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

Learning Associative Spatiotemporal Features with Non-Negative Sparse Coding

Thomas Guthier, Steve Gerges, Volker Willert, Julian Eggert

2013

Preprint:

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

Learning Associative Spatiotemporal Features with Non-Negative Sparse Coding

Thomas Guthier[1], Steve Gerges[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.
Motion features based on optical flow are very powerful in tasks such as the recognition of human actions or gestures. Usually, they are combined with gradient information to form a set of spatiotemporal features. However, humans can recognize gestures and actions and thus derive the implied motion out of static images alone. We model this associative recognition within a learned hierarchy of non-negative sparse coding layers. In the first stages, topology preserving gradient and motion features are processed separately. Afterwards, they are projected onto a combined inner representation, that is learned during the training phase. We show, that during recognition the learned, combined representation improves the recognition of human actions, even in the absence of explicit motion information.

1 Introduction

The human visual system has a remarkable capacity to recognise biological motion, such as human actions or gestures. While the underlying neural process is not fully understood, there is strong evidence, that both explicit motion information as well as form and texture information are involved [7]. The fact that humans can recognize actions out of still images indicates that there exists a combined representation that couples motion and texture information. In this paper, we show that such a combined representation can be learned via non-negative sparse coding (NNSC) [10] and that its associative properties improve the action recognition results on datasets containing different kinds of full-body motions, like walking, running, jumping, etc.

Unsupervised learning algorithms such as latent dirichlet allocation (LDA), deep learning or non-negative matrix factorization (NMF) [1] are often applied to multi-modal learning tasks [5], where the multimodalities are mostly combinations of audio, video or label data. Even though spatiotemporal features are widely used in recognition tasks, only few publications analyse the combined learning of motion and gradient information. In [6] the authors use LDA to train their midlevel features on top of statistical gradient and optical flow features based on orientation histograms. During the detection phase, only the gradient features are used to trigger the pre-learned midlevel representation. They show, that their so called *flobject* (flow-object) analysis improves the results in car detection scenarios.

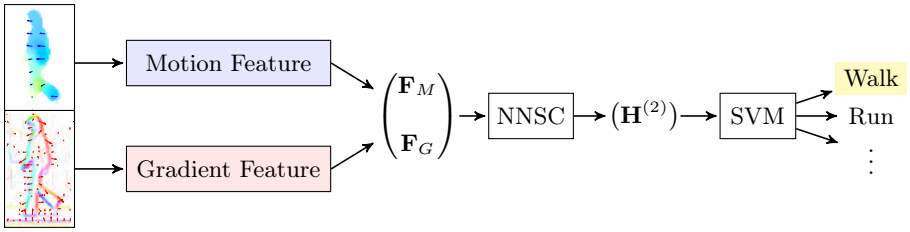

Fig. 1: Classification structure for an example optical flow and gradient vector fields of a walking sequence.

We follow the idea of the flobject analysis but transfer it to the task of human action recognition, where motion information is crucial for the recognition process [2]. In addition, our approach differs in two major aspects. Instead of histogram-based features, like histogram of oriented gradients (HOG) or histogram of optical flows (HOF), that neglect the spatial topology, we use our previously introduced VNMF algorithm [4], to *learn* a feature dictionary for gradient and optical flow input. The VNMF algorithm is based on translation invariant NNSC [11] and the learned features are sparse and preserve the topology information. In the next stage we use an unsupervised learning algorithm to combine the low level VNMF gradient and motion features to midlevel features. For the learning of the combined midlevel features we again apply NNSC instead of LDA. Our combined gradient and motion features outperform a similar approach [3] that combines HOG and HOF features with NMF, stressing the need for topology preserving representations.

