A tensor based on optical flow for global description of motion in videos

Mota V. F.*, Perez E. A.*, Vieira M. B.*, Maciel L. M.*, Precioso, F.† and Gosselin, P. H. ‡
* DCC/ICE, Universidade Federal de Juiz de Fora, Juiz de Fora, Brazil
Email: {virginia.fernandes, eder.perez, marcelo.bernardes, luiz.maurilio}@ice.ufjf.br
† Nice Sophia Antipolis University,Nice, France
Email: precioso@polytech.unice.fr
‡ ETIS , ENSEA, Cergy, France
Email: gosselin@ensea.fr

***Abstract*—Motion is one of the main characteristics that describe the semantic information of videos. In this work, a global video descriptor based on orientation tensors is proposed. This descriptor is obtained by combining polynomial coefficients calculated for each image in a video. The coefficients are found through the projection of the optical flow on Legendre polynomials, reducing the dimension of per frame motion estimations. The sequence of coefficients are then combined using orientation tensors. The global tensor descriptor created is evaluated by a classification of the KTH video database with a SVM classifier.**

***Keywords*-Global descriptor, Orientation tensor, SVM, Optical Flow, Motion modeling.**

I. Introduction

Human action recognition in videos has several applications. Most works approach this problem by a motion analysis and a representation step. Starting from the interesting work of modeling the apparent motion by a projection onto a polynomial basis [1], we propose a new global motion descriptor, based on orientation tensors, for the problem of action recognition in videos. Tensors have been largely used to compute optical flow, but its use is still growing on motion analysis and there are few works that use tensor as a descriptor for human action recognition [2], [3], [4].

The main contribution of this work is the introduction of a video motion indexing scheme based on the modeling of optic flow in video stream. Thus, using statistics on the model hyperparameters, we are able to build a signature (or feature vector) whatever is the width of frames duration of the video. More specifically, we use Legendre polynomial coefficients (which only depends on the chosen degree) to code a per frame optical flow in a tensor that is then accumulated in time defining a global descriptor for video motion.

A. Related work

Through the use of spectral features of a tensor, Jia *et al* [3] proposes a method for action recognition based on multiresolution features. A series of silhouettes are transformed into an image called Serial-Frame from which are extracted features for the construction of an eigenvalues and eigenvector space called Serials-Frame Tensor. Analyzing this space, they can separate useful information to recognize different types of action. The video dataset created by authors has 4 different types of actions: running with a tool lifted, walking, running while hitting with bare hand and walking while waving hand.

Also using the idea of silhouettes, Khadem and Rajam [4] create a set of silhouettes to form a third order tensor comprising three modes: pixels, actions and people. Then they apply a single value decomposition which allows the descriptor to be used to compare the videos of Weizmann database. They argue that this approach is better than the classic principal component analysis.

We can see that these tensor descriptors are all local ones. In general, local descriptors for human action recognition are more explored and achieve greater recognition rates. Hence, there are few references about global descriptors for human action recognition.

Zelnik *et al* presents a global descriptor based on histogram of gradients [5]. This descriptor is applied on the Weizmann video database and is obtained extracting 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 apply a global descriptor on the KTH database, Laptev *et al* [6] apply the Zelnik descriptor [5] in two different ways: using multiple temporal scales as the original and using multiple temporal and spatial scales.

A more recent approach is presented in Kihl *et al* [2] and is based on vector fields comparison. The vector fields are obtained through the projection of optical flow on an orthogonal polynomial basis of a given degree which gives a polynomial approximation of it [1]. From the similarity measure, they can retrieve a vector field within video sequences. This similarity measure is based on the covariance of the highest eigenvalues of a tensor created from the vector field. This descriptor is tested on videos of motions and can be used for interactive user interfaces. Our work is based on the same idea of polynomial approximation, however our work is not related to the descriptor created by them. They propose a per frame motion descriptor. We propose a tensor

descriptor for the whole video or any sequence of frames. Another difference is that they propose a similarity that is based in a tensor decomposition. In our approach, the tensor itself is the descriptor and they are compared by a simple L2 norm.

