hal.inria.fr/inria-00583818

[6] L. Shao and R. Gao, "A wavelet based local descriptor for human action recognition," in *Proc. BMVC*, 2010, pp. 72.1–10, doi:10.5244/C.24.72.

[7] R. Minhas, A. Baradarani, S. Seifzadeh, and Q. J. Wu, "Human action recognition using extreme learning machine based on visual vocabularies," *Neurocomputing*, vol. 73, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231210001517

[8] L. Zelnik-manor and M. Irani, "Event-based analysis of video," in *In Proc. CVPR*, 2001, pp. 123–130.

[9] B. Solmaz, S. M. Assari, and M. Shah, "Classifying web videos using a global video descriptor," *Machine Vision and Applications*, pp. 1–13, Sep. 2012. [Online]. Available: http://dx.doi.org/10.1007/s00138-012-0449-x

[10] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ser. ICCV '99. Washington, DC, USA: IEEE Computer Society, 1999, pp. 1150–. [Online]. Available: http://dl.acm.org/citation.cfm?id=850924.851523

[11] B. Johansson, G. Farnebck, and G. F. Ack, "A theoretical comparison of different orientation tensors," in *Symposium on Image Analysis*. SSAB, 2002, pp. 69–73.

[12] R. Negrel, D. Picard, and P. H. Gosselin, "Using spatial pyramids with compacted vlat for image categorization," in *ICPR*, 2012, pp. 2460–2463.

[13] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," *IEEE International Conference on Computer Vision and Pattern Recognition*, 2009.

[14] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *Conference on Computer Vision & Pattern Recognition*, jun 2009.