A tensor motion descriptor based on histograms of gradients and optical flow

Mota V. F.[1a,b], Perez E. A.[a], Maciel L. M.[a], Vieira M. B.[a], Gosselin, P. H.[c]

[a]*DCC/ICE, Universidade Federal de Juiz de Fora, Juiz de Fora, Brazil*
[b]*DCC/ICEx, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil*
[c]*INRIA Rennes Bretagne Atlantique*

Abstract

This paper presents a new tensor motion descriptor only using optical flow and HOG3D information: no interest points are extracted and it is not based on a visual dictionary. We propose a new aggregation technique based on tensors. This is a double aggregation of tensor descriptors. The first one represents motion by using polynomial coefficients which approximates the optical flow. The other represents the accumulated data of all histograms of gradients of the video. The descriptor is evaluated by a classification of KTH, UCF11 and Hollywood2 datasets, using a SVM classifier. Our method reaches 93.2% of recognition rate with KTH, comparable to the best local approaches. For the UCF11 and Hollywood2 datasets, our recognition achieves fairly competitive results compared to local and learning based approaches.

Keywords: Global motion descriptor, optical flow, histogram of gradients, action recognition

[1]Corresponding author: Tel: +55 32 88856321
E-mail address: virginiaferm@dcc.ufmg.br

1. Introduction

Human action recognition is a very attractive field of research as it is a key part in several areas such as video indexing, surveillance, human-computer interfaces, among others. Most works address this problem by a motion analysis and a representation step. Several descriptors were proposed over the past years, most of them using some motion representation, because it is one of the main characteristics that describe the semantic information of videos. Some examples of motion representations are the histogram of gradients and optical flow.

Usually the optical flow itself is not used as a descriptor. Instead, its histogram is largely associated with other features in order to improve the recognition rate [1, 2]. In our preliminary work, presented in [3], we showed that the modeling of optical flow vector fields gives a consistent global motion descriptor. This descriptor is obtain using the parameters of a polynomial model for each frame of a video. The coefficients were found through the projection of the optical flow on Legendre polynomials, reducing the dimension of the motion estimation per frame. The sequence of coefficients were then combined using orientation tensors.

This work is motivated by the possibility of combining the tensor descriptor presented in [3] with other global features. Indeed, the optical flow projected onto Legendre polynomial basis captures a specific nuance of the underlying motion. Its combination with other motion representations can improve the results and drive a competitive recognition for the problem of human action recognition.

Our main contribution is a new motion descriptor based on orientation

tensor which uses only optical flow [3] and tridimensional histogram of gradients (HOG3D) information [4]: no interest points are extracted and no bag-of-features strategy is used. The global tensor descriptor created is evaluated by a classification of KTH [5], UCF11 (also known as UCF YouTube) [6] and Hollywood2 [7] video datasets with a non-linear SVM classifier.

2. Related work

Laptev et al [2] present a combination of histograms of gradients (HOG) with histogram of optical 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. Then, the signature of the video is computed through a bag-of-features technique.

In [1], HOG, HOF, MBH (motion boundary histogram) and trajectory are combined in order to create a better motion descriptor. For each descriptor type, bag-of-features are computed thanks to a visual codebook. A SVM classifier is then used in the context of action classification for the KTH, Hollywood2, UCF11 and UCF sports datasets.

Also using a bag-of-features strategy, Zhen and Shao [8] presents a new descriptor for action recognition based on Laplacian pyramid coding. The idea is to represent the video by the combination of motion history images and three orthogonal planes, obtained from a set of cuboids extracted from the video sequence. Then, this information is encoded with a Laplacian pyramid model and the final video representation is computed thanks to an improved

version of bag-of-features using the soft-assignment coding and max pooling.

Kobayashi and Otsu [9] propose motion features based on co-occurrence histograms of the space-time 3D gradient orientations. They are employed for frame based features to densely characterize the motion. These frame-based features are extracted from sub-sequences densely sampled along the time axis. Thus, they describe a bag-of-frame-features approach to create the video feature.

