LETTER Communicated by Marcus Lappe

Topological Sparse Learning of Dynamic Form Patterns

T. Guthier
tguthier@rtr.tu-darmstadt.de
V. Willert
vwillert@rtr.tu-darmstadt.de
TU Darmstadt, 64283 Darmstadt, Germany

J. Eggert
Julian.Eggert@honda-ri.de
Honda Research Institute Europe, 63073 Offenbach/Main, Germany

Motion is a crucial source of information for a variety of tasks in social interactions. The process of how humans recognize complex articulated movements such as gestures or face expressions remains largely unclear. There is an ongoing discussion if and how explicit low-level motion information, such as optical flow, is involved in the recognition process. Motivated by this discussion, we introduce a computational model that classifies the spatial configuration of gradient and optical flow patterns. The patterns are learned with an unsupervised learning algorithm based on translation-invariant nonnegative sparse coding called VNMF that extracts prototypical optical flow patterns shaped, for example, as moving heads or limb parts. A key element of the proposed system is a lateral inhibition term that suppresses activations of competing patterns in the learning process, leading to a low number of dominant and topological sparse activations. We analyze the classification performance of the gradient and optical flow patterns on three real-world human action recognition and one face expression recognition data set. The results indicate that the recognition of human actions can be achieved by gradient patterns alone, but adding optical flow patterns increases the classification performance. The combined patterns outperform other biological-inspired models and are competitive with current computer vision approaches.

1 Introduction

The capability of recognizing complex motions such as gestures, human actions, and face movements is crucial for social interactions, predators, prey, or artificial systems interacting in a dynamic environment. The famous point-light-walker experiments (Johansson, 1973) reveal that humans have a highly skilled mechanism dedicated to the analysis of motion information; however, the exact details of this mechanism remain largely unclear.

Neural Computation **27, 42–73** (2014)
doi:10.1162/NECO_a_00670

The point-light-walker experiments show that humans can recognize biological motion even without explicit form information.[1] These observations started an ongoing discussion on how form and motion contribute to the recognition process. While neurophysiological experiments, discussed in Chouchourelou, Golden, Shiffrar, and Chouchourelou (2012), Grossman, Jardine, and Pyles (2010), Grossman and Blake (2002), Giese and Poggio (2003), Theusner, de Lussanet, and Lappe (2014), and Giese (2014), indicate the importance of both form and motion information, there are several open questions, for example, the role of explict low-level motion information such as optical flow (Willert, Toussaint, Eggert, & Korner, 2007; Willert & Eggert, 2009; Sun, Roth, & Black, 2010; Guthier, Willert, Schnall, Kreuter, & Eggert, 2013). Optical flow estimation itself is not selective to form, but by grouping parts with consistent movements, like an upper arm or a torso, the spatial configuration and the movement direction of these parts can be used to identify characteristic motion patterns. Early computational models propose the use of optical flow patterns, for example, in a hierarchical manner (Giese & Poggio, 2003; Jhuang, Serre, Wolf, & Poggio, 2007). To the contrary, motivated by lesion experiments of a patient whose early motion processing areas were impaired but who could nevertheless recognize biological motion Lange and Lappe (2006) and Theusner et al. (2014) suggest that low-level motion plays no major role in the recognition of biological motion. In their related proposed model, motion is incorporated only on a higher level, as the transition between full body poses. Their model is in good accordance with neurophysiological experiments that indicate that early motion processing areas of the brain may not be involved in biological motion perception (Beintema & Lappe, 2002; Servos, Osu, Santi, & Kawato, 2002). However, low-level motion information improves the recognition in the presence of noise (Beintema & Lappe, 2002), which hints of an involvement of low-level motion. As in Lange and Lappe (2006) suggested optical flow could be used to segment the moving person from the background, or, as discussed in Giese (2014), the spatial configuration of midscale optical flow a patterns could be used as a way to describe the human body form alongside static shape or gradient information. Thus, body postures can be defined by the spatial configuration of two, possibly redundant, types of information: static or dynamic (e.g., gradient and optical flow patterns).

Similar to Giese and Poggio (2003), Lange and Lappe (2006), Jhuang et al. (2007), Theusner et al. (2014), and Fleischer, Caggiano, Thier, and Giese (2013), we contribute to the ongoing discussion on a functional level by presenting a computational model that consists of two streams—one for

[1]The term *biological motion* was introduced by Johansson referring to the point-light-walker experiments. Here we use the term to refer to any form of articulated movement from humans, animals, or artificial systems, in contrast to ego-motion-induced global optical flow fields (Pitzalis et al., 2013). The term is further discussed in Chouchourelou et al. (2012).

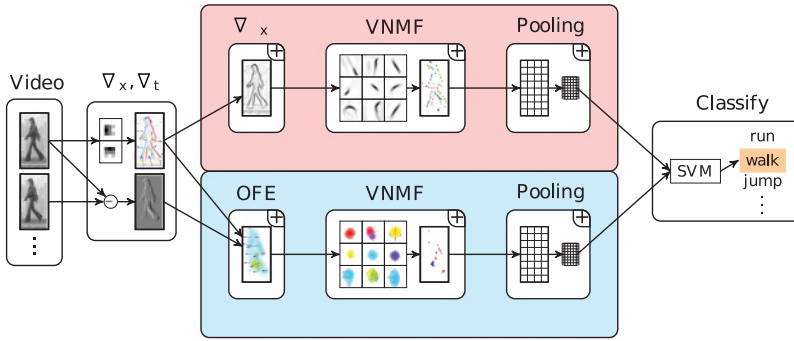

Figure 1: Stages of the hierarchical two-stream biological motion recognition model. First, the spatial and temporal gradients are calculated. In the static stream (red), the spatial gradient amplitudes are projected onto prelearned gradient patterns and subsequently pooled. In the dynamic stream (blue), the spatiotemporal gradients are used to estimate a dense optical flow field (OFE), which is then projected onto prelearned optical flow patterns and subsequently pooled. In the final step, the combined pooled activations are classified using a SVM. The main difference from other models (Giese & Poggio, 2003; Jhuang et al., 2007) is the VNMF algorithm, which we use to learn the midlevel patterns. Boxes marked with a (+) consist of strictly nonnegative components.

dynamic and one for static form information (see Figure 1). We examine the classification performance of the individual and combined streams in complex real-world scenarios to analyze if low-level motion (i.e., optical flow fields), can contribute to the recognition of human actions. Our main contribution is the learning of novel midlevel motion patterns learned with VNMF, a direction-selective, translation-invariant nonnegative sparse coding algorithm (Olshausen & Field, 1997; Lee & Seung, 1999; Hoyer, 2002; Eggert, Wersing, & Koerner, 2004; Eggert & Koerner, 2004; Guthier, Eggert, & Willert, 2012) that includes a lateral competition that enforces topological sparseness. VNMF groups form-invariant motion information into patterns of coherent movement. These motion patterns encode dynamic information like movement direction alongside form information, defined by the shapes of the patterns that resemble body parts. We thus term these patterns dynamic form patterns in contrast to static form patterns that are based on spatial gradients.

We evaluate the model on four real-world computer vision benchmarks for human action (Schuldt, Laptev, & Caputo, 2004; Blank, Gorelick, Shechtman, Irani, & Basri, 2005; Rodriguez, Ahmed, & Shah, 2008) and face expression recognition (Dollar, Rabaud, Cottrell, & Belongie, 2005). We analyze the classification results for the dynamic and the static form patterns learned with the VNMF in comparison to other learning methods as well as

state-of-the-art HOG/HOF descriptors (Dalal, Triggs, & Schmid, 2006). We further show that the overall model is competitive with computer vision approaches (Guha & Ward, 2012; Amiri, Nasiopoulos, & Leung, 2012), while outperforming other biologically inspired computational models (Jhuang et al., 2007; Dean, Washington, & Corrado, 2010).

The outline of the letter is as follows. First, we review the neurophysiological and psychophysical discussion on motion processing in the brain, with a focus on local motion patterns that are relevant for tasks like human action recognition; we then introduce our biological motion recognition model. Next, we discuss desired properties for a pattern learning algorithm and introduce the translation-invariant nonnegative sparse coding algorithm VNMF. We show decomposition results on real-world data and compare the extracted patterns with those extracted with PCA. Next, we present results on human action recognition benchmarks (Weizmann—Blank et al., 2005), KTH (Schuldt et al., 2004), UCF-Sports (Rodriguez et al., 2008) as well as for a face expression recognition data set (Dollar et al., 2005) for different parameter settings. Finally, we briefly discuss the experimental results.

2 Motion Processing Model

The architecture of our biological-motivated motion recognition system is a multistage feed-forward neural network (FFNN) consisting of two parallel streams, related to the motion processing areas in the human visual cortex that we briefly discuss.

FFNNs were introduced as early as 1969 (Minsky and Papert, 1969) and have been continuously refined (Fukushima, 1980) to achieve translation-invariant visual object recognition with a hierarchy of processing stages. Throughout the hierarchy, the complexity of the encoded information, as well as the receptive field size of the simple cells (which represent basic image properties), increases, whereas the generality of the patterns decreases. In object recognition, the receptive fields of the first layer, which models the primary visual cortex (V1), encode basic information such as image gradients or Gabor filters (Olshausen & Field, 1997), while the final layers, which model the IT-complex, contain cell populations that mainly correspond to specific objects (DiCarlo, Zoccolan, & Rust, 2012) and can be regarded as classifiers. There is growing evidence from neurophysiological experiments that a similar feedforward architecture for the recognition of biological motion exists. Unlike static object recognition tasks, which are thought to be mainly located along the ventral stream of the visual cortex, multiple areas are involved in the recognition of biological motion, discussed in Grossman and Blake (2002), Puce and Perrett (2003), Pyles, Garcia, Hoffman, and Grossman (2007), Blake and Shiffrar (2007), and Chouchourelou et al. (2012).

2.1 Brain Areas Involved in Motion Processing. Selected neurons in the superior temporal sulcus (STS) respond to different kinds of biological

motion stimuli, such as human full body movements or facial expressions (Puce & Perrett, 2003; Pyles et al., 2007; Blake & Shiffrar, 2007; Saygin, 2007; Grossman et al., 2010; Chouchourelou et al., 2012) and thus might work as action classifiers. They show invariance to size and position variations (Grossman et al., 2010), while also containing some size- and orientation-specific neurons (Ashbridge, Perrett, Oram, & Jellema, 2000). Different parts of area STS (STSp) are connected to various visual areas, including the motion-sensitive dorsal stream as well as the form- and shape-sensitive ventral stream.

Early motion processing areas such as MT/MST (hMT+ complex) contain cells with receptive fields that pool the information of the V1 cells and to some extent can solve the correspondence problem (related to the aperture problem) that arises at small scales (Willert & Eggert, 2011). Based on neurophysiological experiments, the computational model in Giese and Poggio (2003) interpret the process in this early visual system as optical flow estimation (OFE) (Willert et al., 2007; Willert & Eggert, 2009; Sun et al., 2010; Guthier et al., 2013). However, other researchers (Beintema & Lappe, 2002; Servos et al., 2002) could not find a response of hMT+ to biological motion stimuly.

