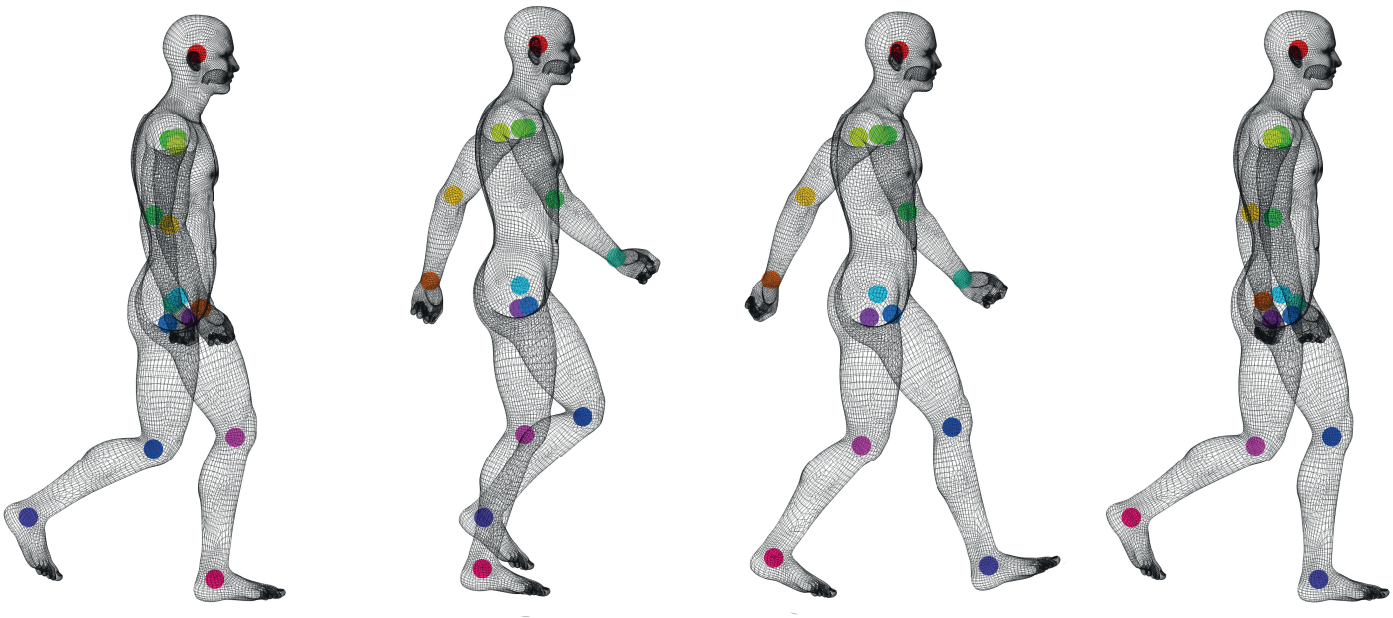
ulm university universität
uulm

Fakultät für Ingenieurwissenschaften, Informatik und Psychologie
Institut für Neuroinformatik

What, Where, and When?
Mechanisms of Learning Biological Motion Representations.

Dissertation zur Erlangung des Doktorgrades Dr. rer. nat. der Fakultät für Ingenieurswissenschaften, Informatik und Psychologie der Universität Ulm



Vorgelegt im Jahre 2018 von:
Georg Layher aus Ludwigsburg

Betreuer:
Prof. Dr. Heiko Neumann

Gutachter:
Prof. Dr. Heiko Neumann
Prof. Dr. Dr.-Ing. Wolfgang Minker

Amtierender Dekan:
Prof. Dr.-Ing. Maurits Ortmanns

Vorsitz Promotionsausschuss:
Prof. Dr.-Ing. Franz Hauck

Tag der Promotion:
05.06.2019

Abstract

The understanding and recognition of human actions is one of the major challenges for technical systems aiming at visual behavior analysis. Evidences from psychophysical and neurophysiological studies provide indications on the feature characteristics and neural processing principles involved in the perception of biological motion sequences. Modeling efforts from the field of computational neuroscience complement these empirical findings by proposing potential functional mechanisms and learning schemes enabling the establishment and recognition of biological motion representations and show how such principles can be transferred to technical domains.

First, results of psychophysical investigations are presented that demonstrate significant increases in the human recognition performance for motion (sub-) sequences containing highly articulated poses, which co-occur with local extrema in the motion energy and the extension of a body. Such key poses thus qualify as candidates to establish biological motion representations.

