Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Beyond Histograms: Why Learned Structure-Preserving Descriptors Outperform Statistical Features

Thomas Guthier, Volker Willert, Julian Eggert

2014

Preprint:

This is an accepted article published in European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN) 2014. The final authenticated version is available online at: https://doi.org/[DOI not available]

Beyond Histograms: Why Learned Structure-Preserving Descriptors Outperform HOG

Thomas Guthier[1], Volker Willert[1] and Julian Eggert[2]

1- TU Darmstadt - Control theory and robotics lab
Landgraf-Georg-Str. 4, 64283 Darmstadt, - Germany

2- Honda Research Institute Europe
Carl-Legien-Str. 30, 63073 Offenbach - Germany

Abstract.

Statistical image descriptors based on histograms (e.g. SIFT [1], HOG [2]) are widely used in image processing, because they are fast and simple methods with high classification performance. However, they discard the local spatial topology and thus lose discriminative information contained in the image. We discuss the relations between HOG and VNMF descriptors, i.e. structure free histograms versus learned structure-preserving patterns. VNMF is a shift-invariant, sparse, non-negative unsupervised learning algorithm [8, 9, 5], that provides a distinct decomposition of the input into its parts. The VNMF descriptor outperforms the statistical HOG descriptor, because it preserves spatial topology leading to better classification results on real-world human action recognition benchmarks [11, 12].

1 Introduction

Since the introduction of the SIFT descriptors by Lowe [1], hand-designed *histogram* based features (e.g. SIFT, HOG, HOF, MBH [2]) are successfully applied in various classification tasks in image processing, i.e. object recognition [1], pedestrian detection [2] or human action recognition [10].

The basic ingredient all these descriptors have in common is the statistical description of gradient structures via a histogram, where each entry corresponds to a discrete gradient direction, a so called bin. The descriptors are able to represent class discriminative structures and are computationally cheap. However, the simplistic histogram description has natural limitations. It discards local spatial relations between structure elements, i.e. the topology of the gradients is neglected, because the explicit spatial occurrence of the gradients is lost in the histogram representation. Furthermore, the number of elements in each descriptor block is limited by the number of bins.

Another class of discriptors is based on structural *patterns* instead of histograms. Here the input is reconstructed by a set of learned patterns, where each pattern (or *basis vector*) represents a part of the structure of the input. The occurrence of the patterns used for the reconstruction of the input, the *activations*, are then applied as descriptors. In contrast to the HOG, the activations preserve the local structure that is encoded in the corresponding patterns and in addition the number of patterns is not as limited as the number of histogram bins. However, the choice of the right patterns is non trivial. To overcome the need of hand-designing a set of patterns for every specific application, unsupervised learning algorithms used to learn natural image statistics [3, 8, 9, 5] can learn the set of generative patterns. In [4] Le et al. propose the use of ISA, a two layered

extension of the well known ICA [3], to extract spatio-temporal features for human action recognition. They show that the learned pattern features outperform the classic HOG/HOF and 3D HOG descriptors on multiple human action recognition datasets.

In this paper, we follow the idea of [4] and apply learned patterns as features for human action recognition. Instead of enforcing independence of the extracted components as in ICA, we use VNMF, a sparse, non-negative and translation-invariant learning algorithm [5], that decomposes the input into distinctive parts. Our main contribution is the analysis of the weakness of the HOG descriptor, which is empirically confirmed by our classification experiments on the Weizmann [11] and UCF Sports [12] human action recognition datasets.

The outline of the paper is as follows: first, the HOG descriptor and our VNMF pattern descriptor are introduced and the functional differences are discussed. Next, we present the classification framework and the experimental results, followed by a short summary and conclusion.

2 Descriptor Comparison

2.1 Histogram of Oriented Gradients (HOG)

The HOG descriptor as introduced by Dalal and Triggs [2] consists of two parts: 1.) a grid of 50% overlapping blocks and 2.) a normalized histogram of the oriented gradients in each of the blocks. The block descriptor is build in three steps: first, each gradient vector (in case of HOF, optical flow vectors) is binned into one of e.g. $b = 8$ ($b :=$ number of bins) distinct directions. Second, for each block the gradient vectors are summed up for each bin, resulting in a histogram with b elements. To achieve invariance to contrast changes, the histograms are normalized using e.g. the Euclidean norm.

