Pattern Recognition Letters 39 (2014) 85–91

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

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

V.F. Mota [a,b,*], E.A. Perez [a], L.M. Maciel [a], M.B. Vieira [a], P.H. Gosselin [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 Research Center, Rennes, France

ARTICLE INFO

*Article history:*
Available online 23 August 2013

Communicated by S. Sarkar

*Keywords:*
Global motion descriptor
Optical flow
Histogram of gradients
Action recognition

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.



## 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 (Wang et al., 2011; Laptev et al., 2008). In our preliminary work, presented in Mota et al. (2012), 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 Mota et al. (2012) 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 (Mota et al., 2012) and tridimensional histogram of gradients (HOG3D) information (E.A. Perez et al., 2012): 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 (Schuldt et al., 2004), UCF11 (also known as UCF YouTube) (Liu and Luo, 2009) and Hollywood2 (Marszałek et al., 2009) video datasets with a non-linear SVM classifier.

## 2. Related work

Laptev et al. (2008) 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 Wang et al. (2011), 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

* Corresponding author. Address: Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, DCC/ICEx, Belo Horizonte, Brazil. Tel.: +55 32 88856321.
*E-mail address:* virginiaferm@dcc.ufmg.br (V.F. Mota).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.08.008

used in the context of action classification for the KTH, Hollywood2, UCF11 and UCF sports datasets.

Also using a bag-of-features strategy, Zhen et al. (2013) 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 (2012) 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 (Zelnik-manor et al., 2001). 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. (2007) apply the Zelnik descriptor (Zelnik-manor et al., 2001) in two different ways: using multiple temporal scales like the original and using multiple temporal and spatial scales.

Solmaz et al. (2012) 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 (Druon et al., 2009; Kihl et al., 2010). 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 $P_{ij}$, 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 $\widetilde{F} = (\widetilde{V}^1(x_1, x_2), \widetilde{V}^2(x_1, x_2))$, using a basis of degree $g$, as:

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

where

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

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 (Westin et al., 1994).

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 (Johansson et al., 2002).

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

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

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 $\tilde{v}_{ij}^1$ and $\tilde{v}_{ij}^2$. 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.