

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

Virginia Fernandes Mota, Eder de Almeida Perez, Silva Maciel Luiz Maurílio Da, Marcelo Bernardes Vieira, Philippe-Henri Gosselin

► **To cite this version:**

Virginia Fernandes Mota, Eder de Almeida Perez, Silva Maciel Luiz Maurílio Da, Marcelo Bernardes Vieira, Philippe-Henri Gosselin. A tensor motion descriptor based on histograms of gradients and optical flow. Pattern Recognition Letters, Elsevier, 2014, 39, pp.85-91. 10.1016/j.patrec.2013.08.008 . hal-00861395

HAL Id: hal-00861395

<https://hal.archives-ouvertes.fr/hal-00861395>

Submitted on 12 Sep 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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/ICE, 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

¹Corresponding 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 obtained 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

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

31 **2. Related work**

32 Laptev et al [2] present a combination of histograms of gradients (HOG)
33 with histogram of optical flow (HOF) to characterize local motion and shape.
34 Histograms of spatial gradient and optical flow are computed and accumu-
35 lated in space-time neighborhoods of detected interest points. Similarly to
36 the SIFT descriptor, normalized histograms are concatenated to HOG and
37 HOF vectors. Then, the signature of the video is computed through a bag-
38 of-features technique.

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

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

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

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

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

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

71 In order to obtain a global descriptor on the KTH dataset, Laptev et al
72 [11] apply the Zelnik descriptor [10] in two different ways: using multiple
73 temporal scales like the original and using multiple temporal and spatial
74 scales.

75 Solmaz et al [12] present a global descriptor based on bank of 68 Gabor
76 filters. For each video, they extract a fixed number of clips and compute the
77 3-D Discrete Fourier Transform. Applying each filter of the 3-D filter bank
78 separately to the frequency spectrum, the output is quantized in fixed sub-
79 volumes. They concatenate the outputs and perform dimension reduction
80 using PCA and classification by a SVM.

81 3. Proposed Method

82 3.1. Tensor based on optical flow approximation

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

$$F : \Omega \subset R^2 \rightarrow R^2 \\ (x_1, x_2) \mapsto (V^1(x_1, x_2), V^2(x_1, x_2))$$

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

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

91 In that way, it reduces the dimension of the optical flow field. Thus, we
92 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:

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

93 where

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

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

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

96 *3.1.1. Orientation tensor: coding frame coefficients*

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

99 Given the vector \vec{v} with m elements, it can be represented by the tensor
100 $T = \vec{v}\vec{v}^T$. It is desired that the eigenvector with the largest eigenvalue of
101 the tensor points out the dominant direction of the signal. A signal with
102 no dominant direction is represented by an isotropic tensor, i.e. the three
103 eigenvalues are approximately equal. It is important to note that the well
104 known structure tensor is a specific case of orientation tensor [16].

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

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

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

114 *3.1.2. Global tensor descriptor*

115 We have to express the motion average of consecutive frames using a series
116 of tensors. This can be achieved by $T^{OF} = \sum_a^b T_f$ using all video frames or
117 an interval of interest. By normalizing T_f with a L_2 norm, we are able to
118 compare different video clips or snapshots regardless their length or image
119 resolution.

120 If the accumulation series diverges, we obtain an isotropic tensor which
121 does not hold useful motion information. But, if the series converge as an
122 anisotropic tensor, it carries meaningful average motion information of the
123 frame sequence. The conditions of divergence and convergence need further
124 studies.

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

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

133 *3.2. Tensor based on histogram of gradients*

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

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

135 or, equivalently, in spherical coordinates $\vec{s}_t(p) = [\rho_p \ \theta_p \ \psi_p]$ with $\theta_p \in [0, \pi]$,
136 $\psi_p \in [0, 2\pi]$ and $\rho_p = \|\vec{g}_t(p)\|$, indicate brightness variation that might be

137 the result of local motion.