The use of local features for human action recognition is more exploited, as they provide higher recognition rates. In general, these approaches use bag-of-features (BoF) strategy. Hence, there are few references about global descriptors which do not rely on a visual dictionary and are uniquely dependent on the video, instead of the whole training set as such in BoF method. Global approaches, however, are much simpler to compute and can achieve fast and fairly high recognition rates.

Zelnik et al presents a global descriptor based on histogram of gradients [10]. This descriptor is applied on the Weizmann video database and is obtained with the extraction of multiple temporal scales through the construction of a temporal pyramid. To calculate this pyramid, they apply a lowpass filter on the video and sample it. For each scale, the intensity of each pixel gradient is calculated. Then, a histogram of gradients is created for each video and compared with others histograms to classify the database.

In order to obtain a global descriptor on the KTH dataset, Laptev et al [11] apply the Zelnik descriptor [10] in two different ways: using multiple temporal scales like the original and using multiple temporal and spatial scales.

Solmaz et al [12] 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 sub-volumes. They concatenate the outputs and perform dimension reduction using PCA and classification by a SVM.

3. Proposed Method

3.1. Tensor based on optical flow approximation

The basic idea of a polynomial based model is to approximate a vector field with a linear combination of orthogonal polynomials [13, 14]. Let us define F an optical flow:

$$F : \begin{array}{l} \Omega \subset R^2 \rightarrow R^2 \\ (x_1, x_2) \mapsto (V^1(x_1, x_2), V^2(x_1, x_2)) \end{array}$$

where the functions $V^1(x_1, x_2)$ and $V^2(x_1, x_2)$ corresponds to the horizontal and vertical displacement of the point $(x_1, x_2) \in \Omega$.

This optical flow is then approximated by projecting the displacement functions onto each polynomial $P_{i,j}$, which belong to an orthogonal basis, as such Legendre basis.

In that way, it reduces the dimension of the optical flow field. Thus, we can express $\tilde{F} = (\tilde{V}^1(x_1, x_2), \tilde{V}^2(x_1, x_2))$, using a basis of degree g, as:

$$\begin{cases} \tilde{V}^1(x_1, x_2) = \sum_{i=0}^{g} \sum_{j=0}^{g-1} \tilde{v}_{i,j}^1 P_{i,j} \\ \tilde{V}^2(x_1, x_2) = \sum_{i=0}^{g} \sum_{j=0}^{g-1} \tilde{v}_{i,j}^2 P_{i,j} \end{cases}$$

where

$$\begin{cases} \tilde{v}^1_{i,j} = \int\int_\Omega V^1(x_1, x_2) P_{i,j} \omega(x_1, x_2) dx_1 dx_2 \\ \tilde{v}^2_{i,j} = \int\int_\Omega V^2(x_1, x_2) P_{i,j} \omega(x_1, x_2) dx_1 dx_2 \end{cases} \quad (1)$$

It is important to note that the number of polynomials which composes a basis of degree g is:

$$n_g = \frac{(g+1)(g+2)}{2}$$

3.1.1. Orientation tensor: coding frame coefficients

An orientation tensor is a representation of local orientation which takes the form of an $m \times m$ real symmetric matrix for m-dimensional signals [15].

Given the vector $\vec{v}$ with m elements, it can be represented by the tensor $T = \vec{v}\vec{v}^T$. It is desired that the eigenvector with the largest eigenvalue of the tensor points out the dominant direction of the signal. A signal with no dominant direction is represented by an isotropic tensor, i.e. the three eigenvalues are approximately equal. It is important to note that the well known structure tensor is a specific case of orientation tensor [16].

