Tensor Clustering for Human Action Recognition

Virgínia Fernandes Mota*†, Gabriel Dutra Dias†, Wisney Tadeu dos Santos†,
Marcelo Bernardes Vieira‡ and Arnaldo de Albuquerque Araújo*
*NPDI/DCC, UFMG, Belo Horizonte, Brazil
†COLTEC, UFMG, Belo Horizonte, Brazil
‡GCG/DCC, UFJF, Juiz de Fora, Brazil

Abstract—**In this work, we present a new technique called Bag-of-tensors. This research aims to create a new method for video description based on tensors which takes into account the nature of the tensor and its anisotropic properties. Therefore, the Bag-of-tensors is composed by three main steps: Feature extraction with tensor creation, coding based on tensor clustering and aggregated pooling. For the task of human action recognition, we used the KTH dataset. Our experiments show that this technique is promising and interesting to be further explored.**

***Keywords*-orientation tensor, tensor clustering, bag-of-features, human action recognition**

I. Introduction

Human action recognition is a research field with application in several areas such as video indexing, surveillance, human-computer interfaces, among others. Several works in literature tackle this problem by extracting a set of descriptors and comparing them throughout a similarity measure. In the past years, several descriptors have been proposed.

The use of tensors is gaining space in image and video classification. An interesting set of works about tensor descriptors for human action recognition is presented in [1]–[6]. All those tensor-based descriptors are obtained by an aggregation of all tensors extracted on each frame.

The drawback of those tensor-based methods is that aggregation of several tensors could lead to an isotropic tensor, which does not have any information of direction. Our hypothesis is that the combination of visual dictionary based methods with orientation tensors in all steps could provide better motion descriptors. Our proposal is, then, to present a new method for video description based on tensors which takes into account the nature of the tensor and its anisotropic properties.

Contributions: The main contribution of this work is to present a new method for video description based on tensors which takes into account the anisotropic properties of tensors. The core of this method is a new tensor clustering technique to create a visual dictionary. The tensor clustering associated with an aggregated pooling leads to the new method called Bag-of-tensors.

A. Related work

Human action recognition can be regarded as the combination of three main steps: (i) feature extraction; (ii) descriptor creation and (iii) action classification. Therefore, techniques for human action recognition can be categorized by various criteria. We could classify the existing methods based on the type of features. Thus, the methods are categorized into the following classes: Global appearance and motion descriptors and Bag-of-features representations.

The majority of tensor approaches can be categorized in global appearance and motion descriptors. An interesting set of works about tensor descriptors for human action recognition is presented in [1]–[6]. Those works explored five different approaches to use tensor descriptors: accumulating optical flow projections [1]; accumulating histogram of gradients [2]; combining optical flow and histogram of gradients tensors [5]; combining multiple histogram of gradients [3]; combining multiple gradient estimators [4]; and accumulating tensors based on variable size block matching algorithm [6].

In all cases, descriptors are obtained by an aggregation of all tensors extracted on each frame. Moreover, all works used tensors as vectors in euclidean space, not relying on its anisotropic properties. However, even with these simple descriptors, they were able to be competitive with state-of-the-art methods achieving up to 93% of recognition rate on KTH dataset [5].

The majority of state-of-the-art methods uses bag-of-features representations with combination of several features, such as trajectories, histogram of gradients, histogram of optical flow and motion boundary histogram, achieving up to 95.3% of recognition rate on KTH dataset [7].

Those works raised us a question: Could the combination of bag-of-features based methods with orientation tensors provide better motion descriptors for human action recognition? That is the question we are aiming to answer with this ongoing research.

The rest of the paper is organized as follows. In Section II, we present the technical background needed to a better understanding of our method. In Section III, we provide a detailed description of our approach. Finally, in Section IV, we carry out experiments on a benchmark action dataset.

II. Technical background

A. Orientation tensor

An orientation tensor is a representation of local orientation which takes the form of an m x 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 [8].

Geometric interpretation of this tensor is very attractive to motion description. Given the set of eigenvalues of a tensor, it can have a spear, a plate or a ball component being dominant, which indicates the dominant direction of the tensor. Thus, for motion description, we are interested in anisotropic tensors, which has information of direction.