138 The gradient of all n points of the image I_j can be compactly represented
 139 by a tridimensional histogram of gradients $\vec{h}_j = \{h_{k,l}\}$, $k \in [1, nb_\theta]$ and
 140 $l \in [1, nb_\psi]$, where nb_θ and nb_ψ are the number of cells for θ and ψ coordinates
 141 respectively. There are several methods for computing the HOG3D and we
 142 chose, for simplicity, an uniform subdivision of the angle intervals to populate
 143 the $nb_\theta \cdot nb_\psi$ bins:

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

144 where $\{p \in I_j \mid k = 1 + \lfloor \frac{nb_\theta \cdot \theta_p}{\pi} \rfloor, l = 1 + \lfloor \frac{nb_\psi \cdot \psi_p}{2\pi} \rfloor\}$ are all points whose angles
 145 map to k and l bins, and w_p is a per pixel weighting factor which can be
 146 uniform or gaussian as in [17]. The whole gradient field is then represented
 147 by a vector \vec{h}_j with $nb_\theta \cdot nb_\psi$ elements.

148 3.2.1. Global tensor descriptor: coding HOG3D coefficients as tensors

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

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

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

154 3.2.2. Global tensor descriptor: subdividing the frame using a grid

155 When the gradient histogram is computed using the whole image, the
 156 cells are populated with vectors regardless their position in the image. This

157 implies in a loss of the correlation between the gradient vectors and their
 158 neighbors. As observed in several works [17], the subdivision of the video
 159 into cubes of frames enhances the recognition rate, using a gaussian weight
 160 for w_p .

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

$$T_j = \sum_{c,r} \vec{h}_j^{c,r} \vec{h}_j^{c,r T}, \quad (2)$$

167 which captures the uncertainty of the direction of the m -dimensional vectors
 168 $\vec{h}_f^{a,b}$ in the frame j . This tensor is normalized using the L_2 norm. The
 169 image subdivision does not change the descriptor size and the accumulation
 170 described above is the same. The global descriptor with image subdivision
 171 and histograms of gradients is then

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

172 Another improvement is to accumulate the tensor obtained with the video
 173 frame flipped horizontally. Therefore, the HOG3D is computed for each
 174 block, the final tensor is computed (Eq. 2) and simply added to the original
 175 frame tensor. This flipped version enforces horizontal gradient symmetries
 176 that occur on the video, even those between multiple frames. In our experi-
 177 ments (Sec. 4) all HOG3D descriptors are obtained using this improvement.

178 *3.3. Combining orientation tensors*

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

$$T = \{T^{OF}, T^{HOG}\}. \quad (3)$$

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

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

189 It is important to note that the nature of these two tensors is different and,
190 as such, need to be equalized. One possible way is to use a power normal-
191 ization in one of the descriptors. Experimentally, best results were obtained
192 by normalizing the HOG3D tensor: the T^{HOG} descriptor in Equation 3 has
193 all of its elements a_k adjusted to a_k^γ , $\gamma \in]0, 1]$.

194 **4. Experimental results**

195 We compute the optical flow using the method described by Augereau
196 et al [18]. This method was chosen because we found experimentally that it
197 computes a more regular optical flow than the one computed by the standard
198 Lucas-Kanade [19].

199 *4.1. KTH Dataset*

200 The KTH actions dataset [5] consists of six human action classes: walk-
201 ing, jogging, running, boxing, waving, and clapping. Each action class is
202 performed several times by 25 subjects. The sequences were recorded in four
203 different scenarios: outdoors, outdoors with scale variation, outdoors with
204 different clothes and indoors. The background is homogeneous and static in
205 most sequences. In total, the data consists of 2391 video samples. We use
206 the same evaluation protocol of the original paper [5], as [1].

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

215 In [4] is reported that the best result is achieved with a grid 8x8 and
216 128 bins obtaining 92.0% of recognition rate. Thus, we concatenate our
217 optical flow tensor descriptor with this HOG3D to form a new global motion
218 descriptor. Table 2 shows the recognition rates for several degrees. The best
219 recognition rate was 93.2% with polynomials of degree 5 (3670 elements)
220 concatenated with a HOG3D of 128 bins (8256 elements). The confusion
221 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

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	HWay	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
HWay	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	0.0	99.3

222 *4.2. UCF11 Dataset*

223 The UCF11 dataset [6] consists in 11 action categories: basketball shoot-
 224 ing, biking/cycling, diving, golf swinging, horse back riding, soccer juggling,
 225 swinging, tennis swinging, trampoline jumping, volleyball spiking, and walk-
 226 ing with a dog. We use the same evaluation protocol of the original paper
 227 [6].