In order to capture the motion variation in time, we can use both the polynomial coefficients $\tilde{v}^1_{i,j}$ and $\tilde{v}^2_{i,j}$ (Eq. 1) and an approximation of their first temporal derivative $\partial\tilde{v}^q_{i,j} = \tilde{v}^q_{i,j}(f) - \tilde{v}^q_{i,j}(f-1)$ with $i + j < g$, to create a vector $\tilde{v_f}$ for each frame f of the video:

$$\tilde{v}_f = [\tilde{v}^1_{0,0}, ..., \tilde{v}^1_{g,0}, \quad \tilde{v}^2_{0,0}, ..., \tilde{v}^2_{g,0}, \quad \partial\tilde{v}^1_{0,0}, ..., \partial\tilde{v}^1_{g,0}, \quad \partial\tilde{v}^2_{0,0}, ..., \partial\tilde{v}^2_{g,0}].$$

Using the vector $\tilde{v_f}$, we generate an orientation tensor $T_f = \tilde{v_f}\tilde{v_f}^T$ for each frame f of the video, which is a $4n_g \times 4n_g$ matrix. This orientation tensor captures the covariance information between $\tilde{v}^1_{i,j}$ and $\tilde{v}^2_{i,j}$. It carries only the information of the polynomial of frame f and its rate of change in time.

3.1.2. Global tensor descriptor

We have to express the motion average of consecutive frames using a series of tensors. This can be achieved by $T^{OF} = \sum_a^b T_f$ using all video frames or an interval of interest. By normalizing T_f 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 motion information. 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.

Instead of using the entire optical flow of the video frames, it is also possible to use only the optical flow from a region with most representative motion. Then, we tested a sliding window with fixed dimensions placed around the subject who is doing the action. The center of mass of global optical flow gives the center of the window.

The accumulated tensor is symmetric, therefore we can use only a triangular superior (or inferior) matrix to represent the video, which reduces the number of coefficients of the final tensor descriptor.

3.2. Tensor based on histogram of gradients

The partial derivatives of the j-th video frame at point p

$$\vec{g}_t(p) = [dx\ dy\ dt] = \left[\frac{\partial I_j(p)}{\partial x}\ \frac{\partial I_j(p)}{\partial y}\ \frac{\partial I_j(p)}{\partial t}\right],$$

or, equivalently, in spherical coordinates $\vec{s}_t(p) = [\rho_p\ \ \theta_p\ \ \psi_p]$ with $\theta_p \in [0,\ \pi]$, $\psi_p \in [0,\ 2\pi)$ and $\rho_p = ||\vec{g}_t(p)||$, indicate brightness variation that might be

the result of local motion.

The gradient of all n points of the image I_j can be compactly represented by a tridimensional histogram of gradients $\vec{h}_j = \{h_{k,l}\}$, $k \in [1, nb_\theta]$ 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. We performed some experiments with the icosahedron discretization [17], however no significant enhancement was detected. Thus, we chose an uniform subdivision of the angle intervals to populate the $nb_\theta \cdot nb_\psi$ bins (Eq. 2), since it achieves good results and it is fast to compute.

$$h_{k,l} = \sum_p \rho_p \cdot w_p, \tag{2}$$

where $\{p \in I_j \mid k = 1 + \left\lfloor \frac{nb_\theta \cdot \theta_p}{\pi} \right\rfloor, l = 1 + \left\lfloor \frac{nb_\psi \cdot \psi_p}{2\pi} \right\rfloor\}$ are all points whose angles map to k and l bins, and w_p is a per pixel weighting factor which can be uniform or gaussian as in [18]. The whole gradient field is then represented by a vector $\vec{h}_j$ with $nb_\theta \cdot nb_\psi$ elements.

3.2.1. Global tensor descriptor: coding HOG3D coefficients as tensors

Analogously to the previous descriptor (Sec. 3.1.2), the HOG3Ds with m bins $\vec{h}_j$, computed for j-th frames, can be combined in a tensor as following:

$$T^{HOG} = \sum_j \vec{h}_j \vec{h}_j^T,$$

using all video frames or an interval of interest. By normalizing T^{HOG} with a L_2 norm, we are able to compare different video clips or snapshots regardless their length or image resolution.