Areas in the ventral stream that are involved in biological motion recognition include the extrastriate body area (EBA), which is mostly activated by images of human bodies. In a recent study (Weiner & Grill-Spector, 2011), the EBA is devided into three limb-selective areas, including ITS and ITG, which is anatomically next to and, to some extent, even overlapping with the motion-selective hMT+ complex. Areas ITS and pITG play a major role in the experiments discussed in Pyles et al. (2007), where it is investigated whether STSp responses are limited to human-articulated motion or if they include so-called creature motion. Their creatures are artificially created random concatenations of limb-like constructs. In their experiments, they measured the fMRI responses of humans observing creature and human movements. The main result of their study is that STSp responded more strongly to human action than to creature action, an indicator for an STSp specialization to human movements. This is in good accordance with its connection to the motor system because the observing humans are not capable of performing the actions of the creatures. In addition, they found that area ITS and area pITG responded to the creature motion as well as to human biological motion, thus playing a role in a more general processing of articulated motion, that is not restricted to human movements. ITS and pITG respond to both creatures and humans, who share common limb forms but not a common global form.

While the role of parts of STSp as a recognition area for higher-level biological motion is well supported in the literature, it remains less clear how low-level motion contributes to biological motion recognition. Due to the anatomical overlap of parts of the EBA and the hMT+ complex, it could be speculated that the limb-selective areas in EBA are not solely

driven by static stimuli, but from optical flow from the hMT+ complex as well. Alternatively, as Servos et al. (2002) suggested on other brain areas, (e.g., the lingual gyrus) might be involved in biological motion recognition. Compared to hMT+, the lingual gyrus might be related to more specific aspects of motion processing, such as form-from-motion (Servos et al., 2002).

2.2 A Biologically Motivated Motion Recognition Model. To investigate whether low-level motion features could improve human action recogntion, we propose a two-stream hierarchical system as depicted in Figure 1. In the first stage, the spatial and temporal gradients of the incoming video are calculated. From there, the data are separated into a dynamic and a static stream. The dynamic stream consists of an optical flow estimation (OFE) step that provides a dense optical flow field as output. The second stage consists of a novel unsupervised nonnegative sparse coding algorithm (VNMF), which is used to learn dynamic form patterns. We keep the direction-selective representation of the early motion estimation stage to retain the nonnegativity properties for the decomposition of the vector fields. For the static stream, the amplitude of the spatial gradients is calculated to achieve a nonnegative representation that is invariant to color changes (bright to dark or dark to bright). Then the VNMF algorithm is used to learn static form patterns. Similar to other hierarchical vision systems (Fukushima, 1980; Wersing & Koerner, 2003; LeCun, Huang, & Bottou, 2004), the neural activations of the learned patterns are subsequently pooled to achieve an invariance against local shifts. The final stage consists of a set of *support vector machines* (SVMs) that are used to classify the different human actions.

3 Unsupervised Pattern Learning

Unsupervised learning algorithms such as PCA, ICA, SC, and NMF are widely applied to gather natural image statistics by learning basic patterns (also referred to as basis vectors). The underlying model assumption is that there exists a limited set of basis vectors whose superposition can reconstruct a given input. There is a broad range of methods for dictionary learning that can be applied to learn the set of basis vectors, such as PCA, ICA, sparse coding (SC) (Olshausen & Field, 1997), nonnegative matrix factorization (NMF) (Lee & Seung, 1999) and several more. For an overview on the methods we refer to Cichocki, Zdunek, Phan, and Amari (2009) and Hyvarinen, Hurri, and Hoyer (2009). Choosing the right learning algorithm is crucial and depends on the desired characteristics for the basis vectors.

There are essentially two counteracting characteristics to consider. The basis vectors restrict the information flow toward higher processing areas; thus, incoming motion information that they cannot represent is discarded in further processing steps. So on the one hand, they should be generic and thus capable of describing the entire natural input space. On the other hand,

the basis vectors should be class discriminative to provide features for a subsequent classification. Parts-based representations are likely candidates for such a set of basis vectors because on a local scale, the movements of body parts consist of coherent motion patterns.

Classical models (Giese & Poggio, 2003; Lange & Lappe, 2006; Jhuang et al., 2007) make use of handcrafted or template patterns rather than learned dictionaries. Others (Fleet, Black, Yacoob, & Jepson, 2000; Casile & Giese, 2005) apply PCA to optical flow fields to learn basis vectors. PCA is known to give rather holistic representations that are good for image compression but not for classification tasks. In more recent work (Cadieu & Olshausen, 2012), motion and shape patterns are learned simultaneously on videos of animal movements. They use a SC algorithm and model the motion information via the phase of a complex-valued representation of the input images. SC applied to natural images is known to extract Gabor filter-like patterns that yield similar responses as patterns found in V1 (Olshausen & Field, 1997). Under the assumption that the cell populations of the visual cortex have a common coding mechanism based on sparsity, it is consistent to use SC for the extraction of motion patterns as well. However, the use of a complex-valued representation as well as the use of negative values is somehow contrary to the nonnegative nature of neural activities. Dean et al. (2010) propose a different representation for motion information by using space-time-volumes around predetected interest points. They apply a SC algorithm and use the activity responses for human action recognition.

Analogous to Fleet et al. (2000), Giese and Poggio (2003), and Casile and Giese (2005), we chose dense optical flow fields to represent the visual motion information. Dense optical flow fields describe the movement of every pixel between two consecutive frames. Similar to Efros, Berg, Mori, and Malik (2003) we further decompose the motion vector field into the four nonnegative directions: right, up, left and down. The nonnegative representation is consistent with an encoding concept based on neural activity, while direction selectiveness is a property found throughout the human visual motion processing areas. Furthermore, a nonnegative encoding is known to lead to parts-based decompositions (Lee & Seung, 1999; Hoyer, 2002; Eggert & Koerner, 2004; Guthier et al., 2012).

The core of our learning algorithm is thus based on nonnegative sparse coding (Hoyer, 2002; Eggert & Koerner, 2004) where as few basis vectors as necessary are used to reconstruct the input, favoring a parts-based representation. The result is a jigsaw puzzle–like decomposition, where for each part of the input, there is only one active basis vector. To explicitly achieve this topological sparsity, a competition between overlapping receptive fields is needed—a competition between the set of possible basis vectors (local competition) as well as between spatial neighboring activities (lateral competition).

Current sparse coding approaches do not address this superposition problem (overlapping receptive fields) directly for two reasons: First, the

sparsity penalty functions based on Olshausen and Field (1997) are global penalties that address the sum rather than the interaction between individual activities. Second, most sparse coding algorithms use only small extracted image patterns and lose the topological information, which then cannot be exploited during learning.

Our unsupervised learning algorithm is a translation-invariant sparse coding algorithm with nonnegativity constraints that is based on Olshausen and Field (1997), Lee and Seung (1999), and (Eggert et al., 2004) with an additional lateral inhibition term (Guthier et al., 2012). The results can be summarized as follows:

1. We use a translation-invariant learning algorithm that is able to place the basis vectors in the entire image and not only on randomly extracted image patches. This enables us to gather the statistics of the entire field of view and address topological effects such as the superposition problem, during learning.
2. A new term that enforces topological sparseness leads to activities that are sharply localized, without any nonlinear postprocessing. The learned basis vectors are parts based and combine the benefits of generative patterns and local templates.

In the following we introduce the concept of nonnegative sparse coding and explore how it is used in our hierarchical model. Then we discuss how translation-invariant learning can address the superposition problem and derive the equations for the translation invariant NMF and our lateral inhibition term.

3.1 Nonnegative Sparse Coding. Sparse coding (SC) (Olshausen & Field, 1997) is based on the idea that each input image $\mathbf{V}_i \in \mathbb{R}^{\mathbf{X}}$ out of a set of images $\mathcal{V} \in \mathbb{R}^{\mathbf{X}\times I}$, with $\mathbf{X} = X \cdot Y$ ($X, Y \mathrel{\widehat{=}}$ number of pixels in horizontal and vertical direction and $I \mathrel{\widehat{=}}$ number of input images) can be represented by a reconstruction $\mathbf{R}_i$ that is generated by a weighted sum of basis vectors $\mathcal{W} \in \mathbb{R}^{\mathbf{X}\times J}$ ($J \mathrel{\widehat{=}}$ number of basis vectors):

$$\mathbf{V}_i \simeq \mathbf{R}_i = \sum_j h_{ij}\mathbf{W}_j. \tag{3.1}$$

The goal of an SC algorithm is to learn the basis vectors $\mathbf{W}_j$ and activity weights h_{ij} that can reconstruct the given set of images $\mathcal{V}$ with as few activities as possible. SC combined with nonnegativity constraints (Hoyer, 2002; Eggert & Koerner, 2004) on $\mathbf{V}_i \geq 0, \mathbf{W}_j \geq 0$ and $h_{ij} \geq 0, \forall i \in I, j \in J$ leads to sparse nonnegative matrix factorization (sNMF). The learning is carried out by minimizing the energy function,

$$E = E_r + \lambda_H E_h = \frac{1}{2}\sum_i \left\| \mathbf{V}_i - \sum_j h_{ij}\bar{\mathbf{W}}_j \right\|_2^2 + \lambda_H \sum_{i,j} h_{ij}, \tag{3.2}$$

with respect to $\mathbf{W}_j$ and h_{ij} by iteratively updating the activities h_{ij} and the basis vectors $\mathbf{W}_j$ via gradient descent. E_r is the reconstruction energy and E_h the sparsity energy term. With the nonnegativity constraints, we can separate each gradient into its positive and negative gradient components $\nabla_A E = (\nabla_A E)^+ - (\nabla_A E)^-$ and apply the multiplicative NMF update rules proposed by Lee and Seung (1999; Seung & Lee, 2001). The sparsity condition for the activities can lead to an undesired scaling effect, where the amplitude of the reconstruction is represented only in the basis vectors and the activities converge to zero. It is therefore necessary to use normalized basis vectors, using, for example, the Euclidean norm $\bar{\mathbf{W}}_j(\mathbf{W}_j)$. The algorithm that minimizes the energy function in equation 3.2 is:

1. Randomly initialize $\mathbf{W}_j$ and h_{ij}, $\quad \forall j \in J, i \in I$.
2. Loop until a local minimum of E is reached:
   (a) Loop over all inputs $i \in I$:
      i. Calculate $\mathbf{R}_i$.
      ii. Calculate activity gradients $(\nabla_{h_{ij}} E_i)^+$, $(\nabla_{h_{ij}} E_i)^-$, $\quad \forall j \in J$.
      iii. Update $h_{ij} = h_{ij} \cdot \frac{(\nabla_{h_{ij}} E_i)^-}{(\nabla_{h_{ij}} E_i)^+}$, $\quad \forall j \in J$.
      iv. Calculate $\mathbf{R}_i$.
      v. Calculate basis vector gradients $(\nabla_{\mathbf{W}_j} E_i)^+$, $(\nabla_{\mathbf{W}_j} E_i)^-$, $\quad \forall j \in J$.
   (b) Update $\mathbf{W}_j = \mathbf{W}_j \circ \frac{\sum_i (\nabla_{\mathbf{W}_j} E_i)^-}{\sum_i (\nabla_{\mathbf{W}_j} E_i)^+}$, $\quad \forall j \in J$.
   (c) Normalize the basis vectors $\bar{\mathbf{W}}_j = \frac{\mathbf{W}_j}{\|\mathbf{W}_j\|_2}$, $\quad \forall j \in J$.