II. Proposed Method

A. Optical flow modeling using polynomials

The basic idea of a polynomial based model is to approximate a vector field with a linear combination of orthogonal polynomials [1], [2]. Let us define F an optical flow:

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

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:

$$\left\{ \begin{array}{l} \tilde{V}^1(x_1, x_2) = \sum_{i=0}^{g} \sum_{j=0}^{g-1} \tilde{v}^1_{i,j} P_{i,j} \\ \tilde{V}^2(x_1, x_2) = \sum_{i=0}^{g} \sum_{j=0}^{g-1} \tilde{v}^2_{i,j} P_{i,j} \end{array} \right. \quad (2)$$

where

$$\left\{ \begin{array}{l} \tilde{v}^1_{i,j} = \int \int_{\Omega} V^1(x_1, x_2) P_{i,j} \omega(x_1, x_2) dx_1 dx_2 \\ \tilde{v}^2_{i,j} = \int \int_{\Omega} V^2(x_1, x_2) P_{i,j} \omega(x_1, x_2) dx_1 dx_2 \end{array} \right. \quad (3)$$

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

$$n_g = \frac{(g+1)(g+2)}{2} \quad (4)$$

B. Orientation tensor: coding frame coefficients

An orientation tensor is a representation of local orientation which takes the form of an $N \times N$ real symmetric matrix for N-dimensional signals [7].

Given the vector $\vec{v}$ with N 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, which is a sphere on three dimensions. It is important to note that the well known structure tensor is a specific case of orientation tensor [8].

An individual tensor for each frame is created using the coefficients $\tilde{v}^1_{i,j}$ and $\tilde{v}^2_{i,j}$ (Eq. 3) calculated through the polynomial approximation of optical flow.

From the optical flow approximation, we create a vector $\tilde{v_f}$ for each frame f of the video:

$$\tilde{v}_f = [\tilde{v}^1_{0,0}, ..., \tilde{v}^1_{g,0}, \tilde{v}^2_{0,0}, ..., \tilde{v}^2_{g,0}] \in R^m \quad (5)$$

Using the vector $\tilde{v_f}$, we create an orientation tensor $T_f = \tilde{v_f}\tilde{v_f}^T$ for each frame f of the video, which is a matrix $2n_g \times 2n_g$. 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.

C. Global tensor descriptor: series of frame tensors

The main idea of our work is to present a new global motion descriptor. We have to express the motion average of consecutive frames using a series of tensors. This can be achieved by $\sum_a^b T_f$ using all video frames or an interval of interest. By normalizing T_f with a $L2$ 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. In our experiments (Sec. III), this basic global descriptor is called ND.

D. Global tensor descriptor: adding the variation of the polynomial coefficients

It is also possible to add more information to the global descriptor. In order to capture the motion variation in time, we can use both the polynomial coefficients (Eq. 3) and an approximation of their first temporal derivative:

$$\partial_t \tilde{v}_f = \left[\frac{\tilde{v}^1_{i,j}(f) - \tilde{v}^1_{i,j}(f-1)}{\Delta t}, \frac{\tilde{v}^2_{i,j}(f) - \tilde{v}^2_{i,j}(f-1)}{\Delta t} \right]_{i+j<g} \quad (6)$$

Hence, the new coefficient vector is $\tilde{v_f}^{new} = [\tilde{v_f}, \partial_t \tilde{v}_f]$. The orientation tensor T_f for each frame is computed as above, resulting in a matrix of $4n_g \times 4n_g$. The global motion descriptor obtained by accumulation and normalization now captures the rate of motion changes in time.

It is important to note that the accumulated tensor created is symmetric, so 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.

In our experiments (Sec. III), this descriptor enhanced with temporal derivative of coefficients is called WD.

III. Experimental results

Validation set.: To validate our tensor descriptor, we use the KTH video dataset [9]. Although it is a base with relatively simple actions, there are so few studies suggesting global descriptors that it is still a good reference for comparison.

Experimental protocol: . The optical flow is computed by a method described in [10]. 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 [10]. We run a multiclass classifier using a one-against-all strategy and a Bayes criterion for model selection. Each class is modeled using a SVM classifier with a triangular kernel function with Euclidian distance.

Results.: In order to study the performance of our descriptor, we evaluate the two types of descriptors as described in the previous section: basic tensor (ND) (Sec. II-C) and with derivative information (WD) (Sec. II-D). 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 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. Both descriptors can be computed inside this window, thus we have two more variations: basic tensor using window (ND-W), and tensor with derivative using window (WD-W).