3.2.2. Global tensor descriptor: subdividing the frame using a grid

When the 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 [18], the subdivision of the video into cubes of frames enhances the recognition rate, using a gaussian weight for w_p.

Suppose the video frame f 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 $\vec{h}_j^{c,r}$, $c \in [1, n_x]$ and $r \in [1, n_y]$, with better position correlation. The tensor for the frame j is then computed as the addition of all block tensors:

$$T_j = \sum_{c,r} \vec{h}_j^{c,r} \; \vec{h}_j^{c,r\,T}, \tag{3}$$

which captures the uncertainty of the direction of the m-dimensional vectors $\vec{h}_f^{a,b}$ in the frame j. This tensor is normalized using the L_2 norm. The image subdivision does not change the descriptor size and the accumulation described above is the same. The global descriptor with image subdivision and histograms of gradients is then

$$T^{HOG} = \sum_{j=1}^{f} T_j.$$

Another improvement is to accumulate the tensor obtained with the video frame flipped horizontally. Therefore, the HOG3D is computed for each block, the final tensor is computed (Eq. 3) and simply added to the original

frame tensor. This flipped version enforces horizontal gradient symmetries that occur on the video, even those between multiple frames. In our experiments (Sec. 4) all HOG3D descriptors are obtained using this improvement.

3.3. Combining orientation tensors

We propose to concatenate the individual tensors, computed with the optical flow approximation (Sec. 3.1) and HOG3D (Sec. 3.2), to form the final descriptor for the input video:

$$T = \{T^{OF}, T^{HOG}\}. \tag{4}$$

Despite other combination methods are possible, concatenation preserves the motion information extracted by each individual descriptor. The information of those descriptors are complementary and can improve the recognition rate.

This descriptor depends only on the video itself, not requiring any recomputation of the previously computed descriptors after the addition of new videos and/or new action categories to the dataset.

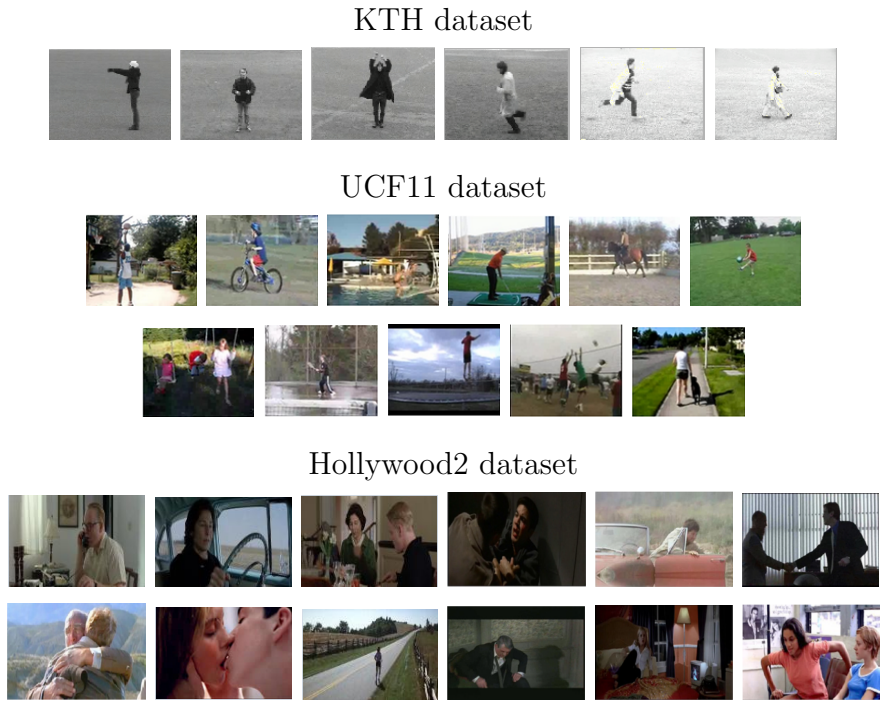
It is important to note that the nature of these two tensors is different and, as such, need to be equalized. One possible way is to use a power normalization in one of the descriptors. Experimentally, best results were obtained by normalizing the HOG3D tensor: the T^{HOG} descriptor in Equation 4 has all of its elements a_k adjusted to a_k^{γ}, $\gamma \in]0, 1]$.