$\mathbf{A} \circ \mathbf{B}$ describes the Hadamard (element wise) multiplication between the two matrices $\mathbf{A}$ and $\mathbf{B}$. The division between the gradient components is the Hadamard division. We consider the inner derivate of $\bar{\mathbf{W}}_j(\mathbf{W}_j) = \frac{\mathbf{W}_j}{\|\mathbf{W}_j\|_2}$ during the calculation of the gradient $\nabla_{\mathbf{W}_j} E_i$ as proposed in Eggert and Koerner (2004), so that the final normalization step in each iteration can be done without altering the energy function.

3.2 The Superposition Problem in Feedforward Neural Networks. To motivate our translation-invariant learning we now discuss the superposition problem. By this, we refer to the simultaneous activation of basis vectors with overlapping receptive fields, which leads to blurry activation patterns that counteract the idea of sparse activations. There already exist effective methods on how to deal with the superposition problem via nonlinear pooling for FFNN, but it is usually ignored during the learning of the simple cell patterns.

Each layer of an FFNN of the neocognitron type consists of two stages: the simple cell activations and the complex cell responses. The first part are

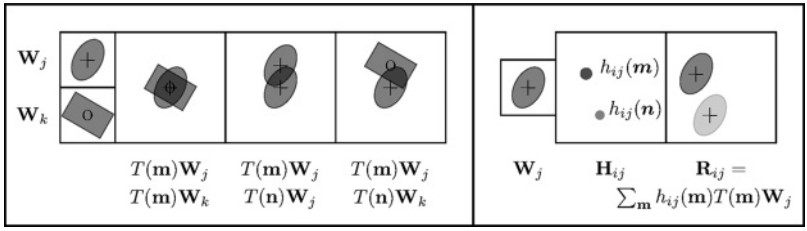

Figure 2: (Left) The superposition problem, which appears when we have overlapping receptive fields for identical or different types of basis vectors. There are three cases for overlapping receptive fields. First is for two different basis vectors $\mathbf{W}_j$ and $\mathbf{W}_k$ on the same center position $\mathbf{m}$, hence when the basis vectors are virtually shifted by the same shift matrix $T(\mathbf{m})$. The second case is when different shifted versions of the same basis vector $\mathbf{W}_j$ overlap. The third case is when shifted versions of different basis vectors overlap. All three cases are penalized by the topological sparsity energy functional E_p in equation 3.11. (Right) The basic principle of translation-invariant learning. For the partial reconstructions $\mathbf{R}_{ij}$, the corresponding basis vector $\mathbf{W}_j$ can be shifted to any position $\mathbf{m}$ in the entire image so that it has its own activation $h_{ij}(\mathbf{m})$. Instead of an activity vector $\mathbf{H}_i$ with scalar values h_{ij} for each basis vector $\mathbf{W}_j$ and input, we have a set of activation images $\mathbf{H}_{ij}$ with scalar activities $h_{ij}(\mathbf{m})$.

in our case the activity responses $\mathbf{H}_j = \text{corr}_2(\mathbf{V}, \mathbf{W}_j)$ of the corresponding basis vectors $\mathbf{W}_j$ to the presented stimulus $\mathbf{V}$ for all $j \in [1, J]$ throughout the entire input space. $\sum_{\mathbf{m}} a(\mathbf{m})b(\mathbf{x}+\mathbf{m}) = \text{corr}_2(\mathbf{A}, \mathbf{B})(\mathbf{x})$ describes the two-dimensional correlation between the two matrices A and B. Here we encounter the superposition problem: because slightly shifted receptive fields of the same or similar basis vectors are overlapping (see Figure 2), we get blurry activity responses.

That is why the simple cells are followed by the complex cells, a nonlinear projection $\tilde{\mathbf{H}}_j = f(\mathcal{H})$ of the activity responses $\mathcal{H} = [\mathbf{H}_1, ..., \mathbf{H}_J]$ with a local and a lateral component. The local competition is between different activities $\mathbf{H}_j$ at one position $\mathbf{x}$, for example, a norm-, a maximum- or a winner-takes-most (Wersing & Koerner, 2003) operator. The lateral competition is achieved via a max-pooling step. Neighboring activities are projected onto a single activity, which leads to activity images with a reduced resolution ($\dim(\mathbf{H}_j) > \dim(\tilde{\mathbf{H}}_j)$). Besides the lateral competition, the pooling has the additional effect of increasing the receptive field size and introducing an invariance to small shifts in the input space.

In summary, the goal of the nonlinear postprocessing on the activities is to arrive at a representation with sparser activities as well as a larger translation invariance and larger receptive fields throughout the hierarchy. A major drawback of the approach is that this nonlinear postprocessing

is not consistent with the learning process, since it is applied only during the detection phase. By this, we mean that the learned basis vectors do not correspond to a topologically sparse decomposition and may not be best suited for the complex cell type sparsifications.

While there are sparse coding algorithms that incorporate local competition into the learning procedure (Li, Hou, Zhang, & Cheng, 2001; Rozell, Johnson, Baraniuk, & Olshausen, 2008), the lateral competition is not addressed. One reason is that the basis vectors are learned on randomly sampled image patches. Due to the sampling process, the neighboring dependencies get lost and cannot be addressed during the learning process. We overcome this problem by using a translation-invariant learning procedure with an additional local and lateral competition penalty function, which we introduce in the next two sections.

3.3 Translation-Invariant Learning. In the translation-invariant approach, each basis vector can be positioned at all pixel positions of the entire image, so instead of reconstructing isolated image patches, the entire input image is reconstructed at once. Shifted local patterns with an identical form can thus be represented by a single basis vector, which helps to eliminate redundant basis vectors. Because each basis vector is shifted throughout the image for the reconstruction, it is also updated depending on its correlations at each shifted pixel position. Thus, the statistics related to each basis vector are gathered throughout the entire image, which reduces the amount of required input data significantly. Furthermore, we can incorporate a lateral competition directly into our learning process, because the receptive fields of each basis vector are used to reconstruct the combined image with overlapping patches and not individual image patches of limited size.

We now introduce the translation-invariant NMF as a special case of the transformation-invariant NMF (Eggert et al., 2004). The reconstruction r_i of an image i (at the two-dimensional pixel coordinate $\mathbf{x}$) is

$$r_i(\mathbf{x}) = \sum_{j,\mathbf{m}} r_{ij\mathbf{m}}(\mathbf{x}) = \sum_{j,\mathbf{m}} h_{ij}(\mathbf{m})\big(T(\mathbf{m})\bar{w}_j(\mathbf{x})\big). \tag{3.3}$$

The set of matrices $T(\mathbf{m})$, $\mathbf{m} \in [\mathbf{1}, \mathbf{X}]$ describes a set of shift operations that are applied to the basis vectors $\bar{\mathbf{W}}_j$.[2] With

$$\big(T(\mathbf{m})\bar{w}_j(\mathbf{x})\big) = \bar{w}_j(\mathbf{x} - \mathbf{m}), \tag{3.4}$$

we get the pixel value of the jth basis vector at the two-dimensional position $(\mathbf{x} - \mathbf{m})$ and shift it by $\mathbf{m}$ to reconstruct the pixel $\mathbf{x}$. $h_{ij}(\mathbf{m})$ is the

[2]The lowercase letters denote scalar elements of the correspoding vectors (e.g., $\bar{w}_j(\mathbf{x})$ is an element of $\bar{\mathbf{W}}_j$).

corresponding activity. The process is illustrated in Figure 2. If we combine equations 3.3 and 3.4 and the two-dimensional convolution $\sum_{\mathbf{m}} a(\mathbf{m})b(\mathbf{x} - \mathbf{m}) = \text{conv}_2(\mathbf{A}, \mathbf{B})(\mathbf{x})$, we get the reconstruction for each pixel $r_i(\mathbf{x})$ and the image reconstruction $\mathbf{R}_i$:

$$r_i(\mathbf{x}) = \sum_{j,\mathbf{m}} h_{ij}(\mathbf{m})\bar{w}_j(\mathbf{x} - \mathbf{m}) = \sum_j \text{conv}_2(\mathbf{H}_{ij}, \bar{\mathbf{W}}_j)(\mathbf{x}), \tag{3.5}$$

$$\mathbf{R}_i = \sum_j \text{conv}_2(\mathbf{H}_{ij}, \mathbf{W}_j). \tag{3.6}$$

In the standard sNMF case from section 3.1, the activity h_{ij} describes the scalar weight of the basis vector $\bar{\mathbf{W}}_j$ for the reconstruction $\mathbf{R}_i$. Now the activity $\mathbf{H}_{ij}$ is a vector with the scalar weight entries $h_{ij}(\mathbf{m})$ for each possible shift $\mathbf{m}$ of the basis vector $\bar{\mathbf{W}}_j$. The dimension of $\mathbf{H}_{ij}$ is equal to the dimension of the input images $\mathbf{V}_i$ and reconstructions $\mathbf{R}_i$. We choose the anchor of the convolution in such a way that the center of the shifted basis vector $T(\mathbf{m})\bar{\mathbf{W}}_j$ is located at the position $\mathbf{m}$ of the corresponding activity $\mathbf{h}_{ij}(\mathbf{m})$. Notice that for the reconstruction, we can now shift each basis vector to each pixel position. Therefore, we can allow restrictions for our basis vectors by setting a maximum receptive field size (mRFS).

The new reconstruction in equation 3.6 leads to the following reconstruction energy term,

$$E_r = \frac{1}{2}\sum_i \|\mathbf{V}_i - \mathbf{R}_i\|_2^2 = \frac{1}{2}\sum_i \|\mathbf{V}_i - \sum_j \text{conv}_2(\mathbf{H}_{ij}, \bar{\mathbf{W}}_j)\|_2^2, \tag{3.7}$$

with the gradients for the activities

$$\nabla_{\mathbf{H}_{ij}} E_r = \underbrace{\text{corr}_2(\mathbf{R}_i, \bar{\mathbf{W}}_j)}_{:=\left(\nabla_{\mathbf{H}_{ij}} E_r\right)^+} - \underbrace{\text{corr}_2(\mathbf{V}_i, \bar{\mathbf{W}}_j)}_{:=\left(\nabla_{\mathbf{H}_{ij}} E_r\right)^-} \tag{3.8}$$

and the gradients for the basis vectors

$$\nabla_{\bar{\mathbf{W}}_j} E_r = \underbrace{\sum_i \text{corr}_2(\mathbf{R}_i, \mathbf{H}_{ij})}_{:=\left(\nabla_{\bar{\mathbf{W}}_j} E_r\right)^+} - \underbrace{\sum_i \text{corr}_2(\mathbf{V}_i, \mathbf{H}_{ij})}_{:=\left(\nabla_{\bar{\mathbf{W}}_j} E_r\right)^-}. \tag{3.9}$$

The derivation of the gradients is in appendix A.