B. Histogram of gradients

The partial derivatives, or gradient, obtained by the filtering of the j-th video frame at pixel p is defined as the vector:

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

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 respectively the number of cells for θ and ψ coordinates. We chose a uniform subdivision of the angle intervals to populate $nb_\theta \cdot nb_\psi$ bins (Eq. 1), since it achieves good results and it is fast to compute.

$$h_{k,l} = \sum_p \rho_p \cdot w_p, \tag{1}$$

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 [9]. The whole gradient field is then represented by a vector $\vec{h}_j$ with $nb_\theta \cdot nb_\psi$ elements.

C. Bag-of-features

The bag-of-features [10] method (BoF), or bag-of-visual-features, is a visual analog to the traditional Bag-of-Words (BoW) representations for text retrieval [11]. The main idea is to represent a histogram of word occurrences in order to provide a compact representation for the text.

Thus, for a bag-of-visual-features technique we consider that an image/video is composed of *visual words*. A visual word is a local segment in an image, defined either by a region (image patch or blob) either by a reference point within its neighborhood. The analysis of visual words occurrences and configurations allows us to detect frequent occurrence patterns.

The standard process to create a feature vector with a BoF based approach follows three steps: (*i*) low-level local descriptor extraction, (*ii*) coding, which performs a pointwise transformation of the descriptors into a representation better adapted to the task (codebook) and (*iii*) pooling, which summarizes the coded features over larger neighborhoods. Classification algorithms are then trained on the final vectors obtained.

III. Proposed Method

In this Section, we present our new technique called Bag-of-tensors composed by three steps: Feature extraction with tensor creation, coding and pooling.

A. Feature extraction with tensor creation

The feature extraction step obtain gradients for each frame. The histogram of gradients (HOG) with m bins $\vec{h}_j$, computed for j-th frames, can, then, be coded in a tensor as follows:

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

Other features can be used in this process. Since it is an ongoing research, we also intend to test other features. One may note that different from tensors descriptors presented in Section IA, tensor accumulation will only appear on the pooling step.

B. Coding

The coding step is an essential part of BoF pipeline as it computes the codebook. The majority of BoF techniques uses the k-means cluster algorithm.

In order to take into account the nature of the tensor and its anisotropic properties, coding step cannot be performed with regular k-means method. It is necessary a tensor clustering step based on eigenvectors and eigenvalues, therefore, orientation tensor can be clustered with its form and energy. To the best of our knowledge, there are no works dealing with tensor clustering of tensor features applied to human action recognition.

In our work, we developed a new tensor clustering composed by two levels: an eigenvector based level and an eigenvalue based level.

Given a set of orientation tensors $\{T_1, T_2, ..., T_n\}$, where each tensor is created from a HOG3D, our tensor clustering is a modified k-means clustering aiming to partition the n tensors into $k(\leq\ n)$ sets $S\ =\ \{S_1, S_2, ..., S_k\}$ so as to minimize the distance d based on eigenvector, i. e., tensors are being clustered by its form. In other words, its objective is to minimize:

$$d(T_1, T_2) = ||I - E_{T1} E_{T2}|| \tag{2}$$

where I is the identity matrix and E_T is the set of eigenvectors of a tensor T.

This first clustering gives sets of tensors grouped by its form. We need now to cluster by its energy, i. e., using the eigenvalues. Therefore, each $m-$order tensor is represented by $[\lambda_1 - \lambda_2, \lambda_2 - \lambda_3, ..., \lambda_n - \lambda_{n+1}, ..., \lambda_m]$ and a k-means is performed in each set S_i from the first level of clustering partitioning in new sets $S' = \{S_{1,1}, S_{1,2}, ..., S_{1,k_2}, ..., S_{k_1,k_2}\}$.

Thus, our codebook is created by a two-level tensor clustering which takes into account eigenvectors and eigenvalues of a set of orientation tensors. The first level has k_1 clusters and the second level cluster each k_1 into k_2 clusters creating $k_1 * k_2$ tensor codewords.

C. Pooling

The pooling step counts the occurrences for every visual word in every video to form the histograms which will constitute the BoF representation for those videos. Standard bag-of-features method loose information on the pooling step when it describes all codeword cluster information with a single value.