2.2 Learned Structure Preserving Descriptor using VNMF

Our learned structure preserving descriptor is a biologically inspired simple cell, complex cell method applied to the gradient amplitudes of input images or dense optical flow fields. The complex cells are realized by a summation pooling block grid, that is identical to the block grid of the HOG descriptor. The local descriptor per block is the simple cell response to a set of patterns, that are learned with VNMF [5], which enforces a parts-based decomposition. Fig. 1 shows 24 basis vectors $\mathcal{W} \in \mathbb{R}^{\bar{X}\times\bar{Y}\times J}$ ($X, Y :=$ pixels in x,y-dimension, $J :=$ number of basis vectors)[1] learned on the gradient amplitudes of images of the Weizmann human action recognition dataset [11]. Each basis vector describes a local gradient structure, such as bars or corners.

The simple cell response or *activation* $h_{jn}(\mathbf{m}) \in \mathbf{H}_{jn} \in \mathbb{R}^{X\times Y}$ of an input $\mathbf{V}_n \in \mathbb{R}^{X\times Y}$ and a basis vector $\mathbf{W}_j$ is calculated in a generative way by minimizing the energy function

$$E_n = \frac{1}{2}\|\mathbf{V}_n - \mathbf{R}_n\|_2^2 + \frac{1}{2}\lambda_P \sum_{j,\mathbf{m}} \mathbf{R}_{j\mathbf{m}n}^{\top}\Big(\mathbf{R}_n - \mathbf{R}_{j\mathbf{m}n}\Big) + \lambda_H \sum_{j,\mathbf{m}} h_{jn}(\mathbf{m}), \quad (1)$$

[1] Due to the translation invariance, the maximum receptive field size of the basis vectors $\bar{X} \times \bar{Y}$ can be smaller then the input size $X \times Y$. For our experiments $\bar{X} \times \bar{Y} = 16 \times 16$.

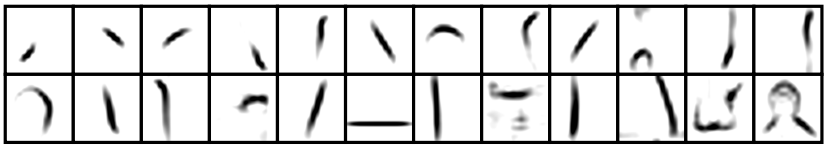

Fig. 1: Set $\mathcal{W}$ of $J = 24$ basis vectors with a maximum receptive field size of $\bar{X} \times \bar{Y} = 16 \times 16$ learned with VNMF [5] on gradient amplitudes of the Weizmann dataset.

with the reconstruction

$$\mathbf{R}_n = \sum_{j,\mathbf{m}} \mathbf{R}_{j\mathbf{m}n} = \sum_{j,\mathbf{m}} \mathrm{conv}_2(h_{jn}(\mathbf{m}), \mathbf{W}_j). \tag{2}$$

All elements are strictly *non-negative*, i.e. $\mathbf{H}_{jn}, \mathbf{W}_j, \mathbf{V}_n, \mathbf{R}_n \geq 0$. The first part of the energy function (1) forces the activations to minimize the difference between the input and the reconstruction (2), while the second part penalizes overlaps between the partial reconstructions $\mathbf{R}_{j\mathbf{m}n}$ coming from different activations $h_{jn}(\mathbf{m})$. The last part favors sparse activations.[2] The activations are calculated on the entire input image by:

1. initialize: $\mathbf{H}_{jn} = \mathrm{corr}_2(\mathbf{V}_n, \mathbf{W}_j)$,
2. loop till convergence: $\mathbf{H}_{jn} \rightarrow \mathbf{H}_{jn} \circ \frac{(\nabla_{\mathbf{H}_{jn}} E_n)^-}{(\nabla_{\mathbf{H}_{jn}} E_n)^+}$,
3. threshold and binarize: $\mathbf{H}_{jn} = \begin{cases} 0, & \text{if } \mathbf{H}_{jn} \leq \tau, \\ 1, & \text{else,} \end{cases}$

with the positive and negative gradient components $(\nabla_{\mathbf{H}_{jn}} E_n)^+$ and $(\nabla_{\mathbf{H}_{jn}} E_n)^-$. The input $\mathbf{V}_n$ is normalized with the infinity-norm and the free parameters are set to $\lambda_P = 0.5$, $\lambda_H = 0.2$ and $\tau = 0.2$.