3.4 Topological Sparsity by Local and Lateral Inhibition. To achieve a topological sparse reconstruction, we now add a local and a lateral competition to the learning algorithm. The basic idea behind the local and lateral competition is to penalize the activation of overlapping receptive fields as depicted in Figure 2. Thus, we want to inhibit overlaps of the partial reconstructions $\mathbf{R}_{ij\mathbf{m}}$ for different basis vectors at the same position and shifted versions of the same and different basis vectors overlapping in their neighborhood. Each of the three cases depicted in Figure 2 can be addressed by a corresponding orthogonality function. The only overlap between the partial reconstruction $\mathbf{R}_{ij\mathbf{m}}$ and the set of all possible partial reconstructions $\mathbf{R}_{ik\mathbf{n}}, \forall k \in [1, ..., J], \forall \mathbf{n} \in [\mathbf{0}, ..., \mathbf{X}]$ that we do not want to penalize is the overlap of each partial reconstruction with itself ($\mathbf{R}_{ij\mathbf{m}}^\top \mathbf{R}_{ij\mathbf{m}}$). This leads us to the following penalty term for the partial reconstruction $\mathbf{R}_{ij\mathbf{m}}$:

$$\mathbf{R}_{ij\mathbf{m}}^\top \left(\sum_{k \neq j} \mathbf{R}_{ik\mathbf{m}} + \sum_{\mathbf{n} \neq \mathbf{m}} \mathbf{R}_{ij\mathbf{n}} + \sum_{k \neq j, \mathbf{n} \neq \mathbf{m}} \mathbf{R}_{ik\mathbf{n}} \right) = \mathbf{R}_{ij\mathbf{m}}^\top \left(\sum_{k,\mathbf{n}} \mathbf{R}_{ik\mathbf{n}} - \mathbf{R}_{ij\mathbf{m}} \right). \tag{3.10}$$

With $\sum_{k,\mathbf{n}} \mathbf{R}_{ik\mathbf{n}} = \mathbf{R}_i$, the new energy function for local and lateral competition becomes

$$E_p = \frac{1}{2} \sum_{i,j,\mathbf{m}} \mathbf{R}_{ij\mathbf{m}}^\top \left(\mathbf{R}_i - \mathbf{R}_{ij\mathbf{m}} \right) = \underbrace{\frac{1}{2} \sum_i \mathbf{R}_i^\top \mathbf{R}_i}_{:=E_{p1}} - \underbrace{\frac{1}{2} \sum_{i,j,\mathbf{m}} \mathbf{R}_{ij\mathbf{m}}^\top \mathbf{R}_{ij\mathbf{m}}}_{:=E_{p2}}, \tag{3.11}$$

with the gradients for the activities

$$\nabla_{\mathbf{H}_{ij}} E_p = \nabla_{\mathbf{H}_{ij}} E_{p1} - \nabla_{\mathbf{H}_{ij}} E_{p2} = \mathrm{corr}_2(\mathbf{R}_i, \tilde{\mathbf{W}}_j) - \mathbf{H}_{ij} \tilde{\mathbf{W}}_j^\top \tilde{\mathbf{W}}_j \tag{3.12}$$

and the gradients for the basis vectors

$$\nabla_{\tilde{\mathbf{W}}_j} E_p = \nabla_{\tilde{\mathbf{W}}_j} E_{p1} - \nabla_{\tilde{\mathbf{W}}_j} E_{p2} = \sum_i \mathrm{corr}_2(\mathbf{R}_i, \mathbf{H}_{ij}) - \tilde{\mathbf{W}}_j \sum_i \mathbf{H}_{ij}^\top \mathbf{H}_{ij}. \tag{3.13}$$

The detailed derivation of the gradients is in appendix B. The gradients for the first part of the competition energy term E_{p1}, equations 3.12 and 3.13, are identical with the positive components of the gradients of the translation invariant reconstruction term, equations 3.8 and 3.9, and therefore do not need to be computed again for the energy function E_p. Thus, the gradients of the competition term come with negligible additional computational costs.

3.5 Relations to Orthogonal-NMF. A common extension of the original NMF algorithm is to enforce orthogonality on the basis vectors (Li et al., 2001; Choi, 2008) with additional constraints for the optimization (Choi, 2008) or with the additional energy function (proposed in Li et al., 2001):

$$E_w = \lambda_w(E_{w1} + E_{w2}) = \lambda_w \left(\sum_j \bar{\mathbf{W}}_j^\top \sum_{k \neq j} \bar{\mathbf{W}}_k + \sum_j \bar{\mathbf{W}}_j^\top \bar{\mathbf{W}}_j \right). \tag{3.14}$$

The learned basis vectors are parts based and more discriminative in face detection tasks (Li et al., 2001) and thus yield similar properties as the proposed lateral competition term E_p. Yet there are three general differences between the approaches. First, for the Euclidean distance in the error function, the influence on the gradients for the E_w term does not scale with the number of input images and the occurrence of each basis vector in the data. Contrariwise, the reconstruction error depends on the number of input images and the activations, which are directly proportional to the occurrence of the basis vectors. The E_p term is based on the reconstruction and thus scales consistently with the reconstruction energy E_r and is thus easier to parameterize than the E_w term. Second, the E_w term penalizes overlapping basis vectors and not the reconstructions. The consequence is that similar shapes, for example, heads of different individuals, will be represented only by a single dominant basis vector, since the similar shapes are not orthogonal. Third, the simple orthogonality measure cannot be integrated into a translation-invariant algorithm, because nonorthogonal basis vectors can simply shift their position inside their window and thus become orthogonal. The E_p is focused on the reconstruction and prevents such a behavior, because the relevant shift is encoded in the corresponding activations.

In summary, it can be said that the E_p term is more focused on topological sparsity of the activations and the parts-based nature of the corresponding basis vectors is a subsequent effect. Contrariwise, the E_w term is focused directly on the basis vectors and is therefore more restrictive.

Yet there is an interesting property of the self-similarity penalty term E_{w2}. Due to the Euclidean normalization $\bar{\mathbf{W}}_j(\mathbf{W}_j) = \frac{\mathbf{W}_j}{\|\mathbf{W}_j\|_2}$, spatially extended basis vectors $\bar{\mathbf{W}}_j$ have higher norm values $\|\mathbf{W}_j\|_2$ and thus lower element values than spatially sparse basis vectors. The gradient $\nabla_{\bar{\mathbf{W}}_j} E_{w2} = \lambda_w \bar{\mathbf{W}}_j$ is proportional to the element values. Consequently, E_{w2} favors spatially extended basis vectors because they have a lower penalty than sparse patterns, like the trivial solution with just one active element. We weight the energy functional with the activations so that it scales alongside the other energy functions E_r and E_p. The new orthogonality term is then

$$E_o = \lambda_o \sum_{i,j,\mathbf{m}} h_{ij}(\mathbf{m}) \bar{\mathbf{W}}_j^\top \bar{\mathbf{W}}_j, \tag{3.15}$$

and with $\bar{\mathbf{W}}_j^\top \bar{\mathbf{W}}_j = 1$ due to the Euclidean norm, the gradients are

$$\nabla_{\bar{\mathbf{W}}_j} E_o = 2\lambda_o \sum_{i,\mathbf{m}} h_{ij}(\mathbf{m}) \bar{\mathbf{W}}_j, \tag{3.16}$$

$$\nabla_{h_{ij}(\mathbf{m})} E_o = \lambda_o \bar{\mathbf{W}}_j^\top \bar{\mathbf{W}}_j = \lambda_o. \tag{3.17}$$

Since E_o favors a sparse representation identical to the linear sparsity term E_h in equation 3.2, and helps to avoid the trivial solution we choose E_o instead of E_h in our algorithm.[3]

3.6 Direction-Selective Motion Coding. Starting from dense optical flow fields, we now want to learn basis vectors that represent dynamic form patterns. Before we can apply our proposed learning algorithm, there is still one thing to consider. Optical flow fields $\mathbf{V}_i^d$, with $d \in \{x, y\}$, can have positive and negative values and therefore violate the nonnegativity constraint. One way of dealing with the negative values is to use the semi-NMF (Ding, Li, & Jordan, 2010), which requires nonnegativity constraints only on the activations. This approach has two drawbacks. The first is that without the explicit nonnegative representation, we lose the main benefit of nonnegative encoding because the reconstruction can consist of a superposition of positive and negative components. Second, a semi-NMF loses the neurophysiological plausibility of the non-negative encoding.

To keep the advantages and neural plausibility of the nonnegativity constraint on the activities and extracted cell populations for flow fields, we follow the proposal of direction-selective cell subpopulations found in the motion pathway. The idea is that different motion directions, such as, up, down, left, and right, are explicitly represented. The cell population of each direction then encodes whether the population is sensitive or insensitive for motion in this particular direction, which fulfills the nonnegativity constraints. We get the direction-selective motion representation by splitting the incoming optical flow fields into the four directions up, down, left, and right, which spawn a nonnegative, orthogonal basis from which each vector can be represented. Each pixel $\mathbf{x}$ of the two vector components x and y of $\mathbf{V}_i^d$ is split into a positive and a negative part:

$$(\mathbf{V}_i^{d+})_\mathbf{x} = \frac{\left|(\mathbf{V}_i^d)_\mathbf{x}\right| + (\mathbf{V}_i^d)_\mathbf{x}}{2}, \quad (\mathbf{V}_i^{d-})_\mathbf{x} = \frac{\left|(\mathbf{V}_i^d)_\mathbf{x}\right| - (\mathbf{V}_i^d)_\mathbf{x}}{2}. \tag{3.18}$$

We obtain the four nonnegative layers for each optical flow field $\mathbf{V}_i^{ds} \geq 0$, with $s \in \{+, -\}$. Also, the reconstruction $\mathbf{R}_i^{ds} \geq 0$ and consequently the basis

[3] If we take the squared activations $h_{ij}(\mathbf{m})^2$ in E_o instead of the linear activity, E_o is identical to E_{p2} with an inverted sign and different weights λ_o and λ_p.

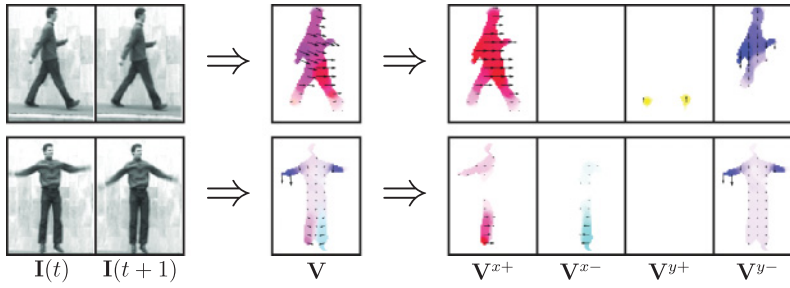

Figure 3: Illustration of the direction selective, nonnegative encoding for two example sequences. First the optical flow (i.e., the movement of every pixel between two consecutive frames) is estimated. Then the optical flow field is decomposed into four nonnegative directions: right, left, up, and down.

vectors $\bar{\mathbf{W}}_j^{ds} \geq 0$ now have four feature layers. This process is illustrated in Figure 3. This leads to the layered reconstruction energy function

$$E_R = \frac{1}{2} \sum_{i,d,s} \left\| \mathbf{V}_i^{ds} - \sum_j \mathrm{conv}_2(\mathbf{H}_{ij}, \bar{\mathbf{W}}_j^{ds}) \right\|_2^2. \tag{3.19}$$

The same activities $\mathbf{H}_{ij}$ are used for each layer ds of the basis vector $\bar{\mathbf{W}}_j^{ds}$, so the single directions are not learned independently but are coupled through the common activation.