Inspired by the work of [12], we could change the pooling step by aggregating tensors based on its distance to the central codeword. We argue that with a codebook created using anisotropic properties we could add more information to the descriptor and improve recognition. The main idea is that after the two-level tensor clustering, we aggregate tensors based on its distance to the central codeword. Therefore, instead of being represented with a histogram of codeword occurrences, the video will be represented by an orientation tensor that represents the tendency of motion of each codeword. Thus, the video descriptor $\vec{v}_d$ is represented by

$$\vec{v}_d = [T_{S_{1,1}}, ..., T_{S_{k_1,k_2}}] \qquad (3)$$

where $T_{S_{i,j}}$ is the accumulation of all tensors from cluster $S_{i,j}$ and $\vec{v}_d$ is the concatenation of $k_1 * k_2$ tensors.

IV. Experimental results

In our experiments, we used the benchmark video dataset KTH [13]. This consists of six human action classes. 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 database consists of 2391 video samples (Figure 1).

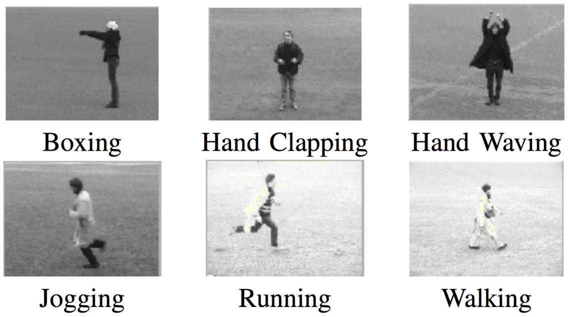

Fig. 1. Examples of videos from KTH dataset [13]

We evaluated our descriptor in a classification task and followed the same evaluation protocol proposed by the authors of the dataset with an SVM classifier.

For each experiment, we vary the percentage of tensors from each video (10, 30, 50) used to create the codebook. All tensors are randomly selected. We have a fixed number of bins of the HOG3D (2x4). In order to have a more local motion information, the HOG3D is extracted from a 8x8 subdivision of each video frame [2]. We vary k_1 and k_2 to analyze the impact of more codewords $(\{4, 2\}, \{8, 4\}, \{16, 8\})$. We compare the standard pooling with our tensor pooling.

Experiments with standard k-means on our tensors did not achieve good results, reaching a maximum of 30% of recognition rate even with the aggregated pooling. This showed that tensor clustering really needs to rely on anisotropic properties. Therefore, all experimental results in this section was performed with our two-level tensor clustering method. Our results for the HOG3D 2x4 are depicted in Tables I, II, III and IV.

TABLE I
Recognition rates (%) for KTH dataset using a HOG3D 2x4 with 8x8 grid and $k_1 = 4$ and $k_2 = 2$, i. e., 8 tensor codewords.

Percentage	**Iterations**	**Standard pooling**	**Tensor pooling**
10	10	34.9	50.0
	20	34.6	48.4
	30	35.9	53.4
30	10	36.8	50.5
	20	35.2	52.9
	30	34.8	50.6
50	10	35.8	53.5
	20	35.6	51.3
	30	34.3	51.7

TABLE II
Recognition rates (%) for KTH dataset using a HOG3D 2x4 with 8x8 grid and $k_1 = 8$ and $k_2 = 4$, i. e., 32 tensor codewords.

Percentage	**Iterations**	**Standard pooling**	**Tensor pooling**
10	10	42.2	60.1
	20	41.5	61.7
	30	45.1	61.8
30	10	44.4	59.8
	20	42.7	60.7
	30	43.7	59.3
50	10	43.9	61.8
	20	45.5	59.2
	30	44.5	58.5

TABLE III
Recognition rates (%) for KTH dataset using a HOG3D 2x4 with 8x8 grid and $k_1 = 16$ and $k_2 = 8$, i. e., 128 tensor codewords.

Percentage	**Iterations**	**Standard pooling**	**Tensor pooling**
10	10	51.4	61.1
	20	50.2	62.6
	30	51.3	63.7
30	10	48.9	62.2
	20	51.3	59.7
	30	52.5	61.5

TABLE IV
Recognition rates (%) for KTH dataset using a HOG3D 2x4 with 8x8 grid and $k_1 = 32$ and $k_2 = 16$, i. e., 512 tensor codewords.

Percentage	**Iterations**	**Standard pooling**	**Tensor pooling**
10	10	50.9	62.8
	20	54.0	64.2
	30	51.6	62.3
30	10	53.5	64.2
	20	54.1	63.6
	30	53.7	64.3

We can note that in all cases, the aggregated pooling is better than the standard pooling, which shows that our tensor pooling carries more information to describe the video.

