Combining gradient histograms using orientation tensors for human action recognition

Eder A. Perez,Virgínia F. Mota, Luiz Maurílio Maciel, Dhiego Sad, Marcelo B. Vieira
Universidade Federal de Juiz de Fora - MG - Brazil
{*eder.perez, virginia.fernandes, luiz.maurilio, dhiego.sad, marcelo.bernardes*}*@ice.ufjf.br*

Abstract

We present a method for human action recognition based on the combination of Histograms of Gradients into orientation tensors. It uses only information from HOG3D: no features or points of interest are extracted. The resulting raw histograms obtained per frame are combined into an orientation tensor, making it a simple, fast to compute and effective global descriptor. The addition of new videos and/or new action cathegories does not require any recomputation or changes to the previously computed descriptors. Our method reaches 92.01% *of recognition rate with KTH, comparable to the best local approaches. For the Hollywood2 dataset, our recognition rate is lower than local approaches but is fairly competitive, suitable when the dataset is frequently updated or the time response is a major application issue.*

1. Introduction

Human action recognition is a field of research very attractive over the past years. It can be considered one of the key prerequisites for video analysis and understanding. One of the methods widely used to create descriptors for actions is the histogram of oriented gradients (HOG). In general, these methods combine HOG with local information. In this work, we present a novel approach using HOG3D and tensors for computing a global motion descriptor.

A global descriptor based on histogram of oriented gradients is presented by Zelnik *et al* [12]. It is applied on Weizmann video database and is obtained extracting multiple temporal scales through the construction of a temporal pyramid. For each scale, the intensity of each pixel gradient is calculated. Then, a HOG 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 [5] apply the Zelnik descriptor [12] in two different ways: using multiple temporal scales as the original and using multiple temporal and spatial scales.

Laptev *et al* [6] proposed the combination of HOG with histogram of optic flow (HOF) to characterize local motion and shape. Histograms of spatial gradient and optic flow are computed and accumulated in space-time neighborhoods of detected interest points. Similarly to the SIFT descriptor [8], normalized histograms are concatenated to HOG and HOF vectors.

Two dimensional features derived from histograms of oriented gradients have been shown to be effective for detecting human actions in videos. However, according to the viewing angle, local features may be often occluded. One alternative was proposed by Kläser *et al* [4] to avoid the problems presented by HOG2D. This work presented the HOG3D, a descriptor based on histograms of 3D gradient orientations and can be seen as an extension of SIFT descriptor.

HOG based methods do not achieve the best recognition rates for KTH and Hollywood2 dataset. Recently, Quoc *et al* [7] proposed an extension of the Independent Subspace Analysis algorithm (ISA) to learn invariant spatio-temporal features from unlabeled video data. In this approach are employed two important ideas from convolutional neural networks: convolution and stacking. These ideas enable the algorithm to learn a hierarchical representation of the data suitable for recognition. This method present a classification results superior to all previous published results on the Hollywood2, UCF, KTH and YouTube action recognition datasets.

The learning step of the previous works extracts singular features of each database to improve recognition rates. The drawback is the time needed for learning and the scalability of the method in function of the addition of new video information.

Our method uses only information from HOG3D: no features or points of interest are extracted. The resulting raw histograms obtained per frame are combined into an orientation tensor, making it a simple global de-

scriptor. Using a SVM classifier, it achieves recognition rates greater than those found by other HOG global techniques on KTH dataset and a competitive recognition rate, in terms of time and space complexity, for Hollywood2 dataset.

2. Proposed Method

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

$$\vec{g}_t(p) = [dx\ dy\ dt] = \left[\frac{\partial I_j(p)}{\partial x}\ \frac{\partial I_j(p)}{\partial y}\ \frac{\partial I_j(p)}{\partial t}\right], \quad (1)$$

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 and we chose, for simplicity, an uniform subdivision of the angle intervals to populate the $nb_\theta \cdot nb_\psi$ bins:

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

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

2.1 Orientation tensor: coding HOG3D 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. 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 [2].