3.7 VNMF Algorithm. The overall energy function is

$$E = E_R + \lambda_P E_p + \lambda_o E_o, = \frac{1}{2} \sum_{i,d,s} \left(\left\| \mathbf{V}_i^{ds} - \sum_j \mathrm{conv}_2(\mathbf{H}_{ij}, \bar{\mathbf{W}}_j^{ds}) \right\|_2^2 \right) \tag{3.20}$$

$$+ \frac{1}{2} \lambda_P \sum_{i,d,s,j,\mathbf{m}} \mathbf{R}_{ij\mathbf{m}}^{ds\top} (\mathbf{R}_i^{ds} - \mathbf{R}_{ij\mathbf{m}}^{ds}) + \lambda_o \sum_{i,d,s,j,\mathbf{m}} h_{ij}(\mathbf{m}) \bar{\mathbf{W}}_j^{ds\top} \bar{\mathbf{W}}_j^{ds}. \tag{3.21}$$

The energy function is minimized by the algorithm introduced in section 3.1. The derivatives for the normalized basis vectors are

$$(\nabla_{\bar{\mathbf{W}}_j^{ds}} E)^+ = \sum_i \left((1 + \lambda_P) \mathrm{corr}_2(\mathbf{R}_i^{ds}, \mathbf{H}_{ij}) + \lambda_o \bar{\mathbf{W}}_j^{ds} \sum_{\mathbf{m}} h_{ij}(\mathbf{m}) \right), \tag{3.22}$$

$$(\nabla_{\bar{\mathbf{W}}_j^{ds}} E)^- = \sum_i \left(\mathrm{corr}_2(\mathbf{V}_i^{ds}, \mathbf{H}_{ij}) + \lambda_P \bar{\mathbf{W}}_j^{ds} \mathbf{H}_{ij}^\top \mathbf{H}_{ij} \right), \tag{3.23}$$

considering the inner derivations due to the Euclidean normalization

$$(\nabla_{\mathbf{W}_j^{ds}} E)^+ = (\nabla_{\bar{\mathbf{W}}_j^{ds}} E)^+ + \bar{\mathbf{W}}_j^{ds} \bar{\mathbf{W}}_j^{ds\top} (\nabla_{\bar{\mathbf{W}}_j^{ds}} E)^-, \tag{3.24}$$

$$(\nabla_{\mathbf{W}_j^{ds}} E)^- = (\nabla_{\bar{\mathbf{W}}_j^{ds}} E)^- + \bar{\mathbf{W}}_j^{ds} \bar{\mathbf{W}}_j^{ds\top} (\nabla_{\bar{\mathbf{W}}_j^{ds}} E)^+, \tag{3.25}$$

and for the activations

$$(\nabla_{\mathbf{H}_{ij}} E)^+ = \sum_{d,s} ((1 + \lambda_p) \mathrm{corr}_2(\mathbf{R}_i^{ds}, \bar{\mathbf{W}}_j^{ds})) + \lambda_o, \tag{3.26}$$

$$(\nabla_{\mathbf{H}_{ij}} E)^- = \sum_{d,s} (\mathrm{corr}_2(\mathbf{V}_i^{ds}, \bar{\mathbf{W}}_j^{ds}) + \lambda_p \mathbf{H}_{ij} \bar{\mathbf{W}}_j^{ds\top} \bar{\mathbf{W}}_j^{ds}). \tag{3.27}$$

Each input $\mathbf{V}_i$ is normalized using the maximum norm. We now define two algorithms. The first is tNMF with the weights $\lambda_o = 0.2$ and $\lambda_p = 0$. The second, VNMF with the parameters $\lambda_o = 0.2$ and $\lambda_p = 0.2$, enforces topological sparsity.[4]

4 Extracted Motion Patterns

We now discuss how far unsupervised learning algorithms are able to extract dynamic form patterns related to body part and face movements. The discussion is focused on

1. The comparison of different algorithms (PCA, sNMF, tNMF, VNMF)
2. Parameter dependencies (pattern size and number of patterns)
3. Different input stimuli (full body movements and face expression movements)

The right choice of the learning data set is crucial, since the unsupervised learning algorithms can represent only what the observations provide. The majority of natural optical flow patterns, however, are due to ego movement and thus related to depth configurations of the observed scene. The related optical flow patterns match the rotational and translational patterns found in area MST. Instead, here we focus our discussion on the analysis of dynamic form patterns as they appear in human action recognition. Hence, we choose data sets without camera motion and a large variety of limb as well as face movements.

The first data set we choose for our analysis is the Weizmann human action recognition data set (Blank et al., 2005), which consists of videos

[4] We tested different parameter setting and found that these settings result in a good trade-off between sparsity and reconstruction quality throughout a variety of different data sets.

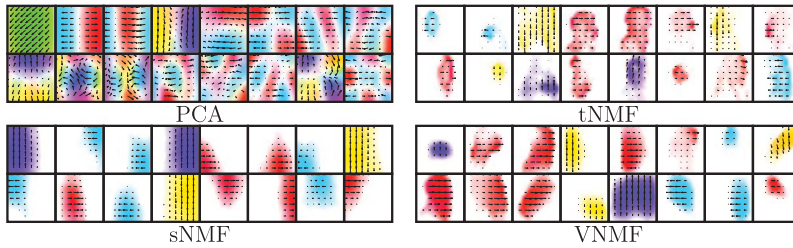

Figure 4: (Left) Both sets are learned on randomly extracted 16 × 16 optical flow patches. (Right) The sets were learned on the entire optical flow fields using the presented translation-invariant algorithms. Unlike the PCA, the three NMF-based algorithms learned parts-based patterns. The patterns extracted with VNMF describe body parts (e.g., a head, arm shapes).

showing nine people, each performing ten natural actions: running, walking, skipping, jumping jack, jumping forward on two legs, jump in place on two legs, gallop sideways, wave two hands, wave one hand and bend. For the face movements, we analyze a facial expression data set (Valstar & Pantic, 2010) with two persons expressing the six basic emotions defined in Ekman and Rosenberg (1997): anger, disgust, fear, happiness, sadness, and surprise.

The experiments on the Weizmann data set were performed with a high-quality OFE algorithm (Sun et al., 2010) based on a global optimization including segmentation information. Unfortunately due to the regularization properties, this OFE is not robust (e.g., on the KTH and UCF-Sports data set or for the face movements), so for the classification, we use the OFE algorithm described in (Guthier et al., 2013). The basis vectors used for the classification experiments (including the Weizmann data set) are all learned on the robust optical flow.

4.1 Patterns extracted with PCA, sNMF, tNMF, and VNMF. In Figure 4, 16 optical flow patterns learned from randomly selected 16 × 16 patches decomposed with PCA and sNMF, as well as patterns learned with the translation-invariant tNMF and VNMF are shown. Basis vectors with horizontal motion dominate those with vertical motion, in good accordance with the intuitive observation that horizontal human movements like walking, and running are statistically more frequent than vertical movements like jumping or hand waving. Due to the nonnegativity and sparsity constraints, all NMF basis vectors are more parts based than the holistic PCA patterns. A further distinction between the NMF and the PCA is that the NMF favors purely translational patterns, even though we did not restrict our basis vectors concerning the distribution throughout the different movement directions. This property is rather a natural restriction, since all elements of

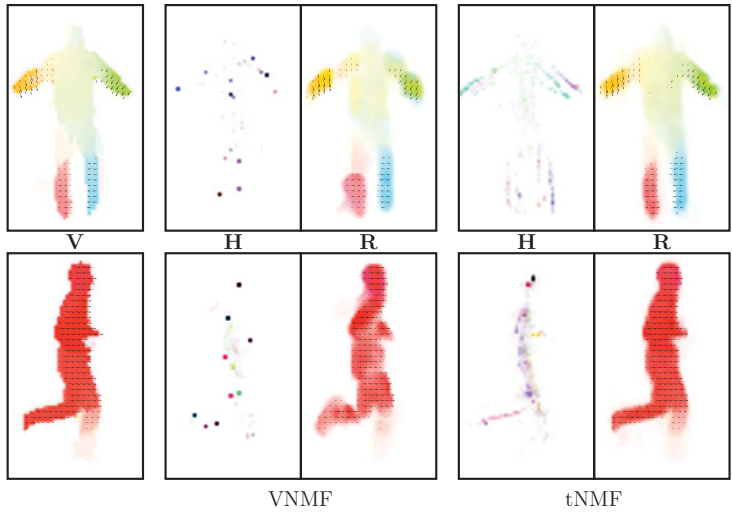

Figure 5: From left to right, the optical flow input $\mathbf{V}_i$, the summed activity image $\mathbf{H}_i = \sum_j \mathbf{H}_{ij}$ (different colors belong to different $\mathbf{H}_{ij}$), and the reconstruction $\mathbf{R}_i$ for VNMF and for tNMF, for two example inputs.

rigid body parts yield consistent translational movements. The main distinction between the different basis vectors is thus the form and overall movement direction.

The effect of the topological sparsity can best be examined by comparing the activity patterns of the two translation-invariant algorithms tNMF and VNMF depicted in Figure 5. The activity images learned with VNMF are topologically sparse and thus yield a small number of dominant and sharply localized activations that are located all over the moving body parts (e.g., on the limbs or the head). Since only a single activity is used to reconstruct a specific area, the corresponding basis vectors tend to represent this specific part (e.g., the head or a limb). The activity patterns obtained without E_p are much more blurry, therefore, the corresponding basis vectors are less distinct.

Unlike the patterns extracted with the other algorithms, the VNMF patterns depict connected, round, or ellipsoidal shapes, which are well suited to describe the form of limbs. Given the topologically sparse activations and the limblike forms of the basis vectors, we conclude that the topology-preserving VNMF algorithm is best suited to learn prototypical dynamic form patterns.