From the number of iterations, we can see that our tensor clustering converges fast reaching good results with only 10 iterations. We believe that with more iterations we could improve the results. However, it is necessary to study the impact of computational time.

It is interesting to note that the percentage of tensors extracted has little impact on recognition rate. In some cases, using only 10% of tensors, we could achieve the same result as using 30% of tensors. In fact, we note again that tensors carry more information and we do not need a big amount of tensors to achieve good performance. Moreover, we have to highlight that accumulation of many tensors could lead to an isotropic tensor, thus, a strategy that uses less tensors could have a better convergence. This is a very important conclusion since all previous tensor descriptors from literature use all tensors of a video, which can be very time-consuming.

As the number of tensor codewords increase, the better is the result. That is expected as more similar tensors are accumulated if we have more clusters.

These first experimental results lead us to a good insight: a vocabulary based method that relies on the properties of a tensor can improve recognition rate for human action recognition. Therefore, this ongoing research has room for improvement and seems to be walking on the right path.

V. Conclusion

In this work we presented an ongoing research which created a new method called Bag-of-tensors. This method, based on tensor clustering, aims to create a video descriptor taking into account the nature of orientation tensors and its anisotropies properties. It is important to note that, tensor properties are often neglected by literature. In general, works that use tensors for video description are not interested in discuss its nature, using tensors as vectors in euclidean space. However, the tensor space must be understood and used regarding its anisotropic properties.

Our experimental results raised some important insights: Tensor clustering cannot be treated in a naive way, the tensor must be understood and used regarding its anisotropic properties; and, tensor aggregation adds more information to video descriptor than histograms. Therefore, we believe that this is a promising approach and can be further compared to state-of-the-art methods. For now, we are far below from state-of-the-art results, comparing our 64.2% with more than 93% of recognition rate. However, we are using only HOG3D with 2x4 bins extracted from only 10% of tensors from a video dataset and only 512 codewords. We believe that as we augment the number of bins and the number of codewords, we could improve our results in order to be competitive.

Therefore, to fulfill the gap between our initial results and the state-of-the-art, we intend to explore other features that can be coded into orientation tensors, such as, histogram of optical flows and dense trajectories. Moreover, we intend to explore other aggregation techniques on pooling step. Finally, the combination of features is also possible with our technique, indeed, we could analyze in which step the combination of features leads to better results.

Acknowledgment

Authors would like to thank CAPES, CNPq, FAPEMIG, UFJF and UFMG for funding.

References

[1] 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. [Online]. Available: http://www.decom.ufop.br/sibgrapi2012/http://www.decom.ufop.br/sibgrapi2012/

[2] 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.

[3] V. Mota, J. Souza, A. de Albuquerque Araújo, and M. B. Vieira, "Combining orientation tensors for human action recognition," in *SIBGRAPI 2013 - Technical Papers ()*, Arequipa, Peru, aug 2013.

[4] D. Sad, V. Mota, L. Maciel, M. B. Vieira, and A. de Albuquerque Araújo, "A tensor motion descriptor based on multiple gradient estimators," in *SIBGRAPI 2013 - Technical Papers ()*, Arequipa, Peru, aug 2013.

[5] V. Mota, E. Perez, L. Maciel, M. Vieira, and P. Gosselin, "A tensor motion descriptor based on histograms of gradients and optical flow," *Pattern Recognition Letters*, vol. 39, no. 0, pp. 85 – 91, 2014, advances in Pattern Recognition and Computer Vision. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865513003036

[6] H. Maia, A. Figueiredo, O. F.L.M., V. Mota, and M. Vieira, "A video tensor self-descriptor based on variable block matching," *Journal of Mobile Multimedia*, 2015.

[7] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, Mar. 2013. [Online]. Available: http://hal.inria.fr/hal-00803241

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

[9] D. G. Lowe, "Object Recognition from Local Scale-Invariant Features," *Computer Vision, IEEE International Conference on*, vol. 2, pp. 1150–1157 vol.2, Aug. 1999. [Online]. Available: http://dx.doi.org/10.1109/ICCV.1999.790410

[10] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 2003, pp. 1470–1477 vol.2.

[11] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.

[12] O. Kihl, D. Picard, and P.-H. Gosselin, "A unified formalism for video descriptor," in *IEEE International Conference on Image Processing*, 2013.

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