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

Unsupervised learning of motion patterns

Thomas Guthier, Julian Eggert, Volker Willert

2012

Preprint:

This is an accepted article published in ESANN 2012 Proceedings. The final authenticated version is available online at: https://doi.org/[DOI not available]

Unsupervised Learning of Motion Patterns

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

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.
Neurophysiological findings suggest that the visual cortex of mammals contains neural populations that are sensitive to specific motion patterns. In this paper, we present a new method to learn such patterns in an unsupervised way. To represent motion, dense optical flow fields of videos containing humans performing several actions like walking and running are estimated. We introduce VNMF, an extension of the translation invariant NMF that works on flow fields, along with a new energy term that enforces parts-basedness. VNMF incorporates three principles found in neural motion processing: Sparsity, non-negativity and direction selectivity. We find that the extracted motion patterns are shaped like body parts, which supports the idea that the representation of biological motion is directly linked to the shape of an object.

1 Introduction

The perception of biological motion is an important capability for humans as well as for artificial vision systems. Therefore it is researched in a variety of disciplines, such as neurophysiology, psychology and computer vision.

There is an ongoing discussion in neurophysiology about how biological motion is represented. The point-light-walker experiments of Johansson [1] seem to indicate that the movement of the joints or body parts alone are sufficient to describe human motion. Other theories favor the idea that human motion can be represented as a time sequence of body shapes [3]. They refer to studies of patients whose early motion processing areas (MT,MST) were damaged, but who could still perceive biological motion. For a survey on the discussion see e.g. Blake et al. [7]. It seems plausible that both, the body shape as well as the specific movement of the body parts, contribute to the recognition of complex motion.

Following this idea we try to learn representations of biological motion based on dense optical flow fields that encode the motion of each pixel in the image. We found that our learned optical flow components are shaped like the underlying body structure of humans and are therefore suited to simultaneously represent the motion as well as the shape, even without using explicit appearance properties such as color or texture.

In related work Black et al. [6] learned optical flow models using principal component analysis (PCA) on videos of walking humans. Their extracted models are holistic rather then parts-based and do not have sparse activity patterns.

Dean et al. [10] used an extended version of the *sparse coding* (SC) algorithm by Olshausen et al. [4] to extract basic space-time volumes out of videos including human motion. Eggert et al. [11] combined SC with the *non-negative matrix factorization* (NMF) of Lee et al. [5] and translation invariance, allowing the reconstruction to position basic components at arbitrary locations. NMF adds non-negativity constraints to the activities and the basis vectors to achieve a more parts-based decomposition of the input. Based on these constraints, Lee et al. [5] developed a multiplicative gradient descent which speeds up the minimization process significantly.

The contribution of this paper is threefold. We extend NMF to work on vector fields, introduce a new orthogonality term to enforce parts-basedness and finally learn basic motion components from human motion sequences which allow a sparse and parts-based composition of motion patterns.

2 Vector NMF (VNMF)

SC is based on the idea that each input image $\mathbf{V}_i \in \mathbb{R}^P$ out of a set of images $\mathcal{V} \in \mathbb{R}^{P \times I}$, with $P \mathrel{\widehat{=}}$ number of pixels and $I \mathrel{\widehat{=}}$ number of input images, can be represented as a weighted sum of basis vectors, $\mathbf{V}_i \simeq \mathbf{R}_i = \sum_j H_{ij}\mathbf{W}_j$. The SC algorithm from Olshausen et al. [4] minimizes the energy function

$$E = E_R + \lambda_H E_H = \frac{1}{2}\sum_i \|\mathbf{V}_i - \sum_j H_{ij}\mathbf{W}_j\|_2^2 + \lambda_H \sum_{i,j} |H_{ij}|_1, \qquad (1)$$

by iteratively updating the activities H_{ij} and the basis vectors $\mathbf{W}_j$ via gradient descent. We now add non-negativity constraints to $\mathbf{V}_i \geq 0, \mathbf{W}_j \geq 0$ and $H_{ij} \geq 0$. Based on these constraints we can use the update rules

$$H_{ij} \leftarrow H_{ij} \odot \frac{(\nabla_{H_{ij}} E)^{neg}}{(\nabla_{H_{ij}} E)^{pos}}, \qquad \mathbf{W}_j \leftarrow \mathbf{W}_j \odot \frac{(\nabla_{\mathbf{W}_j} E)^{neg}}{(\nabla_{\mathbf{W}_j} E)^{pos}}, \qquad (2)$$

with $(\nabla E)^{\{pos,neg\}}$ being the positive or negative contribution to the gradient of $E : \nabla E = (\nabla E)^{pos} - (\nabla E)^{neg}$. This leads to the well-known NMF update equations as introduced by Lee et al. [5] including sparsity.