2.3 Similarities and Differences of both Descriptors

The main similarity is that for both descriptor types, the input image is fragmented in a grid of overlapping blocks (see Fig. 3 right). The block grid captures the *global* spatial relations between the block descriptors, e.g. the upper blocks are more likely to describe head shapes, while the lower blocks reflect features corresponding to leg poses and movements. However, the features differ in the way they describe what is inside the blocks, i.e. the *local* topological information in the image.

Fig. 2 shows the HOG descriptors and the VNMF activations for two 16×16 blocks of an input image. The HOG descriptors for both blocks are identical, because the blocks differ only in the spatial structure of the gradients, and not in the amount of gradient vectors, which is captured by the histograms. In contrast, the pooled activations H_{block} are different, because different basis vectors are used for the reconstruction. This

[2] For more details on the transformation parameter $(\mathbf{m})$, the energy terms and the notation see [8, 9, 5].

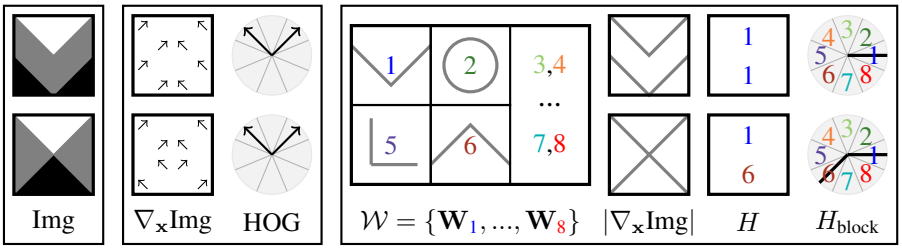

Fig. 2: The HOG and VNMF descriptor for two 16×16 blocks (upper and lower row). From left to right: input block (Img), gradients ($\nabla_{\mathbf{x}}$Img), HOG descriptor (HOG) with $b = 8$ bins, basis vector set ($\mathcal{W}$) with $J = 8$ basis vectors, gradient amplitudes ($|\nabla_{\mathbf{x}}\text{Img}|$), activations for each of the two inputs (H) and pooled activations (H_{block}). The activations 1 and 6 mark the position, where the corresponding basis vectors $\mathbf{W}_1$ and $\mathbf{W}_6$ are placed for the reconstruction of the input.

artificial example illustrates why in principle pattern-based descriptors are able to preserve the *local* topological information in cases where the histogram discriptor discards this information.

Another difference is *how* the image structure is described. The binning approach is simple and computationally cheap. Nevertheless, the number of bins b is limited, because a finer binning makes the HOG descriptor less invariant and may not increase its discriminative properties. On the contrary, the sparsity constraints in the VNMF algorithm allow the learning of an overcomplete basis, so the number of basis vectors J is not as limited as the number of bins, because the more basis vectors are learned, the more image structures can be explicitly represented. Besides, the basis vectors are learned and not hand-crafted as the HOG, so they are easier to adapt to different kinds of input data.

In summary, the VNMF descriptors should outperform the HOG descriptors if the *local topological information* is important for modelling discriminative image descriptors. In the following, this hypothesis is evaluated in classification experiments. In addition, we vary the number of basis vectors and histogram bins and compare the corresponding classification results.

3 Human Action Recognition

The goal of human action recognition is to classify a video depending on the performed action. For the classification we choose a four stage hierarchical system, with 1.) preprocessing, 2.) feature extraction, 3.) dimension reduction and 4.) classification.

In the preprocessing, the person whose action is classified is centered in a 128×128 window for every frame in the video. For each of the windows the spatial gradients and a dense optical flow field [6] are calculated. Then the two different kind of descriptors (VNMF and HOG) are calculated for the gradients and the optical flow field. The feature dimension for each of the 225 overlapping pooling blocks is given by J or b,



Fig. 3: Left: nine images from each class of the UCF-Sports dataset [12]. Right: centered person with 2×2 overlapping blocks and the corresponding descriptors.

so for $J = 8$ we have 1800 features per frame. To better deal with the small amount of training samples, we reduce the dimension down to 100, using non-negative sparse coding [8] for both, the gradient and the optical flow features individually. The features per frame are then temporally pooled so that each video has just one final feature vector, which is classified using a multiclass *Support Vector Machine* (SVM) [7] with radial basis functions as kernels, that is trained in leave-one-out experiments.

Input Type	Gradient (∇_x)			Optical Flow (OF)			∇_x+OF		
J / b	8	16	24	8	16	24	8	16	24
VNMF	0.87	0.83	0.83	0.99	0.98	0.97	0.97	0.97	0.94
HOG	0.70	0.72	0.68	0.92	0.92	0.94	0.90	0.91	0.92

Table 1: Classification results for the VNMF and HOG features for different inputs (gradients and dense optical flow fields) and number of basis vectors/histogram bins ($J, b = \{8, 16, 24\}$) for the 10-class Weizmann dataset.

