Learning Representations for Animated Motion Sequence and Implied Motion Recognition

Georg Layher[1], Martin A. Giese[2], and Heiko Neumann[1]

[1] Institute for Neural Information Processing, Dept. of Engineering and Computer Sciences, Ulm University, Germany
[2] Section for Computational Sensomotorics, Dept. for Cognitive Neurology, University Clinic Tübingen, Germany

Abstract. The detection and categorization of animate motions is a crucial task underlying social interaction and decision-making. Neural representations of perceived animate objects are built into cortical area STS which is a region of convergent input from intermediate level form and motion representations. Populations of STS cells exist which are selectively responsive to specific action sequences, such as walkers. It is still unclear how and to which extent form and motion information contribute to the generation of such representations and what kind of mechanisms are utilized for the learning processes. The paper develops a cortical model architecture for the unsupervised learning of animated motion sequence representations. We demonstrate how the model automatically selects significant motion patterns as well as meaningful static snapshot categories from continuous video input. Such keyposes correspond to articulated postures which are utilized in probing the trained network to impose implied motion perception from static views. We also show how sequence selective representations are learned in STS by fusing snapshot and motion input and how learned feedback connections enable making predictions about future input. Network simulations demonstrate the computational capacity of the proposed model.

1 Introduction

Animated movements in actions, like walking, turning, etc., can be robustly detected and predictions can be derived from such spatio-temporal patterns. Giese & Poggio [8] proposed a hierarchical feedforward network architecture that aims at explaining the computational mechanisms underlying the perception of biological motion, mainly from impoverished stimuli such as point-light walkers. The proposed computational framework utilizes two separate visual pathways for segregated form and motion processing. At the top of this hierarchy prototypical motion patterns are learned (encoded in lateral recurrent asymmetric couplings) to build sequence-selective action prototypes for characteristic optical flow patterns and snapshot sequences. The outputs are finally averaged to get neural responses to biological motion sequences [5]. Several computer vision approaches have been proposed which adopt similar processing strategies of

combining form and motion processing [11, 15] or consider details of the motion processing cascade alone [6]. It is still unclear to a large extent, how motion representations in the medial superior temporal area (MST), form represantiations in the inferior temporal cortex (IT) and sequence-selective patterns in the superior temporal sulcus (STS) interact and which features are used for learning. Also no top-down influences and transfer of information between pathways has been considered so far.

Here, we propose a learning-based hierarchical model for analyzing animated motion to address previously unanswered questions. Prototypes in the form and motion pathway are established using modified Hebbian learning and we suggest how snapshot prototypes are automatically selected from the input video streams. Sequence-selective representations of articulated motions in the cortical area STS are driven jointly by input activations from motion and form prototypes. In addition, feedback connections are learned to enable STS neurons to predict expected input from form-selective IT and motion sensitive MST. We argue that for static articulated postures without continuing motion, STS representations are fed by the corresponding snapshot prototype activations [10]. In turn, STS will send feedback to stages in the segregated pathways generating neural responsiveness in the dorsal pathway on implied motion stimuli [12].

2 Model Architecture

The hierarchical model consists of two separate visual pathways for segregated form and motion processing as inspired by the work of [8] and extends it by building upon our previous work on hierarchical feedforward (FF) and feedback (FB) processing of motion and form along the dorsal and the ventral pathway [9, 3]. Intermediate level form representations (in model IT) and prototypical optical flow patterns (in model MST) are established using a modified competitive Hebbian learning scheme with convergent weight dynamics. A motion-driven reinforcement mechanism automatically selects relevant snapshots in the form path from video input streams. The activities of the prototypical form and motion cells converge in the model complex STS, where correlated temporal activations for specific sequences are learned. Sequence-selective representations are established by combined bottom-up and top-down learning, both based on Hebbian learning. The details are outlined below.

2.1 Form and Motion Processing for Animated Motion Recognition

Processing the raw input data utilizes an initial stage of orientation and direction selective filtering (in model area V1). These responses are fed into separated pathways which are selective to static form representations (areas V2 and IT) and characteristic temporal flow patterns (middle temporal area (MT/V5) and MST). Each model area consists of a three-stage hierarchy of model neurons whose computational properties have been previously reported in, e.g., [9, 3]. The structure of the processing hierarchy along with characteristic simulation results are shown in Fig. 1 and will not be further detailed in this paper.