2.1 Translation Invariance

Dean et al. [10] report that randomly choosing samples in images is inappropriate to learn sparse representations of motion. Therefore they use a feature point extractor to find areas with interesting motion patterns which they use as input for their algorithm. The extracted basis vectors then depend on the reliability of this preprocessing, which is not desirable. We overcome this preprocessing step by using a translation invariant version of the NMF [11]. For this purpose, instead of approximating the reconstruction by $\mathbf{R}_i = \sum_j H_{ij}\mathbf{W}_j$ we use $\mathbf{R}_i = \sum_{j,m} H_{ij}^{(m)} T^{(m)}\mathbf{W}_j$, where each $T^{(m)}\mathbf{W}_j$ describes a translation of the basis vector $\mathbf{W}_j$ by a translation vector m, and each transformed basis vector is weighted by its translation-specific activity $H_{ij}^{(m)}$.

2.2 Non-negativity via Direction Specific Features

Unlike images, flow fields $\mathbf{V}_i^d$, with $d \in \{x, y\}$, can have positive and negative values. The shift invariant SC energy equation for flow fields with $H_{ij}^{(m)} > 0$ is

$$E = \frac{1}{2} \sum_{i,d} \|\mathbf{V}_i^d - \sum_{j,m} H_{ij}^{(m)} T^{(m)} \mathbf{W}_j^d\|_2^2 + \lambda_H \sum_{i,j,m} H_{ij}^{(m)}. \tag{3}$$

We now want to keep the parts-based properties from non-negative decomposition algorithms for flow fields. Our approach is motivated by neurophysiological experiments of Haag et al. [2] who show that early motion processing includes four direction-specific input layers. Each layer represents one of the four main directions of motion: *up*, *down*, *left* and *right*. In analogy to Ding et al. [8] we use a semi-NMF by seperating the two components $\mathbf{V}_i^x$ and $\mathbf{V}_i^y$ of each optical flow field into a positive and negative component,

$$(\mathbf{V}_i^{d+})_\mathbf{x} = \frac{\left|(\mathbf{V}_i^d)_\mathbf{x}\right| + (\mathbf{V}_i^d)_\mathbf{x}}{2}, \quad (\mathbf{V}_i^{d-})_\mathbf{x} = \frac{\left|(\mathbf{V}_i^d)_\mathbf{x}\right| - (\mathbf{V}_i^d)_\mathbf{x}}{2}, \tag{4}$$

for every pixel $\mathbf{x} \in \{1, ..., P\}$ and obtain the four features $\mathbf{V}_i^{ds} \geq 0$, with $s \in \{+, -\}$. The input is now represented as $\mathbf{V}_i^d = \mathbf{V}_i^{d+} - \mathbf{V}_i^{d-}$ and consists of four feature planes. The same yields for the reconstruction $\mathbf{R}_i^{ds}$ and the basis vectors $\mathbf{W}_j^{ds}$.[1] We now introduce two energy functions based on the translation invariant NMF and the direction specific feature planes. The first is

$$E_R^{(1)} = \frac{1}{2} \sum_{i,d,s} \left(\|\mathbf{V}_i^{ds} - \sum_{j,m} H_{ij}^{(m)} T^{(m)} \bar{\mathbf{W}}_j^{ds}\|_2^2 \right) \tag{5}$$

which implies that there is no overlap between the feature planes. We call VNMF1 the algorithm that minimizes $E = E_R^{(1)} + \lambda_H E_H$. VNMF2 is the algorithm minimizing $E = E_R^{(2)} + \lambda_H E_H$ with

$$E_R^{(2)} = \frac{1}{2} \sum_{i,d} \|\mathbf{V}_i^{d+} - \mathbf{V}_i^{d-} - \sum_{j,m} H_{ij}^{(m)} T^{(m)} (\bar{\mathbf{W}}_j^{d+} - \bar{\mathbf{W}}_j^{d-})\|_2^2. \tag{6}$$

For eqs. (5) and (6) we additionally have introduced *normalized* basis vectors $\bar{\mathbf{W}}_j^{ds}$, which following the derivations from [11], lead to a modification of the gradient-based update equations from eq. (2). Here it is important to note that each basis vector has to be normalized over all feature planes, not each plane for itself.