228 The sliding window is not interesting for this dataset because its actions
 229 are more complex than those in KTH dataset. Table 4 shows the recognition
 230 rates for several degrees using the optical flow tensor descriptor. The best
 231 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

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

235 Table 6 shows the recognition rates obtained with the proposed descriptor
 236 with a grid of 32x32 and a HOG3D of 8x16 [4]. A power normalization with
 237 $\gamma = 0.2$ was applied on the final HOG3D tensor. The best recognition rate

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

238 was 72.7%, concatenating the HOG3D with polynomials of degree 13 (88410 elements).

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

239

240 4.3. Hollywood2 Dataset

241 The Hollywood2 dataset [7] consists of 12 action classes: answering the
 242 phone, driving car, eating, fighting, getting out of the car, hand shaking,
 243 hugging, kissing, running, sitting down, sitting up, and standing up. We

244 use the same evaluation protocol of the original paper [7]. The performance
 245 is evaluated by computing the average precision (AP) for each of the ac-
 246 tion classes. For the individual optical flow descriptor, the best results are
 247 achieved with a Gaussian kernel. For the combination, the triangular kernel
 248 shows the best recognition rates.

249 Table 7 shows the recognition rates for several degrees using the optical
 250 flow tensor descriptor. Similar to UCF11 dataset, the sliding window is not
 251 interesting for this dataset because the actions are more complex than on
 252 KTH dataset. We can see that the recognition rates achieved are very low.
 253 In fact, the summation of tensors will tend to be an isotropic tensor because
 254 there are a lot of different motions happening at the same time in the scenes.
 255 The best recognition rate was only 15% with polynomials of degree 2 (300
 256 elements).

257 In [4] is reported that the best result is achieved with a grid 4x4 and
 258 128 bins obtaining 34.03% of recognition rate. Thus, we concatenate our
 259 optical flow tensor descriptor with this HOG3D to form a new global motion
 260 descriptor. Table 8 shows the recognition rates for several degrees. The best
 261 recognition rate was 40.3% concatenating the HOG3D with polynomials of
 262 degree 3 (820 elements).

Table 7: 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

Table 8: 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

263 *4.4. Comparison with the state-of-the-art*

264 A comparison with the state-of-the-art methods is presented in Table 9.

Table 9: 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

265 The proposed method achieves a competitive accuracy with a much sim-
 266 pler global approach, using only the information from optical flow and his-
 267 tograms of gradients, without any bag-of-features strategy [1, 2, 9].

268 In all datasets we improved the performance of the descriptors previously
 269 proposed in [3, 4] and showed better results than other global descriptors
 270 [11, 12].

271 When compared to bag-of-features strategy on KTH dataset, our descrip-
 272 tor shows better performance than those methods which uses HOF and HOG

273 as features [2]. The addition of more information to the descriptor, as MBH
274 and trajectory associated with HOF and HOG [1], induces a better recog-
275 nition than our descriptor. Even though, the recognition rate is very close
276 with a much simpler approach.

277 For UCF11 and Hollywood2 datasets, we note that using several features
278 plays an important role and that learning methods improve overall recog-
279 nition. The performance of our descriptor is lower than these approaches
280 [1, 2, 9] but is fairly competitive.

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

287 **5. Conclusion**

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

295 The main advantage of our method is that it reaches good recognition
296 rates depending uniquely on the video. The addition of new videos and/or

297 new action categories with our approach does not require any re-computation
298 or changes to the previously computed descriptors. Finally, it might be
299 valuable in a scenario where no human action classification method solves all
300 application demands.

301 The drawback of our method is that larger and complex video datasets
302 require higher degree polynomials to give good classification results. As a
303 consequence, the number of coefficients increases exponentially leading to
304 high time complexity. In some cases, increasing the degree does not neces-
305 sarily leads to a better classification, such as in Hollywood2 dataset.

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

310 **6. Acknowledgements**

311 Authors thank to FAPEMIG and CAPES for funding.

312 **References**

- 313 [1] H. Wang, A. Kläser, C. Schmid, L. Cheng-Lin, Action Recognition by
314 Dense Trajectories, in: IEEE Conference on Computer Vision & Pattern
315 Recognition, Colorado Springs, United States, 2011, pp. 3169–3176.
- 316 [2] I. Laptev, M. Marszałek, C. Schmid, B. Rozenfeld, Learning realistic hu-
317 man actions from movies, in: Computer Vision & Pattern Recognition,
318 2008.