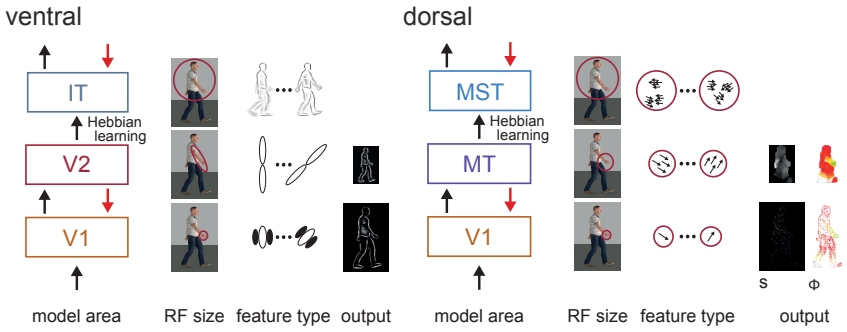

Fig. 1. Structure of hierarchical feedforward and feedback processing along the ventral (left) and dorsal (right) pathway. Each pathway is organized in a homologuous fashion, utilizing receptive fields of increasing size over different model areas. Feedback between stabilizes the feature processing from raw input. Representations of prototypical form and motion responses are established by utilizing unsupervised Hebbian learning. Processing results are shown in the output columns.

2.2 Unsupervised Learning of Form and Motion Prototypes

Hebbian learning in the form and motion pathways. In order to select the image regions that are fed to the learning of prototype representations a region of interest (ROI) is defined which represents a bounding box around the target object. Features within the target region are selected for learning feedforward connection weights in the form and the motion pathway. We employ the modified Hebbian learning rule

$$\Delta w_{ji}^{FF,s} = \eta_s \cdot \bar{v}_i^{post} \cdot (u_j^{pre} - \bar{v}_i^{post} \cdot w_{ji}^{FF,s}) \tag{1}$$

where $\Delta w_{ji}^{FF,s}$ represents the discretized rate of change in the efficacy of the weighted connections with the learning rate η_s; $s \in \{form, motion\}$ indicates that the same core mechanisms are devoted to learning in the form and motion pathway, respectively. The variables $u_j^{pre} = f(x_j)$ and $v_i^{post} = f(y_i)$ are the firing rates driven by the membrane potential of pre- and post-synaptic cells, respectively, henceforth denoted as activity. The activity $\bar{v}_i$ of the postsynaptic cell is calculated by the temporal trace rule $\bar{v}_i^t = (1-\lambda)\bar{v}_i^{t-1} + \lambda v_i^t$ [7], $0 < \lambda < 1$. In this combined temporal trace and instar learning [4] the weighting kernel in the adaptation term (in brackets) is steered by the postsynaptic activity (Oja's rule; [13]) and realizes a steepest descent learning with automatic weight normalization. The post-synaptic cells which gate the learning of their respective input weights are arranged in a layer of neurons competing for the best matching response and their subsequent ability to adapt their input weights.

Reinforcing snapshot learning. The Giese-Poggio model [8] suggests that sequence selectivity for biological motion recognition is driven by sequences of static snapshots. While the original model relies on snapshots that were regularly sampled temporally, we emphasize the automatic selection of snapshots

which correspond to strongly articulated poses. Such snapshot representations are learned in the form channel by utilizing a gating reinforcement signal which is driven by the complementary representation of motion in the dorsal stage MT/MST. Formally, the weighted integration of motion energy over a given neighborhood is calculated by

$$m_e = \int_{\Omega} \underline{u}_{\phi}(x) \cdot \Lambda(x) dx d\phi \tag{2}$$

with $\Lambda(\bullet)$ denoting a spatial kernel for weighting the relative contribution of motion responses $\underline{u}_{\phi}(\bullet)$ at spatial locations x and with direction selectivity ϕ.[3] The motion energy signal itself is a function of time which is used to steer the learning in the form pathway. We suggest that different subpopulations of static form, or snapshot, representations can be learned that correspond to either weakly or strongly articulated postures. Here, we focus on snapshot poses corresponding to highly articulated postures with signatures of maximum limb spreading. Motion energy at limbs drops during phases of high articulation when their apparent direction of motion reverses. We incorporate the function $g(\bullet)$ to control a vigilance in snapshot learning to favor form inputs which co-occur with local motion energy minima, i.e. when $\partial_t m_e = 0$, $\partial_{tt} m_e > 0$. In the weight adaptation, $\Delta w_{ji}^{FF,form}$ in Eqn.1, the learning rate is now gated by the motion dependent reinforcement, $\eta_s \cdot g(m_e)$.