However, the focus on topological sparsity comes at the cost of reconstruction quality. If we compare the reconstructions in Figure 5, we can observe that VNMF does not preserve as many details of the input as the

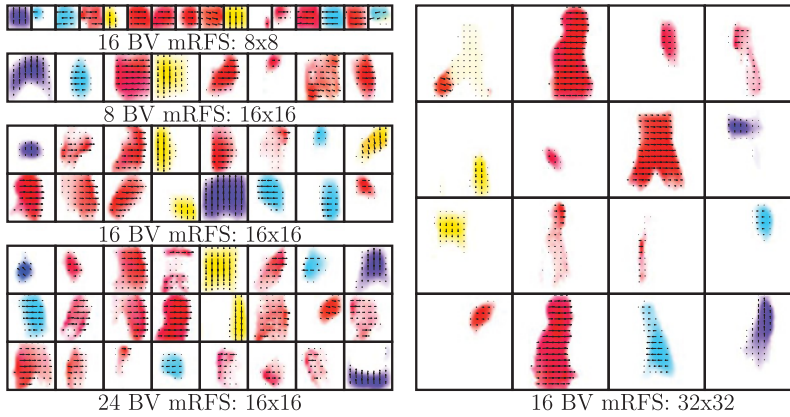

Figure 6: Different basis vector sets for varying number of basis vectors (J) and maximum receptive field size (mRFS).

tNMF algorithm. Even worse, the basis vector is used to reconstruct the head is also used to reconstruct the leg movement in the jumping jack sequence (upper row of Figure 5). Since we restrict our model to a limited number of basis vectors with bounded receptive fields, not all possible patterns can be explicitly represented. In other words, we decrease the degrees of freedom for the learning algorithm by enforcing a nonnegative and topological sparse representation. As compensation, we could relax the restrictions otherwise, for example, by increasing the number of basis vectors, which we discuss in the following.

4.2 Varying Number (J) and Size (mRFS) of the Basis Vectors. Our target is to find prototypical basis vectors, so we have to investigate how many basis vectors (J) are required to represent the different human limb movements present in our data set. The form that can be expressed by the basis vectors is limited by the maximum receptive field size (mRFS); thus, we discuss the variation of the mRFS as well. First, we analyze the parameter variations in terms of visual interpretability.

In Figure 6, example basis vectors extracted for a varying J and mRFS are shown. The main observation is that by increasing the mRFS, more discriminative basis vectors can be learned and therefore the algorithm benefits from an increased J. For the smallest mRFS (8×8) only a few different basis vectors are represented and several basis vectors are redundant. The middle-sized (16×16) basis vectors already make use of the increased J, and up to 16 different basis vectors are extracted. However, a further increase of J produces mostly redundant basis vectors. The small and middle-sized basis vector sets are rather homogeneous concerning the expressed shape

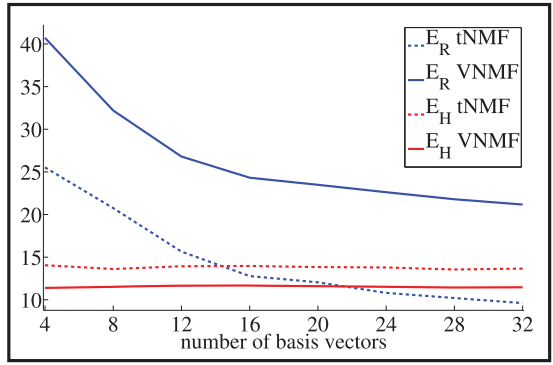

Figure 7: Average reconstruction (E_R) and sparsity (E_H) energy per image, measured over five runs with different initializations. E_H is independent of the number of basis vectors, and VNMF needs fewer activations than tNMF. However, VNMF requires more basis vectors to achieve a similar E_R as the tNMF algorithm.

size of each individual basis vector. When the mRFS is further increased (32×32), two kinds of patterns emerge: on the one hand, large and highly prototypical patterns that describe almost entire human figures and on the other hand small patterns with similar shapes as extracted for the smaller mRFS.

For a quantitative analysis, we compare the average reconstruction error and sparsity per input for the tNMF and VNMF algorithm as depicted in Figure 7. The impressions gained from visual interpretation of the basis vectors are confirmed by the error measurements. On the one hand, the VNMF algorithm needs a larger J to achieve the same reconstruction quality as the tNMF algorithm, since it is enforced to generate a topologically sparse representation. On the other hand, due to the prototypical nature of the VNMF patterns, the VNMF needs fewer activities and is thus sparser.

4.3 Motion Patterns of Face Movements: Facial Action Units. To show the generality of the VNMF algorithm, we applied it with the same parameter settings on a data set showing face movements. The movements of specific face areas, such as the corners of the lips, eyebrows, or the chin, are so-called action units. Facial action units are strongly related to facial expressions, such as the six basic emotions defined in Ekman and Rosenberg (1997). Face movements are thus very important features in interhuman nonverbal communication.

In Figure 8, 15 learned basis vectors and the corresponding activities for one example input motion field are shown. Similar to the activation patterns learned on the human full body movements, the activities are topologically

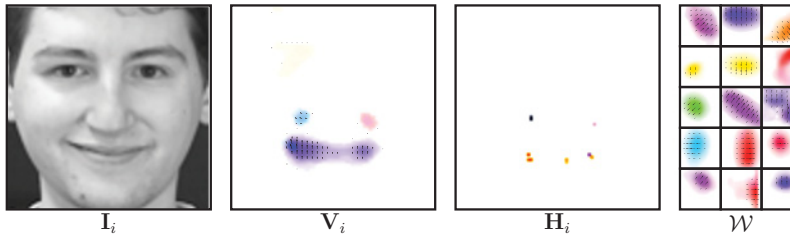

Figure 8: From left to right: Input image $\mathbf{I}_i$ from the MMI data set (Valstar & Pantic, 2010), corresponding optical flow $\mathbf{V}_i$, summed activity image $\mathbf{H}_i = \sum_j \mathbf{H}_{ij}$ (different colors correspond to different $\mathbf{H}_{ij}$) and extracted basis vector set $\mathcal{W}$. The activations are located on moving face parts such as the corners of the lips, corresponding to facial action units.

sparse. Changes in the number of basis vectors and the mRFS have the same effect as for the human full body movements shown in Figure 4. The level of detail preserved by the VNMF algorithm relays highly on the quality of the underlying optical flow estimation, which unfortunately is unable to preserve all detailed movements of the action units. Nonetheless, the movements obtained by the optical flow are represented by the basis vectors and further localized by the corresponding activations.

5 Classification

Next, the descriptive power of the learned dynamic form patterns is shown in classification experiments. We chose a multiclass one versus one implementation of SVMs (Chang & Lin, 2011) for our classification. The classification tasks are human action recognition (Weizmann, KTH, and UCF-Sports) and facial expression recognition, which play an important role in human-human and human-machine interaction. Human action recognition is a vivid and complex research topic because of the variety of possible dynamics of human actions. (For a detailed review on the topic, see Aggarwal & Ryoo, 2011.)

Actions consist of temporal sequences of poses, so the dynamic form patterns extracted from the optical flow should outperform static form patterns extracted from the gradient amplitudes. To evaluate this hypothesis, we show classification results for three cases: using only the dynamic form patterns, using only the static form patterns, and classifying the combination of both patterns. We further evaluate the classification performance of the four learning algorithms (PCA, sNMF, tNMF, and VNMF), different parameter configurations, and state-of-the-art HOG/HOF descriptors (Dalal et al., 2006) and compare the obtained results to related work.

5.1 Preprocessing. Since we are interested only in classifying the different actions and not in the localization of the persons, we use figure-centric input data. For the Weizmann data, we calculate the center from the segmentation masks provided with the data set. For the KTH data set, we define the figure center as the weighted center of the optical flow, and for the UCF-Sports data set, we manually selected the persons. For each video, consider only the frames with a person fully visible and stop the video once a person has left the image. For all actions, the persons (or faces) are automatically cut out around the center and rescaled to a resolution of 128×128 pixels.

The optical flow for each video is normalized using the maximum norm; thus, the highest velocity for each video is 1, independent of the action or emotion shown in the video. Due to this normalization, the information about the absolute amplitude of the velocity between the different classes is lost, which is a drawback, especially for distinguishing between similar classes such as jogging and running. However, in most cases, the exact velocity is already lost during the optical flow estimation.

5.2 Feature Extraction. The dynamic and the static basis vectors are learned by the same algorithms throughout the experiments, and the features for the classification are calculated in the same way for both streams as well. The features for the classification are the simple-cell, complex-cell responses to the learned basis vectors. Since our goal is to find a basis vector set that can describe basic human movements, rather than data set to specific patterns, we applied the basis vectors learned on the Weizmann data set on the KTH and UCF-Sports data set as well. For the face movements, we learned the basis vector set on videos of the MMI data set (Valstar & Pantic, 2010) and evaluated the classification performance on the Dollar FER data set (Dollar et al., 2005).

The simple-cell activity responses $\mathbf{H}_{ij}$ for each learned basis vector $\mathbf{W}_j$ for each input $\mathbf{V}_i$ are calculated in a feedforward manner by using the correlation $\mathbf{H}_{ij} = \mathrm{corr}_2(\mathbf{V}_i, \mathbf{W}_j)$. Our complex-cell response $\mathbf{C}_{ij}(\mathbf{H}_{ij})$ consists of three operations: activities below a fixed threshold are discarded to suppress noise effects; the thresholded activities are summed inside 50% overlapping pooling cells, each with a dimension of 32×32 pixels and thus, the dimension of each simple-cell activity $\mathbf{H}_{ij}$ is reduced from 128×128 to 7×7; and the activities for each basis vector are normalized for each pooling cell using the maximum norm.

5.3 Classification. For the training of the classifiers, the KTH data set is split into a training and a test set in a 16:9 ratio as proposed by the original authors (Schuldt et al., 2004). Since the Weizmann and UCF-Sports data sets have no defined training and testing sets, we perform leave-one-out experiments. The people used for the learning of the basis vectors are

Table 1: Classification Results for the Dynamic (Optical Flow) and Combined (Optical Flow + Gradient) Patterns Extracted with Different Algorithms, with J = 16 and mRFS = 16 × 16.

	Optical Flow				Optical Flow + Gradient			
	PCA	sNMF	tNMF	VNMF	PCA	sNMF	tNMF	VNMF
Weizmann	0.94	0.94	0.94	**0.99**	0.94	0.94	0.94	**0.99**
KTH	0.87	0.90	0.91	0.90	0.89	0.91	**0.93**	**0.93**
UCF	0.71	0.80	0.77	0.89	0.83	0.80	0.80	**0.93**
FER	0.72	0.70	0.76	0.75	0.67	0.66	0.80	**0.82**

Note: The numbers in bold mark the best-performing algorithm.

discarded in the evaluation. The FER data set has only two people, so we use one for training and one for testing.

The SVM is learned with an RBF kernel and a soft-margin parameter to increase robustness. The corresponding parameters are obtained using five-fold cross-validation on the training data. Once the SVM classifiers are trained, each frame of each video is classified individually. The final classification result for each video is the weighted (using the class probabilities provided by the SVM) average of all its frame results.

5.4 Results for Different Learning Algorithms. Table 1 shows the classification results for four learning algorithms: PCA, sNMF, tNMF, and VNMF. All four algorithms show good results on the Weizmann data set.[5] The benefits of the topological sparse decomposition of the VNMF are mainly noticable for the UCF-Sports data set, which, unlike the other data sets, includes strong variations in the viewpoints and how the actions are performed. Here the VNMF outperforms the other methods significantly.

5.5 Results for VNMF with Different Parameters J and mRFS. Table 2 shows the classification results for the dynamic form (optical flow) patterns for a different number (J) and size (mRFS) of basis vectors. Throughout the data sets, there is no clear tendency as to which parameters perform best. Furthermore, the variations among the human action recognition data sets are rather small ($\leq 4\%$), with the exception of the FER, where 16 basis vectors of size 16 × 16 perform best. Thus, a class-specific description can be achieved by as few as eight translation invariant patterns.