2 Classification Structure

Our classification framework is a four stage hierarchy as depicted in Fig. (1). In the first stage the gradients of the input images along with the optical flow fields[1] are computed. The second stage is a VNMF feature extraction, which is divided into two parallel streams, one for the gradient and the other for the motion features. In the third stage, the two feature vectors are combined to form the input of the next NNSC algorithm. The activities of this second learning stage are then classified using a *support vector machine* (SVM)[2]. We get a confidence value for each class for each incoming image pair of the video sequencesand average the confidence values of 20 consecutive image pairs. The final classification label is set to the class with the highest confidence value.

[1] Computed with the publicly available algorithm described in [8].

[2] We use the publicly available libSVM [9] implementation, with radial basis functions as kernel functions, soft margins and one vs one classification for our multi-class problem.

2.1 Feature Extraction

The feature extraction for the gradients and the optical flow fields is the VNMF algorithm [4]. It is an unsupervised learning algorithm based on translation invariant non-negative sparse coding. A given vector field $\mathbf{V}_i^d$, with $i \in \{1, ..., I\}$, $I \mathrel{\widehat{=}}$ number of input frames and $d \in \{x, y\}$, is split up into the non-negative representation $\mathbf{V}_i^f$, with $f \in D$, with $D = \{x+, y+, x-, y-\}$ representing four directions. It is reconstructed $\mathbf{V}_i^f \approx \mathbf{R}_i^f = \sum_{j,m} h_{ij}^{(m)}(T^{(m)}\mathbf{W}_j^f)$ by a translation invariant linear superposition of basis vectors $\mathbf{W}_j^f$, with $j \in \{1, ..., J\}$, $J \mathrel{\widehat{=}}$ number of basis vectors. We use 12 basis vectors for the gradients and 12 basis vectors for the optical flow fields. The weight of each basis vector set $\mathbf{W}_j = \{\mathbf{W}_j^f | \ \forall f \in D\}$ at each position for the input image i is represented by the activity image $\mathbf{H}_{ij}$. The activities $\mathbf{H}_{ij}$ and basis vectors $\mathbf{W}_j^f$ are learned on the training data by minimizing the energy function

$$E = \frac{1}{2}\sum_{i,f} \|\mathbf{V}_i^f - \mathbf{R}_i^f\|_2^2 + \lambda_P \sum_{i,f}\Big(\|\mathbf{R}_i^f\|_2^2 - \sum_{j,m} \|h_{ij}^{(m)}(T^{(m)}\mathbf{W}_j^f)\|_2^2\Big) + \lambda_H \sum_{i,j} |\mathbf{H}_{ij}|_1$$

using multiplicative gradient descent. The parameters are set to $\lambda_P = 0.5$ and $\lambda_H = 0.2$. During detection the pre-learned basis vectors are used and only the activities, thus the occurence of the basic parts in the inputs, are extracted. The basis vectors for the gradient and flow fields of the first dataset are shown in Fig.(2). The motion patterns roughly describe body parts and the gradient patterns consist of connected edge configurations. The topology of the gradient features that form certain shapes of body-parts can also be found in the topology preserving motion patterns. Shape is implicitly coded in the motion patterns via motion discontinuities.

To be invariant to local shifts and to reduce the dimensionality of the activity images, the activities are pooled using overlapping pooling blocks. The activities are pooled inside a person centered window (80x100 pixels) that we extract out of the segmentation masks that are provided with the dataset. We choose 8x10 pooling blocks with a 50% overlap to each of its neighbouring blocks. The maximum activities of each block and every basis vector are stored in the feature vector $\mathbf{F}$. Thus the feature dimension is 960 for the gradient features $\mathbf{F}_G$ and the motion features $\mathbf{F}_M$ each.