To empirically reduce interframe brightness unbalance, the histogram of gradients $\vec{h}_f \in \mathbb{R}^{nb_\theta \cdot nb_\psi}$ of a frame f has all of its elements a_k ajusted to ${a_k}^\gamma$, $\gamma = 0.72$. With this reduction of the relative differences between gradient bins, the frame's tensor is given by:

$$T_f = \vec{h}_f \vec{h}_f^T,$$

that carries the information of the gradient distribution of the frame f. Individually, this tensor has the same information of $\vec{h}_f$, but several tensors can be combined to find component covariances. Since T_f is a symmetric matrix, it can be stored with $\frac{m(m+1)}{2}$ elements.

2.2 Global tensor descriptor: series of frame tensors

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

2.3 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 [8], 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 $\vec{h}_j^{a,b}$, $a \in [1, n_x]$ and $b \in [1, n_y]$, with better position correlation. The tensor for frame f is then computed as the addition of all block tensors:

$$T_f = \sum_{a,b} \vec{h}_f^{a,b}\ {\vec{h}_f^{a,b}}^T,$$

which captures the uncertainty of the direction of the m-dimensional vectors $\vec{h}_f^{a,b}$ for the frame f. Thus, the image subdivision does not change the descriptor size and the accumulation described in Section 2.2 is the same.

Another improvement is to accumulate the tensor obtained with the video frame flipped horizontally. The video frame is flipped, the HOG3D is computed for each block, the final tensor is computed (Eq. 2) 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.

3. Experimental Results

Validation set. To validate our descriptor, we use the KTH [10] and Hollywood2 datasets [9].

Experimental protocol. For the KTH dataset, we run a multiclass classifier using a one-against-all strategy and a Bayes criterion for model selection. For the Hollywood2 dataset we run a monoclass classifier using one-against-all strategy, average precision criterion for model selection and crossvalidation. For both datasets, each class is modeled using a SVM classifier with a triangular kernel function with Euclidian distance.

Results. The performance of our method on the KTH dataset is reported in Table 1. We compare our recognition rate against best reported results in literature for techniques using HOG. We note that our descriptor achieves recognition rates greater than those found by other HOG techniques. When compared to other techniques, our descriptor does not outperfoms the best results for Hollywood2 [3, 7], however it achieves a competitive accuracy with a much simpler approach. Only a few parameters are required: HOG3D resolution (nb_θ and nb_ψ) and 2D grid dimension (n_x and n_y). The confusion matrix is shown in Table 2.

Method	Recognition rate
HOG pyramids [5]	72%
Harris3D + HOG3D [4]	91.4%
Harris3D + HOG/HOF [6]	91.8%
HOG3D + Tensor (our method)	**92.01%**
ISA [7]	93.9%
TCCA [3]	95.33%

Table 1. Comparison of recognition rates for the KTH dataset.

	Box	HClap	HWav	Jog	Run	Walk
Box	93.01	6.25	0.00	0.00	0.00	0.70
HClap	3.50	93.75	1.39	0.00	0.00	0.00
HWav	0.70	0.00	98.61	0.00	0.00	0.00
Jog	0.00	0.00	0.00	86.81	15.28	3.47
Run	0.00	0.00	0.00	9.72	84.03	0.00
Walk	2.80	0.00	0.00	3.47	0.70	95.83

Table 2. Confusion matrix for KTH dataset.

A comparison of our method on the Hollywood2 dataset is reported in Table 3. The average precision for each class is shown on Table 4. We can see on Table 3 that local information plays an important role in this dataset and that learning methods improve overall recognition. Our recognition rate is lower than the local approaches but is fairly competitive. Our approach is fast and new video samples and/or entirely new action

Method	Recognition rate
HOG3D + Tensor (our method)	**34.03%**
Harris3D + HOG3D [4, 11]	43.7%
Harris3D + HOG/HOF [6, 11]	45.2%
ISA [7]	53.3%

Table 3. Comparison of recognition rates for the Hollywood2 dataset.

Action	AP	Action	AP
AnswerPhone	17.40%	DriveCar	70.35%
Eat	14.29%	FightPerson	51.35%
GetOutCar	29.84%	HandShake	12.88%
HugPerson	19.47%	Kiss	49.26%
Run	54.03%	SitDown	52.93%
SitUp	9.94%	StandUp	42.93%
		Mean	**34.03%**

Table 4. Average precision (AP) for each class of Hollywood2 dataset using a HOG 8x16 with a 4x4 grid.