The VNMF2 eq. (6) corresponds to the multiplane version of translation invariant SC eq. (3) extended by an inherent basis vector normalization, with two important additional points to consider. First, the features $\mathbf{W}_j^{d+}$ and $\mathbf{W}_j^{d-}$ are initialized and updated separately. Second, the gradients $\nabla_{\mathbf{W}_j^{ds}} E$ can be split up into $(\nabla_{\mathbf{W}_j^{ds}} E)^{(pos)}$ and $(\nabla_{\mathbf{W}_j^{ds}} E)^{(neg)}$ which allows us to use the update rules from eq. (2).

[1] However we remark that only one common activity $H_{ij}^{(m)}$ is used for all four planes.

2.3 Orthogonality between the Reconstructions

To enforce a more parts-based decomposition we introduce a new energy term

$$E_p = \lambda_p \sum_{i,d,s} \Big(\|\mathbf{R}_i^{ds}\|_2^2 - \sum_{j,m} \|H_{ij}^{(m)} (T^{(m)} \bar{\mathbf{W}}_j^{ds})\|_2^2 \Big). \tag{7}$$

With E_p we penalise the overlap between the partial reconstructions and therefore enforce a parts-based characteristic of our decomposition. Unlike orthogonality between the basis vectors as used by Choi [13], this energy function is driven by the activities. If an activity is present, the constraint on the partial reconstructions suppresses other activities in the spatial surrounding outlined by the corresponding basis vector. This leads to even sparser activity patterns.

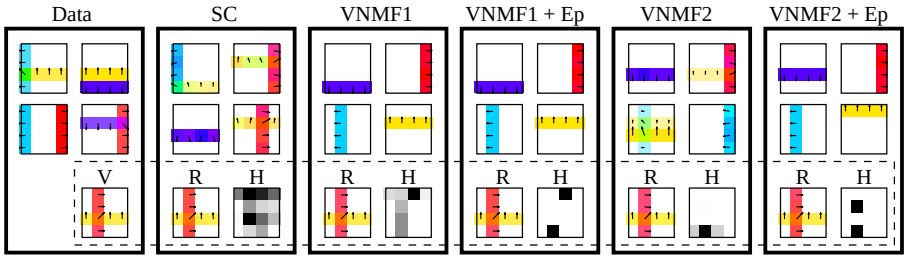

Fig. 1: Example data and extracted basis vectors. The lower row shows the reconstruction R and the superposition of all activities H for one input V.

3 Comparison of the Methods

In the following, we will apply different basis vector decomposition methods on motion data: The translation invariant version of SC eq. (3), VNMF1 eq. (5), VNMF2 eq. (6) and VNMF1 resp. VNMF2 with the new parts-based constraint eq. (7). We use an artificial dataset consisting of 100 vector fields, where each is a superposition of two out of four basic horizontal or vertical bars, with motion vectors into positive or negative orthogonal directions respectively, that are randomly choosen and shifted (see Fig. 1 for example vector fields). To make the effects of the energy functions comparable we used a gradient descent with fixed step size for all experiments[2]. We ran each decomposition until the mean reconstruction error per pixel reached a value of 0.01. Fig. 1 shows the four resulting basis vectors as extracted by each algorithm (top) as well as an example reconstruction along with the superposition of its corresponding activities (bottom). While all methods were able to reconstruct the input, neither SC nor VNMF2 were able to retrieve the original sources. Either the non-negativity constraint due to the direction specific energy function of VNMF1 or the orthogonality constraint eq. (6) on the reconstruction is required.

[2]Parameters: $\lambda_H = 0.1$, $\lambda_p = 0.5$, Stepsize: $\Delta_H = 0.1$, $\Delta_W = 0.01$.