2.3 Unsupervised Learning of Sequence-Selective Representations

Categorial representations in the form and motion pathway, namely in IT and MST, which were learned at the previous stage, feed forward their activations to the stage of STS. In order to stabilize the representations and activity distributions, even in the case of partial loss of input signals, the STS sequence-selective representations send top-down signals to their respective input stages.

Feedforward learning of sequence-selective motion representations. Prototypical representations with spatio-temporal sequence selectivity are learned by using a modified Hebbian instar learning mechanism similar to the learning of form and motion prototypes (Eqn.1),

$$\Delta w_{ji}^{in,FF} = \eta_{seqFF} \cdot \bar{v}_i^{post} \cdot (u_j^{pre} - \bar{v}_i^{post} \cdot w_{ji}^{in,FF}). \tag{3}$$

The weighting kernel $w_{ji}^{in,FF}$ represents convergent IT $\rightarrow$ STS and MST $\rightarrow$ STS bottom-up input to a postsynaptic STS cell (instar). η_{seqFF} denotes the learning rate and u_j and v_i are the firing rates of the pre- and post-synaptic neurons, respectively (the post-synaptic activity is again calculated via a temporal trace mechanism). The pre-synaptic activity is generated by concatenating form and motion output activations, namely $\mathbf{u} = \mathbf{u}^{\mathbf{IT}} \cup \mathbf{u}^{\mathbf{MST}}$.

[3] For whole body motion considered here, we simply integrated the motion energy over the entire ROI without subdividing the image region. An analysis at smaller scales might necessitate an integration over smaller overlapping patches.

Learning feedback connections. Sequence-selective prototypes in STS in turn learn the output weights back to the segregated form and motion prototype representations, namely STS → IT + MST. Unlike the FF learning mechanisms, the learning here is gated by the pre-synaptic cell (in STS) for their top-down weights, realizing an outstar mechanism [4]. The learning equation reads

$$\Delta w_{ji}^{out,FB} = \eta_{seqFB} \cdot \bar{v}_i^{pre} \cdot (u_j^{post} - w_{ji}^{out,FB}) \tag{4}$$

with the same components as in the bottom-up learning in Eqn.3. The bottom-up and top-down learning schemes differ in the definition of the competitive terms. As a consequence, the sum of weights $\sum_i w_{ji}^{out,FB}$ approach the mean activity u_j^{post} so that the sequence selective units v_i^{pre} memorize an expected pattern of their driving input configurations.

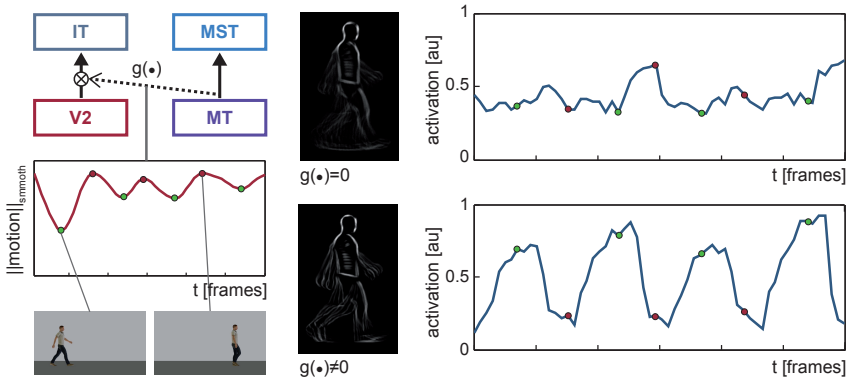

Fig. 2. IT prototypes trained using disabled and enabled reinforcement signal. Minima and maxima in motion energy correspond to articulated and non-articulated postures (bottom left). Continuous learning of IT prototypes leads to activation profiles with low selectivity (top right). Motion driven reinforcement leads to IT prototypes which signal snapshot poses in synchrony with the gait (bottom right; for details, see text).