Second, based on these findings, a neural model for the learning of biological motion representations is presented. The model combines hierarchical feedforward and feedback processing along the ventral (form; what) and dorsal (motion; where) pathways with an unsupervised Hebbian learning mechanism for the learning of prototypical form and motion representations. More specifically, gated learning in the form pathway realizes the selective learning of highly articulated postures. Sequence selective representations are established using temporal association learning driven by motion and form input. The proposed model shows how the unsupervised learning of key poses can form the basis for the establishment of biological motion representations and gives a potential explanation for empirically observed phenomena, such as implied motion perception.

Third, as a transfer to technical application scenarios, a real-time biologically inspired action recognition system is presented which automatically selects key poses in action sequences and employs a deep convolutional neural network (DCNN) to learn class-specific pose representations. The network is mapped onto a neuromorphic platform, enabling the real-time (~1000 fps) and energy-efficient (~70 mW) assignment of key poses to action classes.

Last, it is shown how an associative learning scheme similar to the one applied in the neural model for the learning of biological motion representations can be used for the learning of visual category and subcategory representations. Here, instar learning is used to learn representations of visual categories, while outstar learning on the other hand is applied to establish representations of the expected input distribution. The category specific pattern is propagated back to the preceding stage where a residual signal reflecting the difference to the current input signal is derived. This difference is emphasized by modulation of the input with the residual signal and a subsequent normalization. If the difference is large enough, a new subcategory representation is established.

Acknowledgements

First of all I want to thank my supervisor Prof. Dr. Heiko Neumann for his exceptional supervision, his way of sharing his outstanding expertise with others, his always open mind, his time, his patience (I really carried it to the extremes) and his commitment and engagement in our common research activities and projects. The countless enlightening discussion we had – about and apart from scientific topics – will always influence the way I see the world. Thanks for some of the most interesting and exciting years I had.
I gratefully thank my very own biological motions – Susan, Nora, Mathilda and Imke – for their patience, their empathy, their energy and their love. You are what keeps me non-rigid and articulated.
Thanks to my co-workers, co-authors, colleagues and students who spend more than several days (and nights) with me discussing, implementing, designing and writing. Thanks Stefan, Michael, Felix, Martin, Martin A., Martin V., Tobias, Fabian, Stephan, Christian, Max, Yan, ...
Last but not least I want to thank the reviewers of this thesis, as well as the doctoral committee for the time and the effort they invested in this thesis.

The research presented within this thesis has been supported by the Transregional Collaborative Research Centre SFB/TRR 62 “A Companion Technology for Cognitive Technical Systems” funded by the German Research Foundation (DFG). I gratefully thank all my colleagues within the project for numerous valuable discussions and their continuous support.

Further, I gratefully acknowledge the support via a field test agreement between Ulm University and IBM Research Almaden, and particularly the TrueNorth team at IBM for their support. In addition, the support of NVIDIA Corporation with the donation of a Tesla K40 GPU used for parts of the presented research is also gratefully acknowledged.

Contents

Abstract **3**

Acknowledgements **4**

1 Introduction **7**

2 Articulated & Biological Motion - A Perceptual Study **11**
2.1 Introduction and Motivation 11
2.2 Main Contributions 11

3 Learning Representations of Articulated Motion Sequences - A Neural Model **15**
3.1 Introduction and Motivation 15
3.2 Main Contributions 15

4 Biologically Inspired Action Recognition from Key Poses **22**
4.1 Introduction and Motivation 22
4.2 Main Contributions 22

5 Learning of Category & Subcategory Representations - A Neural Model **26**
5.1 Introduction and Motivation 26
5.2 Main Contributions 27

6 Summary & Conclusion **30**

Bibliography **33**

I Individual Contributions & Publications **37**
Individual Contributions 37
List of Publications 43

II Articles of the Thesis **49**
[Layher and Neumann, 2018] 53
[Layher et al., 2014a] 67
[Layher et al., 2017a] 83
[Layher et al., 2014b] 107

Declaration of Authorship **127**

1 Introduction

The term *biological motion* was first introduced by Gunnar Johansson to describe the *"motion patterns characteristic of living organisms in locomotion"* [Johansson, 1973; p. 201]. From a perceptual perspective, biological motion could be defined as the visual appearance of moving animate – or living – beings. In contrast to the motion patterns observed for inanimate rigid objects, biological motion involves a deformation of the object's shape over time. After projection onto an image plane, both, the deformation of the body shape and the whole body movement (translatory/rotational) are reflected in static form and dynamic motion (optical flow) information in an image sequence. However, the question whether and how form and motion information contribute to the representation and recognition of biological motion sequences is yet not finally answered.