Degree	1	4	8	17	22
Rate	70.02%	77.31%	78.24%	79.74%	**81.13%**

TABLE I
RECOGNITION RATES FOR SEVERAL DEGREES FOR ND DESCRIPTOR

	Walk	Jog	Run	Box	HClap	HWav
Walk	86.81	21.53	4.86	0.00	0.00	0.00
Jog	10.42	54.17	11.11	0.00	0.70	0.00
Run	2.78	24.31	84.03	3.5	0.00	0.00
Box	0.00	0.00	0.00	96.50	9.03	4.86
HClap	0.00	0.00	0.00	0.00	88.19	18.05
HWav	0.00	0.00	0.00	0.00	2.08	77.08

TABLE II
CONFUSION MATRIX FOR THE ND DESCRIPTOR AND BASIS DEGREE OF 22: FINAL TENSOR DESCRIPTOR WITH 609960 ELEMENTS.

Table I shows the recognition rates of the basic descriptor using several polynomial degrees (Sec. II-C). Note that higher degree is, better is the recognition rate. The best recognition rate was 81.13% with polynomial degree of 22. The resulting confusion matrix is shown on Table II. As expected, the worst recognition rate is found for jogging. This can be explained by high similarities between the optical flow of this class and the ones of walking and running.

Degree	1	4	8	17	22
Rate	79.16%	**84.83%**	83.45%	83.21%	82.86%

TABLE III
RECOGNITION RATES FOR SEVERAL DEGREES FOR WD DESCRIPTOR

	Walk	Jog	Run	Box	HClap	HWav
Walk	90.28	22.22	2.08	0.00	0.00	0.00
Jog	9.03	61.81	8.33	0.00	0.00	0.00
Run	0.70	15.97	89.58	1.40	0.00	0.00
Box	0.00	0.00	0.00	98.60	9.72	5.56
HClap	0.00	0.00	0.00	0.00	89.58	15.28
HWav	0.00	0.00	0.00	0.00	0.69	79.17

TABLE IV
CONFUSION MATRIX FOR THE WD DESCRIPTOR AND BASIS DEGREE OF 4: FINAL TENSOR DESCRIPTOR WITH 1830 ELEMENTS.

Table III shows the recognition rates for the descriptor coding the derivative of polynomial coefficients. The best recognition was 84.83% with a basis of degree 4. The confusion matrix is shown on Table IV. We can see that frame coherence reduces significantly the mislabelling of jogging, walking and running actions. The speed of the motion (captured by the added derivative) is the main difference between them. The high recognition rate with low polynomial degree is remarkable. With degree 4, the final descriptor for (WD) has only 1830 elements. This is much better compared to the best result for (ND) with 609960 elements for degree 22 (Table II).

Degree	Recognition rate ND-W	Recognition rate WD-W
1	70.00%	76.60%
8	76.60%	82.39%
12	81.01%	84.94%
17	83.79%	**86.44%**
23	**84.71%**	85.75%

TABLE V
RECOGNITION RATES FOR SEVERAL DEGREES USING A SLIDING WINDOW WITH DIMENSIONS 60X100.

	Walk	Jog	Run	Box	HClap	HWav
Walk	93.75	11.81	4.86	7.69	0.70	0.00
Jog	2.78	74.31	6.94	0.00	0.00	0.00
Run	3.47	13.89	88.19	2.10	0.00	0.00
Box	0.00	0.00	0.00	90.21	2.78	5.56
HClap	0.00	0.00	0.00	0.00	90.28	22.92
HWav	0.00	0.00	0.00	0.00	6.25	71.53

TABLE VI
CONFUSION MATRIX FOR THE ND-W DESCRIPTOR AND BASIS DEGREE OF 23: FINAL TENSOR DESCRIPTOR WITH 180300 ELEMENTS.

	Walk	Jog	Run	Box	HClap	HWav
Walk	95.83	11.81	4.17	6.29	0.00	0.00
Jog	3.47	76.39	6.25	0.70	0.70	0.00
Run	0.70	11.81	89.58	2.10	0.00	0.00
Box	0.00	0.00	0.00	90.91	5.56	4.86
HClap	0.00	0.00	0.00	0.00	93.75	22.92
HWav	0.00	0.00	0.00	0.00	0.00	72.22

TABLE VII
CONFUSION MATRIX FOR THE WD-W DESCRIPTOR AND BASIS DEGREE OF 17: FINAL TENSOR DESCRIPTOR WITH 234270 ELEMENTS.