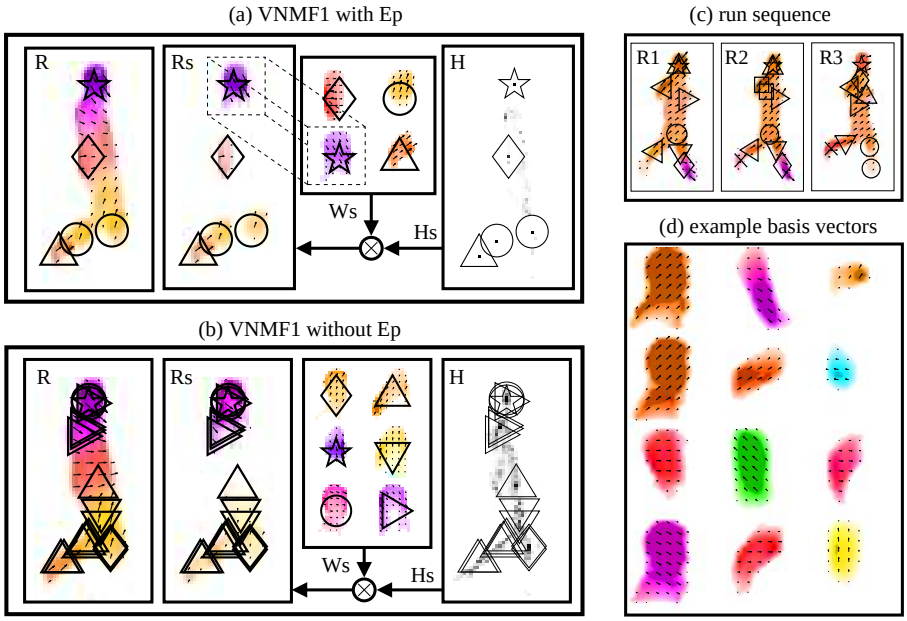

Fig. 2: (a) Decomposition of a single walking frame with VNMF1+E_p. H displays the superposition of all activities in one image. The marked activities H_s fulfill $H_s > 0.2 * \max_{j,m}(H_{ij}^{(m)})$. (b) The decomposition of the same frame, but without the parts-based term E_p. (c) Reconstructions of three consecutive frames with marked H_s. (d) Enlarged view of a subset of the basis vectors $\mathbf{W}_j$ learned with VNMF1+E_p.

4 Extraction of sharply localized Motion Components

From Fig. 1 it can be seen that, even though the reconstructions as well as the basis vectors are retrieved by both VNMF1 and VNMF1+E_p, VNMF1 exhibits spurious activity traces. That is, the activity is sparse but not sharply localized, whereas for VNMF1+E_p, we get almost binary activities which are better suited for parts-based representations. To confirm these observations for more realistic data, we tested the VNMF1 and VNMF1+E_p algorithm with the NMF update rules of eq. (2) on the Weizmann human action recognition dataset [12] which consists of videos with 9 persons performing 10 different actions, like walking, running, waving, jumping, etc. As a preprocessing step, we estimated the optical flow field using the algorithm of Sun et al. [9]. A subset of the learned basis vectors and results for a typical parts-based decomposition of a motion frame are shown in Fig. 2.

The dominant activities H_s in Fig. 2(a) are localized on different body parts, in particular the head (★), the arm (◊), the upper legs (◯) and a foot (△).

The corresponding basis vectors represent parts of the flow field, and thus are shaped like body parts. The effect of the parts-based energy term E_p eq. (7) can be seen by comparing Figs. 2(a) and (b). While the reconstruction is very similar, the basis vectors and especially the activities differ. Fig. 2(a) shows a smaller number of dominant, but sharply localized activities, which due to the E_p term suppress other activities in their surrounding, whereas the activities in Fig. 2(b) are much more distributed.

5 Summary

We have shown that the shape as well as the motion of rigid body parts can be extracted by an extension of the NMF class of algorithms, which includes sparseness, translation invariance and parts-basedness. The gained activities result in a sparse representation similar to point light stimuli, but which encode localized moving body parts, as shown in Fig. 2(d).

References

[1] G. Johansson, Visual perception of biological motion and a model for its analysis, *Percept. Psychophys.*, vol. 14, pp. 201-211. 1973.

[2] J. Haag and A. Borst, Neural mechanism underlying complex receptive field properties of motion-sensitive interneurons, *Nature Neuroscience*, vol. 7, pp. 628-634, 2004.

[3] J. Lange and M. Lappe, A model of biological motion perception from configural form cues, *The Journal of Neuroscience*, 26(11) pp. 2894-2906, 2006.

[4] B. Olshausen and D.J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, *Nature*, vol. 381, pp. 607-609, 1996.

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

[6] D.J. Fleet, M.J. Black, Y. Yacoob and A.D. Jepson, Design and use of linear models for image motion analysis, *Int. J. of Computer Vision*, vol. 36, no. 3, pp. 171-193, 2000.

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

[8] C. Ding, T. Li, M.I. Jordan, Convex and Semi-Nonnegativ Matrix Factorizations, *Pattern Analysis and Machine Intelligence* (PAMI), vol. 32 pp. 45-55, 2010.

[9] 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.

[10] T. Dean, R. Washington, G. Corrado, Recursive Sparse, Spatiotemporal Coding, *IEEE Int. Symposium on Multimedia* (ISM), pp. 645-650, 2009.

[11] J. Eggert, H. Wersing and E. Koerner, Transformation-invariant representation and NMF, *IEEE Int. Joint Conf. on Neural Networks* (IJCNN), pp. 2535-2539 vol. 4, 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), pp. 1395-1402, 2005.

[13] S. Choi, Algorithms for orthogonal nonnegative matrix factorization, *IEEE World Congress on Computational Intelligence* (IJCNN), pp. 1828-1832, 2008.