Table 1 shows the results for both descriptors on the Weizmann dataset. Our main assumption, that the VNMF should outperform the HOG descriptor, is confirmed by the experiments. The difference is particular high in case of the gradient descriptors ($\sim$15%) and still significant for the optical flow and the combined descriptors ($\sim$7%). The feature dimension per block (J,b) has no significant influence for both descriptor types, which is surprising for the VNMF. The results imply that the local discriminative gradient statistics can be captured by as few as 8 translation invariant patterns.

VNMF			HOG			Related Work	
∇_x	OF	∇_x+OF	∇_x	OF	∇_x+OF	Le [4]	Wang [10]
0.81	0.80	0.87	0.77	0.78	80	0.87	0.88

Table 2: Classification results for the UCF Sports [12] dataset for VNMF and HOG descriptors ($J, b = 8$) compared to state-of-the-art algorithms [4, 10].

Table 2 shows the classification result for the VNMF and HOG descriptors compared to two state-of-the-art bag-of-word methods [4, 10] on the challenging UCF

Sports dataset (see Fig. 3), which contains different view-point, camera motion and strong variations in the individual performance of the actions. Again, the VNMF outperforms the HOG by $\sim$7%. The VNMF is competitive with the two state-of-the-art methods, that both use high dimensional features ($>$100 elements) calculated in $16 \times 16 \times 10$ space-time-volumes; to the contrary we only use 8 patterns of gradients and optical flow for a single 16×16 block.[3]

4 Summary & Conclusion

We show that our VNMF descriptors outperform the HOG and thus confirm the results reported in [4], i.e. learned image features are more discriminative than hand-designed histograms. Unlike the ISA, that is applied in [4], the VNMF has *sparsity* and *non-negativity* constraints, that, combined with the translation-invariant learning [9], yield a small set of generative patterns. Besides its lower dimensionality, the VNMF set is as discriminative as the high dimensional set extracted with ISA. Because of our distinctive comparison between the HOG and VNMF descriptor (see Fig. 2), we are able to verify that preserving the local topological information is the essence for increasing the discriminativity of image descriptors.

References

[1] D.G. Lowe, Distinctive image features from scale-invariant keypoints, *Int. Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[2] N. Dalal, B. Triggs and C. Schmid, Human detection using oriented histograms of flow and appearance, *European Conf. on Computer Vision* (ECCV), pp. 428-441, 2006.

[3] A. Hyvearinen, J. Hurri and P.O. Hoyer, Natural Image Statistics, Springer, 2009.

[4] 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 *IEEE Int. Conf. on Computer Vision and Pattern Recognition* (CVPR), 2011.

[5] T. Guthier, J. Eggert and V. Willert, Unsupervised learning of motion patterns, *European Symposium on Artificial Neural Networks* (ESANN), 2012.

[6] T. Guthier, A.Schnall, K.Kreuter, V. Willert and J. Eggert, Non-negative Sparse Coding for Motion Extraction, *IEEE Int. Joint Conf. on Neural Networks* (IJCNN), 2013.

[7] C.C. Chang and C.J. Lin., LIBSVM : a library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, 2011.

[8] J. Eggert and E. Koerner, Sparse coding and NMF, *IEEE Int. Joint Conf. on Neural Networks* (IJCNN), vol. 4, pp. 2529-2533, 2004.

[9] J. Eggert, H. Wersing and E. Koerner, Transformation-invariant representation and NMF, *IEEE Int. Joint Conf. on Neural Networks* (IJCNN), 2004.

[10] H. Wang, A. Klaser, C. Schmid, and C.L. Liu, Dense trajectories and motion boundary descriptors for action recognition. *Int. Journal of Computer Vision*: pp. 1-20, 2013.

[11] M. Blank, L. Gorelick, E. Shechtman, M. Irani and R. Basri, Actions as Space-Time Shapes, *IEEE Int. Conf. on Computer Vision* (ICCV), 2005.

[12] M.D. Rodriguez, J.Ahmed, and M.Shah, Action MACH: A Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition, *IEEE Int. Conf. on Computer Vision and Pattern Recognition* (CVPR), 2008.

[3] The comparison is somewhat problematic, because the bag-of-words approach does not use a figure centric representation.