2.2 Combined Midlevel Representation

In the third stage the pooled activities $\mathbf{F}$ of the feature stage are used as inputs for the next layer of non-negative sparse coding which projects them onto the activities $\mathbf{H}^{(2)}$. The main difference to the VNMF algorithm of the feature layer is, that there is no translation invariance, because the pooling at the first stages already compensates for local shifts. The energy function of this second stage NMF is

$$E = \frac{1}{2}\sum_{i} \|\mathbf{F}_i - \sum_{j} \mathbf{H}_{ij}^{(2)}\mathbf{W}_j^{(2)}\|_2^2 + \lambda_H \sum_{i,j} |\mathbf{H}_{ij}^{(2)}|_1$$

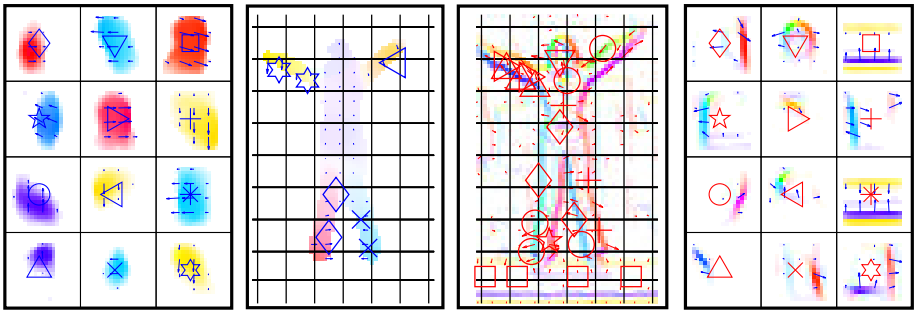

Fig. 2: The left part shows the 12 basic motion patterns and an example flow field marked with the most prominat activities $\mathbf{H}_i > 0.2 * \max(\mathbf{H}_i)$. The grid on the flow field shows the pooling cells, where four neighbouring cells form one block. On the right the 12 basic gradient patterns and an example gradient image is shown. While the motion patterns are formed like body parts, the gradient patterns describe edge configurations that either belong to body parts or the background structure.

with $\lambda_H = 0.1$. We learn 100 basis vectors $\mathbf{W}^{(2)}$, which encode combinations of the pooled activities $\mathbf{F}$ and which represent occurrences of the basic motion patterns or gradient configurations. Since the basic motion patterns encode body parts, the combined representations $\mathbf{W}^{(2)}$ are likely to encode body part configurations, i.e. body poses at different points in time during an action cycle.

3 Flow-object

The idea of the flow-object analysis is, that the discriminative properties of the midlevel representation $\mathbf{W}^{(2)}$ improves for the gradient features, when $\mathbf{W}^{(2)}$ is learned with the corresponding motion representation. Thus, the co-occurrence of action specific motion patterns and action unspecific gradient patterns leads to action specific combined representations $\mathbf{W}^{(2)}_{GM}$. When a still image of an action is shown, only the gradient features can be used to trigger the midlevel representation. But since the combined representation $\mathbf{W}^{(2)}_{GM}$ is itself action specific, the flow-object features should have superior classification results over the midlevel features $\mathbf{W}^{(2)}_{G}$, that are trained with gradient information alone.

For our experiments we thus distinguish between three kinds of features: First $\mathbf{H}^{(2)}_{G}$, here $\mathbf{F}_G$ is used for both, the learning of $\mathbf{W}^{(2)}_{G}$ and for the detection. The combined features $\mathbf{F}_{GM}$ are used for learning the midlevel representation $\mathbf{W}^{(2)}_{GM}$ as well as for detecting $\mathbf{H}^{(2)}_{GM}$. Finally, the flow-object feature $\mathbf{H}^{(2)}_{Flob}$, where $\mathbf{W}^{(2)}_{Flob}$ is learned with the combined features $\mathbf{F}_{GM}$ and for the detection only the gradient features $\mathbf{F}_G$ are used. $\mathbf{W}^{(2)}_{Flob}$ is almost equivalent to $\mathbf{W}^{(2)}_{GM}$, except that the projection dimensions for the flow features $\mathbf{F}_M$ are discarded.