4. Experimental results

We compute the optical flow using the method described by Augereau et al [19]. This method was chosen because we found experimentally that it

computes a more regular optical flow than the one computed by the standard Lucas-Kanade [20]. The descriptor is evaluated by a classification of KTH, UCF11 and Hollywood2 datasets (Figure 1), using a SVM classifier.

Figure 1: Examples of videos from KTH, UCF11 and Hollywood2 datasets.



4.1. KTH Dataset

The KTH actions dataset [5] consists of six human action classes: walking, jogging, running, boxing, waving, and clapping. Each action class is performed several times by 25 subjects. The sequences were recorded in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. The background is homogeneous and static in

most sequences. In total, the data consists of 2391 video samples. We use the same evaluation protocol of the original paper [5], as [1].

The best optical flow descriptor for KTH dataset was obtained with a sliding window with fixed dimensions put around the subject who is doing the action. The center of mass of global optical flow gives the center of the window. It works for KTH scenes because they have only one person acting and a nearly static background. Table 1 shows the recognition rates for this descriptor using a sliding window of 60x100 pixels. The best recognition rate was 87.8% with polynomials of degree 8, leading to a descriptor with 16290 elements.

In [4] is reported that the best result is achieved with a grid 8x8 and 128 bins obtaining 92.0% of recognition rate. Thus, we concatenate our optical flow tensor descriptor with this HOG3D to form a new global motion descriptor. Table 2 shows the recognition rates for several degrees. The best recognition rate was 93.2% with polynomials of degree 5 (3670 elements) concatenated with a HOG3D of 128 bins (8256 elements). The confusion matrix is presented in Table 3.

Table 1: Recognition rates of KTH dataset for several degrees using only the optical flow descriptor with a sliding window with dimensions 60x100.

Degree	1	2	3	4	5	6	7	**8**	9	10
Rate (%)	81.8	85.5	85.9	86.7	85.8	87.6	87.6	**87.8**	87.5	87.1

4.2. UCF11 Dataset

The UCF11 dataset [6] consists in 11 action categories: basketball shooting, biking/cycling, diving, golf swinging, horse back riding, soccer juggling,

Table 2: Recognition rates of KTH dataset for several degrees using a sliding window with dimensions 60x100 concatenated with a HOG3D [4] of 128 bins with $\gamma = 1$.

Degree	1	2	3	4	**5**	6	7	8	9	10
Rate (%)	92.6	92.1	92.5	92.7	**93.2**	92.1	92.0	91.3	91.8	90.6

Table 3: Confusion matrix of KTH dataset for the best result, 93.2% using polynomials of degree 5 and a HOG3D of 8x16 with grid 8x8 [4].

	Box	HClap	HWav	Jog	Run	Walk
Box	95.8	0.0	0.0	0.0	0.0	4.2
HClap	6.2	93.8	0.0	0.0	0.0	0.0
HWav	0.7	4.2	95.1	0.0	0.0	0.0
Jog	0.0	0.0	0.0	90.3	6.9	2.8
Run	0.0	0.0	0.0	14.6	84.7	0.7
Walk	0.0	0.0	0.0	0.7	0.00	99.3

swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog. We use the same evaluation protocol of the original paper [6].

The sliding window is not interesting for this dataset because its actions are more complex than those in KTH dataset. Table 4 shows the recognition rates for several degrees using the optical flow tensor descriptor. The best recognition rate was 57.8% with polynomials of degree 12 (66430 elements).

Table 4: Recognition rates of UCF11 dataset for several degrees using only the optical flow descriptor.