- 319 [3] V. F. Mota, E. A. Perez, M. B. Vieira, L. M. Maciel, F. Precioso, P.-H.
320 Gosselin, A tensor based on optical flow for global description of motion
321 in videos, in: SIBGRAPI 2012 (XXV Conference on Graphics, Patterns
322 and Images), Ouro Preto, MG, Brazil, 2012, pp. 298–301.
- 323 [4] E. A. Perez, V. F. Mota, L. M. Maciel, D. Sad, M. B. Vieira, Combining
324 gradient histograms using orientation tensors for human action recog-
325 nition, in: International Conference on Pattern Recognition, 2012, pp.
326 3460–3463.
- 327 [5] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: A local
328 svm approach, in: In Proc. ICPR, 2004, pp. 32–36.
- 329 [6] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos in the
330 wild, IEEE CVPR, 2009.
- 331 [7] M. Marszałek, I. Laptev, C. Schmid, Actions in context, in: Conference
332 on Computer Vision & Pattern Recognition, 2009.
- 333 [8] X. Zhen, L. Shao, A local descriptor based on laplacian pyramid coding
334 for action recognition, Pattern Recognition Letters (0) (2012) –, in Press.
335 doi:10.1016/j.patrec.2012.10.021.
- 336 [9] T. Kobayashi, N. Otsu, Motion recognition using local auto-correlation
337 of spacetime gradients, Pattern Recognition Letters 33 (9) (2012) 1188
338 – 1195. doi:10.1016/j.patrec.2012.01.007.
- 339 [10] L. Zelnik-manor, M. Irani, Event-based analysis of video, in: In Proc.
340 CVPR, 2001, pp. 123–130.

- 341 [11] I. Laptev, B. Caputo, C. Schuldt, T. Lindeberg, Local velocity-adapted
342 motion events for spatio-temporal recognition, *Comput. Vis. Image Un-*
343 *derst.* 108 (2007) 207–229. doi:10.1016/j.cviu.2006.11.023.
- 344 [12] B. Solmaz, S. M. Assari, M. Shah, Classifying web videos using a
345 global video descriptor, *Machine Vision and Applications* (2012) 1–
346 13doi:10.1007/s00138-012-0449-x.
- 347 [13] M. Druon, Modélisation du mouvement par polynômes orthogonaux :
348 application à l'étude d'écoulements fluides, Ph.D. thesis, Université de
349 Poitiers (02 2009).
- 350 [14] O. Kihl, B. Tremblais, B. Augereau, M. Khoudair, Human activities dis-
351 crimination with motion approximation in polynomial bases, in: *IEEE*
352 *International Conference on Image Processing*, Hong-Kong, 2010, pp.
353 2469–2472.
- 354 [15] C.-F. Westin, A tensor framework for multidimensional signal process-
355 ing, Ph.D. thesis, Linköping University, Sweden, S-581 83 Linköping,
356 Sweden, dissertation No 348, ISBN 91-7871-421-4 (1994).
- 357 [16] B. Johansson, G. Farnebeck, G. F. Ack, A theoretical comparison of
358 different orientation tensors, in: *Symposium on Image Analysis*, SSAB,
359 2002, pp. 69–73.
- 360 [17] D. G. Lowe, Object recognition from local scale-invariant features, in:
361 *Proceedings of the International Conference on Computer Vision - Vol-*
362 *ume 2, ICCV '99*, IEEE Computer Society, Washington, DC, USA, 1999.

- 363 [18] B. Augereau, B. Tremblais, C. Fernandez-Maloigne, Vectorial computa-
364 tion of the optical flow in color image sequences., in: Thirteenth Color
365 Imaging Conference, 2005, pp. 130–134.
- 366 [19] B. Lucas, T. Kanade, An iterative image registration technique with an
367 application to stereo vision (ijcai), in: Proceedings of the 7th Interna-
368 tional Joint Conference on Artificial Intelligence (IJCAI 81), 1981, pp.
369 674–679.