3 Results

The model has been tested in various computational experiments, not all of which we can present here. In a first experiment, we probed the properties of snapshot selection from the input streams and their signature concerning static articulations. The latter property has been motivated by the fact that extremal articulation indicates configurations of implied motion, in turn, predictive for future motions. Results shown in Fig. 2 demonstrate that input activations in V2 with strongly articulated shapes cohere with local motion minima. Such minima

drive the reinforcement signal for learning whole body form prototypes. Temporal response signatures for IT prototypes are shown for disabled reinforcement, $g(m_e) = 0$, and when it is enabled, $g(m_e) \neq 0$. We studied the response properties of STS representations and their motion sequence selectivity. Initially, a prototypical sequence-selective representation is learned at the level of STS for a walker that is traversing from left to right. The input representations from the form and motion pathway were established in model IT and MST. The network is subsequently probed by three different movement scenarios: a forward moving walker with same profile and movement direction as in the training phase (*recall*), a forward moving walker traversing from right to left (*opposite*), and a backward moving walker (*reverse*). Form/motion prototypes and the sequence representation are triggered maximally in the *recall* case while in the *opposite* case form and motion prototypes only respond minimally, and so do the sequence-selective cells. In the *reverse* case the form prototypes selectively match the input at high articulation configurations, while the motion responses remain minimal. As a consequence, the sequence-selective representations respond at an intermediate level (Fig. 3). This evidence is in line with the experimental findings by [14] and recent observations by [16]. We have further investigated the direction and speed

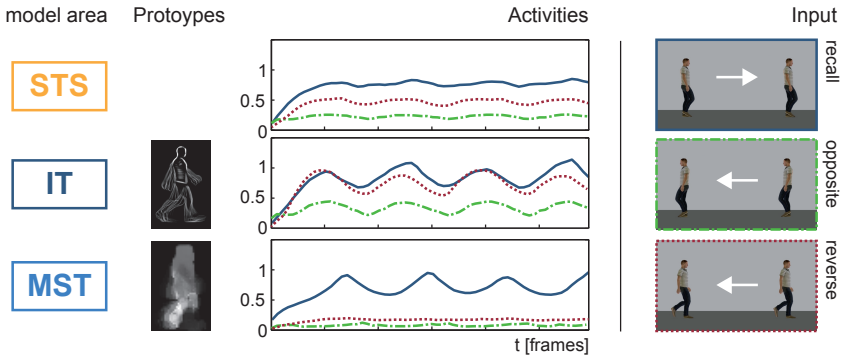

Fig. 3. Response behavior of IT snapshot neurons, MST motion pattern neurons, and sequence-selective STS cells trained by video input for a walker moving from left to right. Activations in the model areas are shown for different input conditions for recall of the training sequence (top), opposite walker movement (middle), and walker displayed in reverse motion (bottom). For details and brief discussion, see text.

tuning of the sequence-selective prototypes. Here, we configured different walkers with varying movement directions and speeds with reference to a previously learned representation of a rightward moving walker at a speed of 1 m/s. Walking directions in the test cases were rotated by $\pm\{5°, 10°, 20°, 40°, 80°\}$. Model simulations result in a direction tuning with half amplitude of approximately ± 45 deg. A similar tuning is observed for walkers proceeding at different speeds (results not shown).

In a further experiment we investigated the selective lesioning of the model architecture, particularly the effects of cutting connections between model areas and the activity flow between learned representations. The fully connected model with learned IT/MST and STS feedforward and feedback connections was used as reference. When bottom-up connections from motion input (MST) or from snapshot input (IT) were cut off the sequence-selective neuron responses in STS drop to approximately half their response amplitude. Feedback from STS invokes an amplification of activities in IT and MST representations. We observe that FF activation from IT alone can drive sequence neurons (in accordance with [1]). Snapshot representations in IT drive the STS sequence neurons which, in turn, send feedback signals to the stages of IT and MST prototype representations. In the motion pathway such feedback elicits an increase in presynaptic activation. We argue that this reflects the induction of increased fMRI BOLD response in human MT+ following the presentation of static implied motion stimuli [12].