Degree	1	2	3	4	5
Rate (%)	34.4	45.4	49.9	50.7	53.7
Degree	6	7	8	9	10
Rate (%)	56.1	56.3	56.5	56.3	57.4
Degree	11	**12**	13	14	15
Rate (%)	57.3	**57.8**	56.9	57.8	56.6

Table 5 presents recognition rates for several parameter sets for HOG3D descriptor (3.2). The best recognition rate was 68.9% using a grid 32x32 and a HOG3D of 128 bins.

Table 6 shows the recognition rates obtained with the proposed descriptor with a grid of 32x32 and a HOG3D of 8x16 [4]. A power normalization with $\gamma = 0.2$ was applied on the final HOG3D tensor. The best recognition rate was 72.7%, concatenating the HOG3D with polynomials of degree 13 (88410 elements). The confusion matrix for this result is presented in Table 7.

Table 5: Recognition rates of UCF11 dataset for several parameter sets for the HOG3D descriptor [4].

Parameters	**Rate (%)**
Grid 4x4 HOG3D 8x16	65.0
Grid 8x8 HOG3D 8x16	67.5
Grid 16x16 HOG2D 8x16	68.4
Grid 32x32 HOG3D 8x16	**68.9**

Table 6: Concatenating the optical flow tensor descriptor with a grid of 32x32 and a HOG3D of 8x16 [4] for UCF11 dataset. A power normalization was applied on the final HOG3D tensor with $\gamma = 0.2$.

Degree	1	2	3	4	5
Rate (%)	69.3	68.0	70.0	70.0	71.2
Degree	6	7	8	9	10
Rate (%)	70.7	71.8	71.5	71.4	72.2
Degree	11	12	**13**	14	15
Rate (%)	71.9	72.5	**72.7**	72.4	71.8

4.3. Hollywood2 Dataset

The Hollywood2 dataset [7] consists of 12 action classes: answering the phone, driving car, eating, fighting, getting out of the car, hand shaking, hugging, kissing, running, sitting down, sitting up, and standing up. We use the same evaluation protocol of the original paper [7]. The performance is evaluated by computing the average precision (AP) for each of the action classes. For the individual optical flow descriptor, the best results are

Table 7: Confusion matrix of UCF11 dataset for the best result, 72.7% using polynomials of degree 13 and a HOG3D of 8x16 with grid 32x32 [4].

	Bike	Dive	Golf	Juggle	Jump	Ride	Shoot	Spike	Swing	Tennis	WDog
Bike	72.4	0.7	1.0	0.0	0.7	7.5	0.0	1.0	3.0	1.0	12.8
Dive	0.0	89.2	3.2	0.0	0.0	1.8	0.6	1.2	0.6	0.7	2.7
Golf	0.0	4.9	86.8	3.0	0.0	0.0	0.0	0.0	2.7	1.0	1.6
Juggle	0.0	0.7	6.1	62.9	7.2	1.6	2.8	1.9	6.8	5.2	4.7
Jump	0.0	0.0	3.0	8.3	74.1	0.0	0.7	0.7	10.6	0.0	2.7
Ride	3.2	0.0	0.0	0.0	0.0	81.1	0.7	0.9	1.0	0.7	12.4
Shoot	7.0	3.4	0.0	8.6	2.2	0.0	55.4	14.5	1.5	3.8	3.6
Spike	0.0	3.0	0.0	1.7	0.0	1.0	5.5	84.9	3.0	0.0	1.0
Swing	6.8	0.8	3.9	2.5	4.5	2.3	2.0	1.1	65.9	1.0	9.2
Tennis	1.9	0.0	5.5	4.1	5.1	1.2	7.0	2.6	0.6	69.7	2.3
WDog	7.0	1.0	1.8	3.4	0.0	17.9	1.8	0.7	5.2	3.9	57.4

achieved with a Gaussian kernel. For the combination, the triangular kernel shows the best recognition rates.