Figure 1.1 gives a naïve impression of how form and motion information might support a human observer in unraveling the nature of the displayed biological motion. In the first row, six static frames sampled from an image sequence showing a person performing a jumping jack are displayed. In the second row, the respective optical flow patterns are illustrated by encoding the direction and the speed through color value and color opacity. Examined in isolation, the six static images provide different indications on the performed action. Images 2, 4, and 6 (from left to right) directly imply that the observed subject is performing the action of jumping jack, whereas images 1 and 5 both support the conclusion that the subject is standing still. Incorporating the motion information displayed in the second row reveals the difference between these two snapshots. In the first image the subject has not started the execution yet, while in the fifth image the subject is moving upwards in between two repetitions of the jumping jack. Furthermore, relating the optical flow patterns in the second row to the static snapshots in the first (e.g. by contour assignment) directly discloses the non-rigidity of the displayed object in motion. The contour of the subject's body in the second picture, for instance, includes motion components in various different directions, directly providing valuable information on how the subject transits from one particular pose to another as part of the performed action.

Following these observations, it seems obvious that within a set of body pose configurations, *not all poses are equal* in their expressiveness with respect to the biological motion they constitute in succession. Even when presented as static images in isolation, some key poses allow the direct recognition of an action, while others are meaningless or highly ambivalent. Moreover, additional motion information allows to resolve ambiguities and provides indications on the non-rigidity and thus the animacy of the observed object in motion.

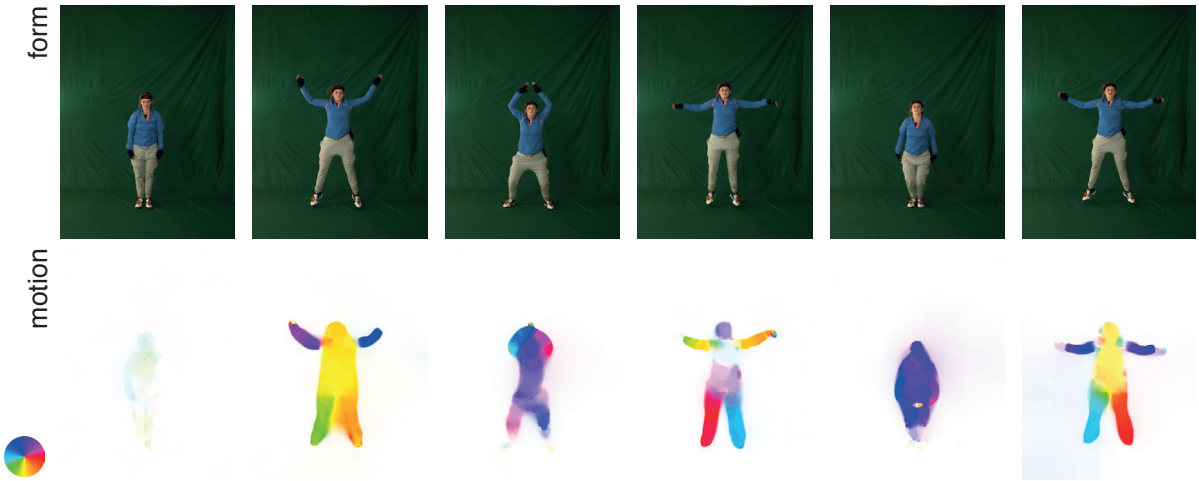


Figure 1.1: Form and motion information both contribute to the recognition of biological motion sequences. Six static images extracted from a image sequence showing a subject performing the action of jumping jack are displayed in the first row. In isolation, images 2, 4, and 6 (from left to right) directly imply the correct action, whereas snapshots 1 and 5 can both be perceived as showing a subject standing still. In the second row, the associated optical flow patterns are displayed color-encoded (direction: color value, speed: color opacity). Taking the motion information into account reveals that only in case of the first image, the subject is actually not moving. Written informed consent for the publication of exemplary images was obtained from the displayed subject. Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Articulated Motion for Social Signal Processing by Georg Layher, Michael Glodek, Heiko Neumann ©2017 [Layher et al., 2017b].

This, however, raises the following related questions:

- Do highly expressive poses share common characteristics which render them distinctive and allow a direct and robust recognition of a biological motion?
- How can highly expressive poses be identified within a biological motion sequence?
- How can representations which capture the defining features of highly expressive poses be learned?

Focusing on the feature characteristics and properties within an an observed biological motion sequence, these questions can be rephrased to yield core statements as inquiries for the investigations summarized within this article thesis:

What kind of body shapes can be observed in a biological motion sequence and how does the derived *form information* support the establishment of biological motion representations?

Where are the individual parts of the observed body pose moving to and how does the *motion information* about the transition between body poses contribute to the learning of biological motion sequences?