Proposed Method	Result	Related Work	Result
Grad. $\mathbf{H}_G^{(2)}$	80%	HOG+NMF [3]	$\sim$ 74%
Flow-object $\mathbf{H}_{Flob}^{(2)}$	90%	HOG/HOF+NMF [3]	94,5%
Grad.+Motion $\mathbf{H}_{GM}^{(2)}$	98%	Blank et. al. [12]	99,6%

Table 1: Results on the Weizmann human action recognition dataset for our proposed method and related work. Note that the related work approaches apply leave-one-out experiments with 8 training persons, while we have used a fixed set of 4 persons for the training of the features and the classifier.

4 Experiments

We evaluate our algorithms on the Weizmann human action recognition dataset [12], which consist of 9 persons performing 10 different actions. The two VNMF algorithms as well as the NNSC and the SVM model are learned on the training data. We choose four persons for training and evaluated the classification on the remaining five persons. For the reported experiments we choose the first four persons for training. We also permuted the training persons and found no qualitative differences in the results.

The free SVM parameters, the soft margin and the gamma value of the radial basis functions, were optimized using 5-fold cross-validation on the training data. The detection rates are given in Table 1 and show two main results:

First, the combined VNMF features reach 98% detection rate and are thus competitive with the state-of-the-art. It is interesting to note, that our topology preserving features outperform a similar method [3], that combines HOG and HOF features with NMF and SVM classification. In both cases, whether only gradient information is used or gradient and motion features are applied, we get a significantly improvement of the recognition rate.

Second, the flow-object features beat the gradient-only features by 10%. Since the only difference between the two approaches lies in the training of the midlevel representation $\mathbf{W}^{(2)}$, we can conclude that the discriminant motion features are succesfully associated with the corresponding gradient features and that this information is stored in the combined representation $\mathbf{W}^{(2)}$.

5 Summary & Conclusion

Our learned topology preserving gradient and motion features outperform the commonly applied statistical features based on histograms on a activity recognition dataset. This is a strong indication for the discriminative power of topology information in recognition tasks. However, further experiments have to evaluate if the spatial configuration of gradient and optical flow patterns improve recognition results in general or whether the robustness of histogram based representations is necessary to deal with noise and invariances.

In addition, our experiments show that a learned, combined midlevel representation of gradient and motion features can improve the classification results even in the absence of the motion features during detection. This flow-object analysis shows that even static classification can benefit from motion information during the training phase.

Finally we conclude that sparsity and non-negativity constraints may be desirable properties that lead to a suitable representational basis for the combination of appearance and motion information for movement recognition.

References

[1] D.D .Lee and S. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature*, vol. 401, pp. 788-791, 1999.

[2] J.K. Aggarwal and M.S. Ryoo, Human activity analysis: A review, *ACM Computing Surveys*, vol. 43, no. 3, 2011.

[3] T. Mauthner, P.M. Roth and H. Bischof, Instant Action Recognition, *Scandinavian Conf. on Image Analysis*, pp. 1-10, 2009.

[4] T. Guthier, J. Eggert and V. Willert, Unsupervised learning of motion patterns, *European Symposium on Artificial Neural Networks* (ESANN), 2004.

[5] J. Driesen, H. van Hamme and W.B. Kleijn, Learning from images and speech with non-negative matrix factorization enhanced by input space scaling, *IEEE Spoken Language Technology Workshop* (SLT), pp. 1-6, 2010.

[6] P.S. Li, I.E. G. and B.J. Frey, Learning better image representations using 'Flobject Analysis', *IEEE Conf. on Computer Vision and Pattern Recognition,* (CVPR), pp. 2721-2728, 2011.

[7] R. Blake and M. Shiffrar, Perception of Human Motion, *Annual Review of Psychology*, vol. 58, pp. 47-73, 2007.

[8] D. Sun, S. Roth and M.J. Black, Secrets of optical flow estimation and their principles, *IEEE Conf. on Computer Vision and Pattern Recognition* (CVPR), pp. 2432-2439, 2010.

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

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

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

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