Table 8 shows the recognition rates for several degrees using the optical flow tensor descriptor. Similar to UCF11 dataset, the sliding window is not interesting for this dataset because the actions are more complex than on KTH dataset. We can see that the recognition rates achieved are very low. In fact, the summation of tensors will tend to be an isotropic tensor because there are a lot of different motions happening at the same time in the scenes. The best recognition rate was only 15% with polynomials of degree 2 (300 elements).

In [4] is reported that the best result is achieved with a grid 4x4 and 128 bins obtaining 34.03% of recognition rate. Thus, we concatenate our optical flow tensor descriptor with this HOG3D to form a new global motion descriptor. Table 9 shows the recognition rates for several degrees. The best recognition rate was 40.3% concatenating the HOG3D with polynomials of degree 3 (820 elements). The average precision for each class is presented in Table 10.

Table 8: Recognition rates of Hollywood2 dataset for several degrees using only the optical flow descriptor.

Degree	1	**2**	3	4	5	6	7	8	9	10
Rate (%)	12.0	**15.0**	13.2	13.3	12.1	13.0	14.5	13.6	13.0	13.3

4.4. Comparison with the state-of-the-art

A comparison with the state-of-the-art methods is presented in Table 11.

Table 9: Concatenating the optical flow descriptor with a grid of 4x4 and a HOG3D of 8x16 [4] for Hollywood2 dataset. A power normalization was applied on the final HOG3D tensor with $\gamma = 0.2$.

Degree	1	2	**3**	4	5	6	7	8	9	10
Rate (%)	39.5	39.9	**40.3**	40.2	40.3	39.8	40.1	39.7	39.9	40.3

Table 10: Average precision (AP) for each class for the best result on Hollywood2 dataset.

Action	**AP (%)**
AnswerPhone	26.3
DriveCar	71.4
Eat	37.5
FightPerson	50.4
GetOutCar	32.0
HandShake	18.6
HugPerson	29.1
Kiss	47.9
Run	54.9
SitDown	57.5
SitUp	8.8
StandUp	48.7
Mean	40.3

Table 11: Comparison with state-of-the-art for KTH, UCF11 and Hollywood2 datasets.

KTH		**UCF11**		**Hollywood2**	
Laptev et al [11]	72.0	Perez et al [4]	68.9	Perez et al [4]	34.0
Laptev et al [2]	91.8	Wang et al [1]	84.2	Laptev et al [2]	45.2
Solmaz et al [12]	92.0			Kobayashi and Otsu [9]	47.7
Zhen and Shao [8]	92.0			Wang et al [1]	58.3
Perez et al [4]	92.0				
Wang et al [1]	94.2				
Kobayashi and Otsu [9]	95.6				
Our Method	93.2	Our Method	72.7	Our Method	40.3

The proposed method achieves a competitive accuracy with a much simpler global approach, using only the information from optical flow and histograms of gradients, without any bag-of-features strategy [1, 2, 9].

In all datasets we improved the performance of the descriptors previously proposed in [3, 4] and showed better results than other global descriptors [11, 12].

When compared to bag-of-features strategy on KTH dataset, our descriptor shows better performance than those methods which uses HOF and HOG as features [2]. The addition of more information to the descriptor, as MBH and trajectory associated with HOF and HOG [1], induces a better recognition than our descriptor. Even though, the recognition rate is very close with a much simpler approach.

For UCF11 and Hollywood2 datasets, we note that using several features plays an important role and that learning methods improve overall recognition. The performance of our descriptor is lower than these approaches [1, 2, 9] but is fairly competitive.

Thereby, we can conclude that our descriptor aggregates useful information of optical flow and HOG3D, enhancing the recognition rate. Moreover, our descriptor only depends on the video, no learning method is required. The addition of new videos and/or new action categories with our approach does not require any re-computation or changes to the previously computed descriptors.

5. Conclusion

In this paper, we presented a novel approach for motion description in videos combining optical flow and HOG3D information. It is an effective approach reaching 93.2% of recognition rate with KTH, comparable to the best local and learning-based methods. However, for the UCF11 and Hollywood2 datasets we note points of interest play an important role and that learning methods improve overall recognition. Our recognition rate is lower than the approaches based on codebook but is fairly competitive in both datasets.