5.6 Comparison with State-of-the Art Features and Algorithms. To make the comparison of the learned basis vectors to state-of-the-art features

[5] All misclassified videos involve the jump on one leg class that is often discarded by other researchers (e.g., Jhuang et al., 2007), leading to the reduced Weiz.9 data set.

Table 2: Classification Results Using the Optical Flow Patterns for Different J and mRFS.

mRFS	8×8			16×16			24×24		
J	8	16	24	8	16	24	8	16	24
Weizmann	0.98	0.99	0.99	0.99	0.99	0.97	0.97	0.99	**1.00**
KTH	0.89	0.88	**0.90**	0.88	**0.90**	0.88	0.88	0.88	**0.90**
UCF	0.89	0.88	0.87	0.89	0.89	0.90	0.89	**0.91**	0.89
FER	0.74	0.73	0.72	0.73	**0.75**	0.73	0.71	0.68	0.73

Note: The bold numbers represents the best-performing algorithm.

Table 3: Classification Results on the Weizmann (Blank et al., 2005), KTH (Schuldt et al., 2004), UCF-Sports (Rodriguez et al., 2008), and FER (Dollar et al., 2005) for the VNMF Algorithm ($J = 16$, mRFS $= 16 \times 16$) Compared to State-of-the-Art HOG/HOF Features and Related Work.

		KTH	Weizmann	UCF	FER
VNMF	Gradient	0.71	0.80	0.87	0.31
	Optical flow	0.90	**0.99**	0.89	0.75
	Optical flow + gradient	0.93	**0.99**	**0.93**	**0.82**
HOG	Gradient	0.67	0.74	0.77	0.39
	Optical flow	0.80	0.86	0.78	0.63
	Optical flow + gradient	0.82	0.87	0.80	0.71
Related work	Jhuang et al. (2007)	0.92	0.96	-	-
	Dean et al. (2010)	0.86	-	-	-
	Guha and Ward (2012)	-	**0.99**	-	**0.82**
	Klaser, Marszalek, and Laptev (2010)	-	-	0.90	-
	Amiri et al. (2012)	**1.00**	-	-	-

Note: The bold numbers mark the best-performing algorithm.

extractors independent of the preprocessing (figure-centering and optical flow estimation) we calculated HOG/HOF features (Dalal et al., 2006) on the same data as used for the learned basis vectors. The cell/block building of the HOG features is identical to the overlapping summation pooling blocks we use in our complex cell response, so we used the same pooling sizes for both types of features.

The results are depicted in Table 3. Throughout all data sets, the dynamic form patterns (optical flow) outperform the static form patterns (gradient), while the combined features (optical flow + gradient) perform best. This result is of particular interest because it shows that the dynamic information contributes more to the recognition of biological motion than the static information. However, each stream on its own is able to recognize some of

the actions, and the information from both streams is complementary, since the results improve considerably when combining form and motion.

In addition, the learned pattern features outperform the designed state-of-the-art HOG/HOF descriptors significantly. Table 3 further shows results of other related work, such as the biological-inspired feedforward system based on templates (Jhuang et al., 2007) or using sparse coding on space-time-volumes (Dean et al., 2010). Our system outperforms the other biological-inspired systems and has equal results to state-of-the-art algorithms in computer vision on three of the four data sets.

6 Summary and Discussion

The proposed topological sparse VNMF algorithm is capable of extracting template-like optical flow patterns that describe forms with consistent movement, thus dynamic form patterns. The corresponding activations are sharply localized and located on human body parts. The experiments show that it is possible to extract prototypical forms of human body parts in an unsupervised fashion by enforcing reasonable restrictions on the nature of the decomposition. In addition, the learned patterns have more discriminative power than the designed HOG/HOF descriptors, which indicates that a learned overcomplete basis can represent and discriminate between a larger variety of possible optical flow and gradient combinations than the histogram/binning approach.

One of the appealing properties of the algorithm is that all additional energy functions scale in the same way as the reconstruction energy, which makes the parameterization easy and interpretable. We applied the same parameters to the four data sets and for two unrelated input types (optical flow and gradient amplitudes) and in all cases extracted prototypical parts-based basis vectors. The optimization itself is fast and parameter free, except for the size (mRFS) and number of basis vectors (J). The free parameters can be set based on the desired reconstruction quality, after visual inspection of the extracted patterns or based on the obtained classification result. We found no strong parameter dependency with respect to the classification results in our experiments.

While both static and dynamic information contribute to the classification, the dynamic patterns are more efficient. The spatial configuration of the static and dynamic form patterns, encoded by the pooled activations, describes the full body pose. While the static patterns can encode poses, a vast number of poses are themselves not action specific. In those cases, only the dynamics (e.g., described by optical flow fields) can discriminate similar classes like jogging and running. Another way to include the dynamics is on a larger scale by encoding transitions between poses, as proposed in Lange and Lappe (2006) and Theusner et al. (2014).

Our experiments indicate that the spatial configuation of parts-based optical flow patterns can improve the recognition of natural human actions.

The evaluation is focused on the classification of multiple actions in the presence of viewpoint variations and cluttered background. These results are consistent with reports that show that low-level motion cues can improve the recognition of biological motion in the presence of noise. Due to the static form patterns, our model is capable of recognizing human actions, even if the optical flow is disabled. It would be interesting to extend our robust, parts-based dynamic pose description with an additional pose transition model and analyze how it relates, for example, to the inversion effect or the prominent role of feet motion in biological motion recognition (Troje & Westhoff, 2006).

Appendix A: Gradients for Translation Invariant Learning

$$E_r = \frac{1}{2}\sum_i \|\mathbf{V}_i - \mathbf{R}_i\|_2^2 = \frac{1}{2}\sum_{i,\mathbf{x}} \left(v_i(\mathbf{x}) - \sum_{j,\mathbf{m}} h_{ij}(\mathbf{m})\bar{w}_j(\mathbf{x}-\mathbf{m}) \right)^2 \tag{A.1}$$

Gradients for the activities:

$$\nabla_{h_{ij}(\mathbf{m})} E_r = \sum_{\mathbf{x}} \left(-\bar{w}_j(\mathbf{x}-\mathbf{m})(v_i(\mathbf{x}) - r_i(\mathbf{x})) \right) \tag{A.2}$$

$$= \sum_{\mathbf{x}} r_i(\mathbf{x})\bar{w}_j(\mathbf{x}-\mathbf{m}) - \sum_{\mathbf{x}} v_i(\mathbf{x})\bar{w}_j(\mathbf{x}-\mathbf{m}) \tag{A.3}$$

$$\nabla_{\mathbf{H}_{ij}} E_r = \underbrace{\operatorname{corr}_2(\mathbf{R}_i, \bar{\mathbf{W}}_j)}_{:=\left(\nabla_{\mathbf{H}_{ij}} E_r\right)^+} - \underbrace{\operatorname{corr}_2(\mathbf{V}_i, \bar{\mathbf{W}}_j)}_{:=\left(\nabla_{\mathbf{H}_{ij}} E_r\right)^-} \tag{A.4}$$

Gradients for the basis vectors, with the substitution $\mathbf{x}' = \mathbf{x} - \mathbf{m}$:

$$\nabla_{\bar{w}_j(\mathbf{x}')} E_r = \sum_{i,\mathbf{x}} \left(-h_{ij}(\mathbf{x}-\mathbf{x}')(v_i(\mathbf{x}) - r_i(\mathbf{x})) \right) \tag{A.5}$$

$$= \sum_i \left(\sum_{\mathbf{x}} r_i(\mathbf{x})h_{ij}(\mathbf{x}-\mathbf{x}') - \sum_{\mathbf{x}} v_i(\mathbf{x})h_{ij}(\mathbf{x}-\mathbf{x}') \right) \tag{A.6}$$

$$\nabla_{\bar{\mathbf{W}}_j} E_r = \underbrace{\sum_i \operatorname{corr}_2(\mathbf{R}_i, \mathbf{H}_{ij})}_{:=\left(\nabla_{\bar{\mathbf{W}}_j} E_r\right)^+} - \underbrace{\sum_i \operatorname{corr}_2(\mathbf{V}_i, \mathbf{H}_{ij})}_{:=\left(\nabla_{\bar{\mathbf{W}}_j} E_r\right)^-} \tag{A.7}$$

Appendix B: Gradients for Inhibition Energy Function

$$E_p = \frac{1}{2}\sum_{i,j,\mathbf{m}} \mathbf{R}_{ij\mathbf{m}}^{\top}\left(\mathbf{R}_i - \mathbf{R}_{ij\mathbf{m}}\right) = \underbrace{\frac{1}{2}\sum_{i,j,\mathbf{m}} \mathbf{R}_{ij\mathbf{m}}^{\top}\mathbf{R}_i}_{:=E_{p1}} - \underbrace{\frac{1}{2}\sum_{i,j,\mathbf{m}} \mathbf{R}_{ij\mathbf{m}}^{\top}\mathbf{R}_{ij\mathbf{m}}}_{:=E_{p2}} \tag{A.8}$$

$$E_{p1} = \frac{1}{2}\sum_{i,\mathbf{x}}\left(\sum_{j,\mathbf{m}} \bar{w}_j(\mathbf{x}-\mathbf{m})h_{ij}(\mathbf{m})\right)^2 \tag{A.9}$$

$$E_{p2} = \frac{1}{2}\sum_{i,\mathbf{x},j,\mathbf{m}} \bar{w}_j^2(\mathbf{x}-\mathbf{m})h_{ij}^2(\mathbf{m}) \tag{A.10}$$

Gradients for the basis vectors, with the substitution $\mathbf{x}' = \mathbf{x} - \mathbf{m}$:

$$\nabla_{\bar{w}_j(\mathbf{x}')} E_{p1} = \sum_{i,\mathbf{x}} h_{ij}(\mathbf{x}-\mathbf{x}')r_i(\mathbf{x}) \tag{A.11}$$

$$\nabla_{\bar{\mathbf{W}}_j} E_{p1} = \sum_i \mathrm{corr}_2(\mathbf{R}_i, \mathbf{H}_{ij}) \tag{A.12}$$

$$\nabla_{\bar{w}_j(\mathbf{x}')} E_{p2} = \sum_i \bar{w}_j(\mathbf{x}')\sum_{\mathbf{x}} h_{ij}^2(\mathbf{x}-\mathbf{x}') \tag{A.13}$$

$$\nabla_{\bar{\mathbf{W}}_j} E_{p2} = \bar{\mathbf{W}}_j \sum_i \mathbf{H}_{ij}^{\top}\mathbf{H}_{ij} \tag{A.14}$$

$$\nabla_{\bar{\mathbf{W}}_j} E_p = \nabla_{\bar{\mathbf{W}}_j} E_{p1} - \nabla_{\bar{\mathbf{W}}_j} E_{p2} = \sum_i \mathrm{corr}_2(\mathbf{R}_i, \mathbf{H}_{ij}) - \bar{\mathbf{W}}_j \sum_i \mathbf{H}_{ij}^{\top}\mathbf{H}_{ij} \tag{A.15}$$

Gradient for the activities:

$$\nabla_{h_{ij}(\mathbf{m})} E_{p1} = \sum_{\mathbf{x}} \bar{w}_j(\mathbf{x}-\mathbf{m})r_i(\mathbf{x}) \tag{A.16}$$

$$\nabla_{\mathbf{H}_{ij}(\mathbf{m})} E_{p1} = \mathrm{corr}_2(\mathbf{R}_i, \bar{\mathbf{W}}_j) \tag{A.17}$$

$$\nabla_{h_{ij}(\mathbf{m})} E_{p2} = h_{ij}(\mathbf{m})\sum_{\mathbf{x}} \bar{w}_j^2(\mathbf{x}-\mathbf{m}) \tag{A.18}$$

$$\nabla_{\mathbf{H}_{ij}} E_{p2} = \mathbf{H}_{ij}\bar{\mathbf{W}}_j^{\top}\bar{\mathbf{W}}_j \tag{A.19}$$

$$\nabla_{\mathbf{H}_{ij}} E_p = \nabla_{\mathbf{H}_{ij}} E_{p1} - \nabla_{\mathbf{H}_{ij}} E_{p2} = \mathrm{corr}_2(\mathbf{R}_i, \bar{\mathbf{W}}_j) - \mathbf{H}_{ij}\bar{\mathbf{W}}_j^{\top}\bar{\mathbf{W}}_j \tag{A.20}$$

References

Aggarwal, J., & Ryoo, M. S. (2011). Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, *43*(3), 16.