When should representations of the observed form and motion patterns be established and how can particularly salient and distinctive *key poses* be identified within a biological motion sequence?

In the first three of the four publications which form this article thesis, we approach these questions from three different perspectives, all contributing to a better understanding of the principles and mechanisms behind the learning and processing of biological motion representations. First, the results of an empirical study presented in [Layher and Neumann, 2018], which are briefly summarized in Chapter 2, provide evidence on the type of features which enable a significant increase of the human performance in recognizing biological motion sequences displayed as *point-light stimuli* (PLS). A novel masking technique is proposed which reduces the influence of high level processes on the perception of the displayed PLS, such as skeleton fitting or the utilization of a priori knowledge about the performed action. Replicating and extending previous findings of Thirkettle et al., 2009, the reported results show that the human recognition performance is strongly correlated to the lateral and longitudinal extension of the displayed body poses. Due to the nature of human kinetics, the body extension in turn is correlated to the motion energy profile of an observed biological motion sequence. Local extrema in specifically tuned feature channels, such as the body extension and the motion energy, might be candidates to identify moments of highly expressive and distinctive body poses which enable a high performance in recognizing biological motion (sub-) sequences.
Second, a neural model for the unsupervised learning of biological motion representations is presented in [Layher et al., 2014a] and briefly recapitulated in Chapter 3. The model combines hierarchical feedforward and feedback processing along the ventral (form; what) and dorsal (motion; where) pathways with a component of correlation-based unsupervised Hebbian learning for the establishment of prototypical form and motion representations. This component is augmented by a gating mechanism driven by a signal derived from the motion pathway to steer the form-based learning of snapshot representations. Together, the learning is realized as a three-factor rule of Hebbian weight adaptation. The proposed interaction between the motion and the form pathway gives a potential explanation of how representations of highly expressive poses (as observed in the perceptual study) might be learned by unsupervised adaptation mechanisms. Sequence selective representations are established by combining a feedback mechanism with temporal association learning driven by the activities of the prototypical representations in the form and motion pathways. The proposed feedback mechanism potentially gives an explanation for the functional mechanism behind activities in motion sensitive cortical areas measured in physiological studies while static images of highly articulated poses were presented (implied motion) [Kourtzi and Kanwisher, 2000].
Third, in [Layher et al., 2017a], parts of the processing principles and mechanisms summarized in Chapter 2 and Chapter 3 are transferred to a technical domain and adapted to realize a real-time action recognition system. A feature-driven mechanism is implemented to automatically select highly articulated key poses within an image stream by evaluating the lateral and longitudinal extension of an observed body around local temporal extrema in the motion energy (in accordance to the findings summarized in Chapter 2). The proposed selection principle is in line with the feature characteristics identified in the perceptual study presented in [Layher and Neumann, 2018] and reflects the interaction between form and motion processing pathways described in [Layher et al., 2014a]. A deep convolutional neural network (DCNN) is employed

to learn pose representations of combining form and motion information. The trained network is mapped onto the IBM Neurosynaptic System platform, which enables the real-time (~1000 fps) and energy-efficient (~70 mW) assignment of key poses to action classes. The proposed action recognition system was trained and evaluated on two different datasets. The results show, that the recognition performance of the proposed approach is on par with other key pose based methods described in the literature. In Chapter 4, a brief summary of the proposed action recognition system is provided.

Together, the three articles summarized in Chapters 2-4 demonstrate how the results of an empirical study and a modeling effort complement each other and how an application in a technical domain can benefit from the gained insights about the observed principles and the proposed processing mechanisms.

In [Layher et al., 2014b], the last of the four articles which constitute this article thesis, the learning scheme applied in [Layher et al., 2014a] is adapted for the unsupervised learning of visual category and subcategory representations. In a three-layered architecture, instar learning is used to establish representations of visual categories, while outstar learning is applied to learn representations of the expected average input distribution. The category specific pattern is propagated back to the preceding stage where a residual signal reflecting the difference to the current input signal is derived. This difference is emphasized by modulation of the input with the residual signal and a subsequent normalization. If the difference between the category representation and the input distribution is large enough, a new subcategory representation is established. Overall, the model network demonstrates how gross category representations can be established which are further spezialized and fine tuned in an adaptive fashion. By combining modulatory feedback in the spirit of biased competition with mechanisms of predictive coding the formation of subcategorial representations is achieved purely stimulus driven without any teaching signal or user interaction. The proposed recurrent network architecture is briefly summarized in Chapter 5.