The main advantage of our method is that it reaches good recognition rates depending uniquely on the video. The addition of new videos and/or new action categories with our approach does not require any re-computation or changes to the previously computed descriptors. Finally, it might be valuable in a scenario where no human action classification method solves all application demands.

The drawback of our method is that larger and complex video datasets require higher degree polynomials to give good classification results. As a consequence, the number of coefficients increases exponentially leading to high time complexity. In some cases, increasing the degree does not neces-

sarily leads to a better classification, such as in Hollywood2 dataset.

In order to improve the recognition rate of our descriptors, we intend to further analyze the spectral characteristics of the proposed orientation tensor. Furthermore, we need to study the conditions of divergence and convergence of the tensor accumulation.

6. Acknowledgements

Authors thank to FAPEMIG and CAPES for funding.

References

[1] H. Wang, A. Kläser, C. Schmid, L. Cheng-Lin, Action Recognition by Dense Trajectories, in: IEEE Conference on Computer Vision & Pattern Recognition, Colorado Springs, United States, 2011, pp. 3169–3176.

[2] I. Laptev, M. Marszałek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Computer Vision & Pattern Recognition, 2008.

[3] V. F. Mota, E. A. Perez, M. B. Vieira, L. M. Maciel, F. Precioso, 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, 2012, pp. 298–301.

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

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

[6] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos in the wild, IEEE CVPR, 2009.

[7] M. Marszałek, I. Laptev, C. Schmid, Actions in context, in: Conference on Computer Vision & Pattern Recognition, 2009.

[8] X. Zhen, L. Shao, A local descriptor based on laplacian pyramid coding for action recognition, Pattern Recognition Letters (0) (2012) –, in Press. doi:10.1016/j.patrec.2012.10.021.

[9] T. Kobayashi, N. Otsu, Motion recognition using local auto-correlation of spacetime gradients, Pattern Recognition Letters 33 (9) (2012) 1188 – 1195. doi:10.1016/j.patrec.2012.01.007.

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

[11] I. Laptev, B. Caputo, C. Schuldt, T. Lindeberg, Local velocity-adapted motion events for spatio-temporal recognition, Comput. Vis. Image Underst. 108 (2007) 207–229. doi:10.1016/j.cviu.2006.11.023.

[12] B. Solmaz, S. M. Assari, M. Shah, Classifying web videos using a global video descriptor, Machine Vision and Applications (2012) 1–13doi:10.1007/s00138-012-0449-x.

[13] M. Druon, Modélisation du mouvement par polynômes orthogonaux :

application à l'étude d'écoulements fluides, Ph.D. thesis, Université de Poitiers (02 2009).

[14] O. Kihl, B. Tremblais, B. Augereau, M. Khoudeir, Human activities discrimination with motion approximation in polynomial bases, in: IEEE International Conference on Image Processing, Hong-Kong, 2010, pp. 2469–2472.

[15] C.-F. Westin, A tensor framework for multidimensional signal processing, Ph.D. thesis, Linköping University, Sweden, S-581 83 Linköping, Sweden, dissertation No 348, ISBN 91-7871-421-4 (1994).

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

[17] A. Kläser, M. Marszałek, C. Schmid, A spatio-temporal descriptor based on 3d-gradients, in: British Machine Vision Conference, 2008, pp. 995–1004.

[18] D. G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the International Conference on Computer Vision - Volume 2, ICCV '99, IEEE Computer Society, Washington, DC, USA, 1999.

[19] B. Augereau, B. Tremblais, C. Fernandez-Maloigne, Vectorial computation of the optical flow in color image sequences., in: Thirteenth Color Imaging Conference, 2005, pp. 130–134.

[20] B. Lucas, T. Kanade, An iterative image registration technique with an

application to stereo vision (ijcai), in: Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI 81), 1981, pp. 674–679.