Table V shows the results for both descriptors using the sliding window. We have tested several sizes of window and the best for KTH database is 60x100 pixels. This size can capture the most representative motions of the scenes.

We can see that the best result was found for the descriptor with derivative and sliding window, which has a recognition rate of 86.44% with a basis degree of 17, having 234270 elements. Its confusion matrix is shown in Table VII. A similar performance of 84.83% is obtained by the descriptor (WD) (Sec. II-D) without using the sliding window on the degree 4 with 1830 coefficients. The large amount of coefficients is not advantageous in terms of time processing. The first result took about 10 minutes and the former 89 minutes in a platform based on Intel Core i7-870 2930Mhz with 8Gb of 1333Mhz/DDR3. An overview of recognition rates for all descriptor variants (ND, WD, ND-W, WD-W) in function of the polynomial degree is shown in Figure 1. Table VI shows the confusion matrix for the descriptor without derivatives and with sliding window, which has a recognition rate of 84.71% with a basis of degree 23.

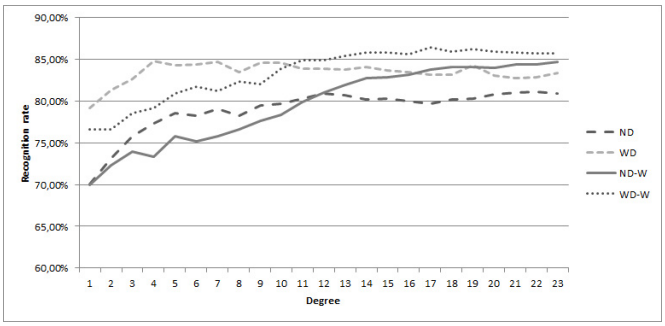

Fig. 1. Overview of recognition rates for all four types of descriptors.

Finally, the descriptor proposed does not outperform the best recognition rates with local information found for KTH database as shown in Table VIII ([11], [12]). However, its performance is close to the state-of-the-art methods, without tracking salient objects. Note that it outperforms the best global descriptor we have found in the literature [6].

[6]	**ND**	**ND-W**	**WD**	**WD-W**	[11]	[12]
72%	81.13%	84.71%	84.83%	86.44%	94.53%	95.33%

TABLE VIII
COMPARISON BETWEEN OUR GLOBAL DESCRIPTORS AND OTHER STATE-OF-THE-ART APPROACHES.

IV. CONCLUSION

In this work, we proposed a global motion tensor descriptor using a polynomial representation of the optical flow. We use Legendre polynomial coefficients to code a per frame optical flow in a tensor that is then accumulated in time. Although our descriptor does not reach the recognition rate found by local descriptors on KTH, we argue that global descriptors can achieve a good balance between descriptor size, recognition rate and time complexity, and should be more investigated.

Our method beats with a 86.83% recognition rate the global descriptor found in literature [6], that has almost 72% of rate based on histogram of gradients, and is a promising motion representation that can be further improved.

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

ACKNOWLEDGMENT

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

REFERENCES

[1] M. Druon, "Modélisation du mouvement par polynômes orthogonaux : application à l'étude d'écoulements fluides," Ph.D. dissertation, Université de Poitiers, 02 2009.

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

[3] C. Jia, S. Wang, X. Xu, C. Zhou, and L. Zhang, "Tensor analysis and multi-scale features based multi-view human action recognition," in *International Conference on Computer Engineering and Technology*, ser. ICCET'10. IEEE Press, 2010.

[4] B. Khadem and D. Rajan, "Appearance-based action recognition in the tensor framework," in *international conference on Computational intelligence in robotics and automation*, ser. CIRA'09. IEEE Press, 2009, pp. 398–403.

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

[6] I. Laptev, B. Caputo, C. Schuldt, and T. Lindeberg, "Local velocity-adapted motion events for spatio-temporal recognition," *Comput. Vis. Image Underst.*, vol. 108, pp. 207–229, December 2007.

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

[8] 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.

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

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

[11] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *CVPR*, 2010.

[12] T. Kim, S. Wong, and R. Cipolla, "R.: Tensor canonical correlation analysis for action classification," in *In: CVPR 2007*, 2007.