Parameters	Hor. Flip	Original
Grid 4x4 HOG 4x8	77.07%	77.65%
Grid 4x4 HOG 8x16	89.80%	88.88%
Grid 8x8 HOG 4x8	76.95%	77.19%
Grid 8x8 HOG 8x16	92.01%	89.34%
Grid 16x16 HOG 4x8	77.07%	77.07%
Grid 16x16 HOG 8x16	88.76%	89.22%

Table 5. Recognition rate for KTH dataset for several parameter sets.

Parameters	Hor. Flip	Original
Grid 4x4 HOG 4x8	31.45%	30.45%
Grid 4x4 HOG 8x16	34.03%	33.64%
Grid 16x16 HOG 4x8	30.65%	30.08%
Grid 16x16 HOG 8x16	33.81%	33.62 %

Table 6. Recognition rate for Hollywood2 dataset for several parameter sets.

categories can be added without any recomputation.

Several combinations of grid size, number of bins and horizontal flip of the image were tested. The recognition rates on KTH dataset are described on Table 5. We can see that these three parameters have a high influence on the recognition rate. The grid size affects how many tensors will be computed per frame, whereas the number of HOG3D bins affects the size of the descriptor. The most interesting influence is from the horizontal flipping, which achieves a recognition rate more than 3% higher compared to the others. The same behavior holds for Hollywood2 dataset with originally 33.64% of recognition rate and 34.03% with horizontal flip as best result (Tab. 6).

In terms of time complexity, the descriptors were computed with an average of 23 frames per second (HOG3D computed twice for horizontal flipping) for the whole Hollywood2 database in an Intel I7 2930MHz processor with 8Gb of memory. For comparison, only the feature extraction step in the work of [3] performs at 1.6 frames per second for Hollywood2. Comparing with [7], its best result is 10 frames per second using a GTX270 GPU for Hollywood2 database. The computation of the space-time derivatives dominates the time complexity of our method. This means that it is highly scalable and suitable to parallel improvements using SIMD instructions, multicore processors and GPUs.

4. Conclusion

In this paper, we presented a method for human action recognition based on the combination of Histograms of Gradients into orientation tensors. The resulting tensor descriptor is a simple but effective approach for video classification. It is simple because of its low complexity in terms of time and space. Only a few parameters are needed, resulting in compact tensor descriptors: a 8×16 HOG3D (128 bins) results in a tensor of only 8256 elements, independently of the frame dimension, video length or grid size. The computation of the space-time derivatives dominates the time complexity of our method, thus it is highly scalable and suitable to parallel improvements using SIMD instructions, multicore processors and GPUs.

It is an effective approach because it reaches 92.01% of recognition rate with KTH, comparable to the best local approaches [7, 3] which have much higher time complexity. For the Hollywood2 dataset, however, we note that local information plays an important role and that learning methods improve overall recognition. Our recognition rate is lower than the local approaches but is fairly competitive. Higher misrecognition may be acceptable when the dataset is frequently updated or the time response is a major application issue. The addition of new videos and/or new action cathegories with our approach does not require any recomputation 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 [11].

5. Acknowledgements

Authors thank Fundação de Amparo à Pesquisa do Estado de Minas Gerais/FAPEMIG, CAPES, and UFJF for funding. This work uses RETIN SVM classifier from ENSEA-UCP, France [1].

References

[1] J. Fournier, M. Cord, and S. Philipp-Foliguet. Retin: A content-based image indexing and retrieval system. *Pattern Analysis and Applications*, 4:153–173, 2001.

[2] B. Johansson, G. Farnebck, and G. F. Ack. A theoretical comparison of different orientation tensors. In *Symposium on Image Analysis*, pages 69–73. SSAB, 2002.

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

[4] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference*, pages 995–1004, sep 2008.

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

[6] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision & Pattern Recognition*, jun 2008.

[7] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, pages 3361–3368, 2011.

[8] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ICCV '99, pages 1150–, Washington, DC, USA, 1999. IEEE Computer Society.

[9] M. Marszałek, I. Laptev, and C. Schmid. Actions in context. In *Conference on Computer Vision & Pattern Recognition*, jun 2009.

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

[11] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

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