Amiri, S. M., Nasiopoulos, P., & Leung, V. (2012). Non-negative sparse coding for human action recognition. In *Proceedings of the 2012 19th IEEE International Conference on Image Processing* (pp. 1421–1424). Piscataway, NJ: IEEE.

Ashbridge, E., Perrett, D. I., Oram, M. W., & Jellema, T. (2000). Effect of image orientation and size on object recognition: Responses of single units in the macaque monkey temporal cortex. *Cognitive Neuropsychology*, *1*(17), 13–34.

Beintema, J. A., & Lappe, M. (2002). Perception of biological motion without local image motion. *Proceedings of the National Academy of Sciences*, *8*(99), 5661–5663.

Blake, R., & Shiffrar, M. (2007). Perception of human motion. *Annu. Rev. Psychol.*, *58*, 47–73.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., & Basri, R. (2005). Actions as space-time shapes. In *Proceedings of the Tenth IEEE International Conference on Computer Vision* (Vol. 2, pp. 1395–1402). Piscataway, NJ: IEEE.

Cadieu, C. F., & Olshausen, B. A. (2012). Learning intermediate-level representations of form and motion from natural movies. *Neural Computation*, *24*(4), 827–866.

Casile, A., & Giese, M. A. (2005). Critical features for the recognition of biological motion. *Journal of Vision*, *5*(4), 711–720.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(27), 1–27. http://www.csie.ntu.edu.tw/cjlin/libsvm.

Choi, S. (2008). Algorithms for orthogonal nonnegative matrix factorization. In *Proceedings of the IEEE International Joint Conference on Neural Networks* (pp. 1828–1832). Piscataway, NJ: IEEE.

Chouchourelou, A., Golden, A., Shiffrar, M., & Chouchourelou, A. (2012). *What does "biological motion" really mean? Differentiating visual percepts of human, animal, and non-biological motions*. New York: Oxford University Press.

Cichocki, A., Zdunek, R., Phan, A. H., & Amari, S.-i. (2009). *Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation*. Hoboken, NJ: Wiley.

Dalal, N., Triggs, B., & Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *Computer Vision–ECCV 2006* (pp. 428–441). New York: Springer.

Dean, T., Washington, R., & Corrado, G. (2010). Sparse spatiotemporal coding for activity recognition. In *Proceedings of the 11th IEEE International Symposium Multimedia* (pp. 645–650). Piscataway, NJ: IEEE.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Ding, C. H., Li, T., & Jordan, M. I. (2010). Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(1), 45–55.

Dollar, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *Proceedings of the 2nd Joint IEEE International*

Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (pp. 65–72). Piscataway, NJ: IEEE.

Efros, A. A., Berg, A. C., Mori, G., & Malik, J. (2003). Recognizing action at a distance. In *Proceedings of the Ninth IEEE International Conference on Computer Vision* (pp. 726–733). Piscataway, NJ: IEEE.

Eggert, J., & Koerner, E. (2004). Sparse coding and NMF. In *Proceedings of the 2004 IEEE International Joint Conference on Neural Network* (Vol. 4, pp. 2529–2533). Piscataway, NJ: IEEE.

Eggert, J., Wersing, H., & Koerner, E. (2004). Transformation-invariant representation and NMF. In *Proceedings of the 2004 IEEE International Joint Conference on Neural Networks* (Vol. 4, pp. 2535–2539). Piscataway, NJ: IEEE.

Ekman, P., & Rosenberg, E. L. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS)*. New York: Oxford University Press.

Fleet, D. J., Black, M. J., Yacoob, Y., & Jepson, A. D. (2000). Design and use of linear models for image motion analysis. *International Journal of Computer Vision*, *36*(3), 171–193.

Fleischer, F., Caggiano, V., Thier, P., & Giese, M. A. (2013). Physiologically inspired model for the visual recognition of transitive hand actions. *Journal of Neuroscience*, *15*(33), 6563–6580.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202.

Giese, M. A. (2014). Biological and body motion perception. In J. Wagemans (Ed.), *Oxford handbook of perceptual organization*. New York: Oxford University Press.

Giese, M. A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience*, *4*(3), 179–192.

Grossman, E. D., & Blake, R. (2002). Brain areas active during visual perception of biological motion. *Neuron*, *35*, 1167–1175.

Grossman, E. D., Jardine, N. L., & Pyles, J. A. (2010). FMR-adaptation reveals invariant coding of biological motion on the human STS. *Frontiers in Human Neuroscience*, *4*.

Guha, T., & Ward, R. K. (2012). Learning sparse representations for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(8), 1576–1588.

Guthier, T., Eggert, J., & Willert, V. (2012). Unsupervised learning of motion patterns. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* (Vol. 20, pp. 323–328). New York: Springer-Verlag.

Guthier, T., Willert, V., Schnall, A., Kreuter, K., & Eggert, J. (2013). Non-negative sparse coding for motion extraction. In *Proceedings of the IEEE International Joint Conference on Neural Networks*. Piscataway, NJ: IEEE.

Hoyer, P. O. (2002). Non-negative sparse coding. In *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing* (pp. 557–565). Piscataway, NJ: IEEE.

Hyvarinen, A., Hurri, J., & Hoyer, P. O. (2009). *Natural image statistics*. New York: Springer.

Jhuang, H., Serre, T., Wolf, L., & Poggio, T. (2007). A biologically inspired system for action recognition. In *Proceedings of the IEEE 11th International Conference on Computer Vision* (pp. 1–8). Piscataway, NJ: IEEE.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, *14*(2), 201–211.

Klaser, A., Marszalek, M., Laptev, I., & Schmid, C., et al. (2010). *Will person detection help bag-of-features action recognition?* Grenoble: INRIA Center.

Lange, J., & Lappe, M. (2006). A model of biological motion perception from configural form cues. *Journal of Neuroscience*, *26*(11), 2894–2906.

LeCun, Y., Huang, F. J., & Bottou, L. (2004). Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 2, pp. II–97. Piscataway, NJ: IEEE.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.

Li, S. Z., Hou, X. W., Zhang, H. J., & Cheng, Q. S. (2001). Learning spatially localized, parts-based representation. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 1, pp. I–207. Piscataway, NJ: IEEE.

Minsky, M., & Papert, S. (1969). *Perceptrons*. Cambridge, MA: MIT Press.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, *37*(23), 3311–3325.

Pitzalis, S., Sdoia, S., Bultrini, A., Committeri, G., Di Russo, F., Fattori, P., . . . Galati, G. (2013). Selectivity to translational egomotion in human brain motion areas. *PloS One*, *8*(4), e60241.

Puce, A., & Perrett, D. (2003). Electrophysiology and brain imaging of biological motion. *Philosophical Transactions of the Royal Society of London, B: Biological Sciences*, *358*(1431), 435–445.

Pyles, J. A., Garcia, J. O., Hoffman, D. D., & Grossman, E. D. (2007). Visual perception and neural correlates of novel biological motion. *Vision Research*, *47*(21), 2786–2797.

Rodriguez, M. D., Ahmed, J., & Shah, M. (2008). Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–8). Piscataway, NJ: IEEE.

Rozell, C. J., Johnson, D. H., Baraniuk, R. G., & Olshausen, B. A. (2008). Sparse coding via thresholding and local competition in neural circuits. *Neural Computation*, *20*(10), 2526–2563.

Saygin, A. P. (2007). Superior temporal and premotor brain areas necessary for biological motion perception. *Brain*, *9*(130), 2452–2461.

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: A local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition* (Vol. 3, pp. 32–36). Piscataway, NJ: IEEE.

Servos, P., Osu, R., Santi, A., & Kawato, M. (2002). The neural substrates of biological motion perception: An fMRI study. *Cerebral Cortex*, *7*(12), 772–782.

Seung, D., & Lee, L. (2001). Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.) *Advances in neural information processing systems*, *13* (pp. 556–562). Cambridge, MA: MIT Press.

Sun, D., Roth, S., & Black, M. J. (2010). Secrets of optical flow estimation and their principles. In *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2432–2439). Piscataway, NJ: IEEE.

Theusner, S., de Lussanet, M., & Lappe, M. (2014). Action recognition by motion detection in posture space. *Journal of Neuroscience*, *3*(34), 909–921.

Troje, N. F., & Westhoff, C. (2006). The inversion effect in biological motion perception: Evidence for a "life detector"? *Current Biology*, *8*(16), 821–824.

Valstar, M., & Pantic, M. (2010). Induced disgust, happiness and surprise: An addition to the mmi facial expression database. In *Proc. Int. Conf. Language Resources and Evaluation, Workshop on Emotion* (pp. 65–70).

Weiner, K. S., & Grill-Spector, K. (2011). Not one extrastriate body area: Using anatomical landmarks, hMT+, and visual field maps to parcellate limb-selective activations in human lateral occipitotemporal cortex, *Neuroimage*, *56*(4), 2183–2199.

Wersing, H., & Koerner, E. (2003). Learning optimized features for hierarchical models of invariant object recognition. *Neural Computation*, *15*(7), 1559–1588.

Willert, V., & Eggert, J. (2009). A stochastic dynamical system for optical flow estimation. In *Proceedings of the IEEE 12th International Conference on Computer Vision Workshops* (pp. 711–718). Piscataway, NJ: IEEE.

Willert, V., & Eggert, J. (2011). Modeling short-term adaptation processes of visual motion detectors. *Neurocomputing*, *74*(9), 1329–1339.

Willert, V., Toussaint, M., Eggert, J., & Korner, E. (2007). Uncertainty optimization for robust dynamic optical flow estimation. In *Proceedings of the Sixth International Conference on Machine Learning and Applications* (pp. 450–457). Piscataway, NJ: IEEE.

Received October 28, 2013; accepted June 2, 2014.