4 Discussion

We propose a model for learning articulated motion patterns for animated motion recogntition. The model builds upon neurophysiological knowledge about the cortical sites and specific neuronal representations which contribute to articulated motion and implied motion perception. Based on feedforward/feedback processing in segregated pathways the form and motion prototypes are learned. Cross-channel interaction enables to select key poses in action sequences. We argue that such a mechanism is responsible for the development of snapshot representations corresponding to signatures of high articulation. Form and motion responses converge at the stage of STS to learn sequence-selective representations. Unlike previous approaches as, e.g., in [8, 5], we employ Hebbian learning of sequence-selective representations in STS by combining both form *and* motion, while snapshot and motion pattern prototypes do not employ individual sequence-selectivity.

Feedback connections learn to represent the expected input and, in turn, enable the network of IT/MST $\rightarrow$ STS and STS $\rightarrow$ IT/MST to predict the optical flow patterns and the associated snapshot sequences. As a result, temporal action sequences are represented (through learning) in a distributed network of recurrently connected sites (model IT, MST, and STS) to robustly recall the saliencies and regularities in presentations of articulated motions. The model predicts that the presentation of static key poses from previously learned sequences alone leads to enhanced activation in STS sequence selective neurons as observed in [10]. The model also hypothesizes how the presentation of static articulated poses lead to the emergence of predictive motion perception and enhanced neural activations in the motion pathway [12]. Next we will test the current model with input point-light stimuli as in biological motion perception. We predict that motion driven inputs will activate STS movement-selective neurons while responses in the form pathway will initially respond only at a minor level.

Feedback from STS cells will then enhance IT snapshot responses representing meaningful point configurations.

Acknowledgements. GL and HN have been supported by the Transregional Collaborative Research Centre "A Companion Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

References

[1] C.I. Baker, C. Keysers, T. Jellema, B. Wicker, and D.I. Perrett. Neuronal representation of disappearing and hidden objects in temporal cortex of the macaque. Exp. Brain Res. **140** (2001) 375–381

[2] N.E. Barraclough, D. Xiao, M.W. Oram, and D.I. Perrett. The sensitivity of primate STS neurons to walking sequences and to the degree of articulation in static images. Progress in Brain Res. **154** (2006) 135–148

[3] P. Bayerl and H. Neumann. Disambiguating visual motion through contextual feedback modulation. Neural Computation **16** (2004) 2041–2066

[4] Carpenter, G.A. and Grossberg, S. (eds.). Pattern recognition by self-organizing neural networks. Cambridge, MIT Press (1991)

[5] A. Casile and M.A. Giese. Roles of motion and form in biological motion recognition. ICANN/ICONIP 2003, Springer, LNCS 2714 (2003) 854–862

[6] M.-J. Escobar, G.S. Masson, T. Vieville, and P. Kornprobst. Action recognition using a bio-inspired feedforward spiking network. Intl. J. of Computer Vision **83** (2009) 284–301

[7] P. Földiak. Learning invariances from transformation sequences. Neural Computation **3** (1991) 194–200

[8] M.A. Giese and T. Poggio. Neural mechanisms for the recognition of biological movements. Nat. Rev. Neurosci. **4** (2003) 179–192

[9] T. Hansen and H. Neumann. A recurrent model of contour integration in primary visual cortex. J. of Vision **8**(8):8 (2008) 1–25

[10] T. Jellema and D.I. Perrett. Cells in monkey STS responsive to articulated body motions and consequent static posture: a case of implied motion? Neuropsychologia **41** (2003) 1728–1737

[11] H. Jhuang, T. Serre, L. Wolf, and T. Poggio A biologically inspired system for action recognition. Proc. IEEE 11th Intl. Conf. on Computer Vision, ICCV'07, Rio de Janeiro, Brasil, Oct.14-20 (2007)

[12] Z. Kourtzi and N. Kanwisher. Activation of human MT/MST by static images with implied motion. J. of Cognitive Neuroscience **12** (2000) 48–55

[13] E. Oja A simplified neuron model as a principal component analyzer J. of Math. Biology, **15** (1982) 267–273

[14] M.W. Oram and D.I. Perrett. Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey. J.l of Neurophysiology **76**(1) (1996) 109–129

[15] K. Schindler and L. Van Gool. Action snippets: How many frames does human action recognition require? Proc. IEEE Conf. on Computer Vision and Pattern Recognition, CVPR08, Anchorage, Alaska, USA, June 22-24, 2008

[16] J.M. Singer and D.L. Sheinberg. Temporal cortex neurons encode articulated actions as slow sequences of integrated poses. J. of Neuroscience **30**(8) (2010) 3133–3145