After a concluding summary and a contextualization of this thesis in Chapter 6, a description of the individual contributions of the articles' authors, as well as a full list of publications is provided in Chapter I. For a fully detailed description of the contributions summarized in Chapters 2-5 refer to the copies of the original articles included in Chapter II of this thesis. The article thesis is organized using arabic numbers to structure the summaries of the articles in the overall context (Chapters 1-6) and roman numbers to subdivide the individual contributions, the list of publication, as well as the original articles into chapters (Chapter I and II).

2 Articulated & Biological Motion - A Perceptual Study

Within the following chapter, the contributions presented in [Layher and Neumann, 2018] to the overall context of this thesis are briefly summarized. In Section 2.2 figures extracted from [Layher and Neumann, 2018] are included.

2.1 Introduction and Motivation

Experimentally analyzing the contributions of form and motion information on the perception of biological motion sequences requires different visual parameters of the motion to be isolated from each other. In his pioneering work, Johansson proposed a method of reducing the display of biological motion stimuli to a set of spots moving in coherence with the joints of a subject's skeleton [Johansson, 1973] (see Figure 2.1 for an example). Perceptual experiments showed that human observers are capable of recognizing biological motions on the basis of such *point-light stimuli* (PLS) within temporal windows equal or greater to 200 ms. Stimulus onset intervals of larger than 400 ms even allowed the subjects to reliably discriminate different types of motions presented as PLS [Johansson, 1976]. Since PLS contain both motion and – at least structural – form information, the debate on the contributions of form and motion information to the observed recognition performance is still ongoing [Garcia and Grossman, 2008]. Over the last decades, evidence from experimental studies indicated a strong influence of motion based features [Mather and Murdoch, 1994], as well as highlighted the importance of structural form information [Hiris, 2007; Lange and Lappe, 2006]. In [Layher and Neumann, 2018] further evidences on the characteristics of the features underlying biological motion representations, as well as indications on the learning principles are presented. The central contributions of the conducted perceptual study to the overall context of this thesis are briefly summarized below. The detailed description of the proposed stimuli, the experimental design, and the results, are included into this thesis on page 53 [Layher and Neumann, 2018].

2.2 Main Contributions

Thirkettle et al., 2009 showed that human observers are capable of distinguishing a PLS of a walking human and a control stimulus even when only 50 ms sequences are presented, but that the average recognition performance is strongly dependent on the lateral extension of the

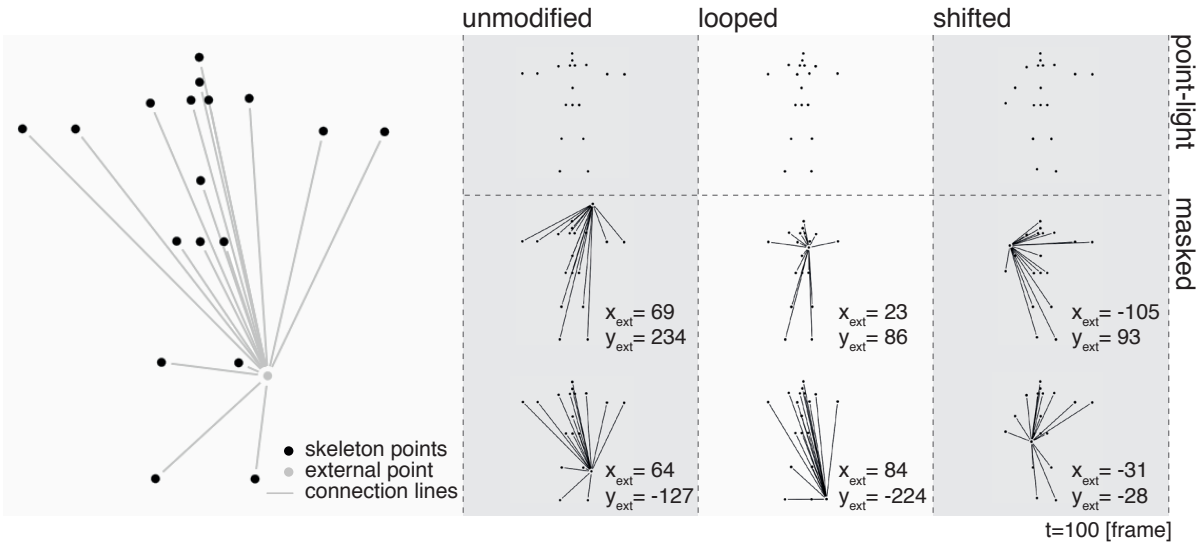


Figure 2.1: Masking biological motion point-light stimuli – a novel approach. PLS stimuli were generated on the basis of motion capture data averaged among multiple subjects performing several jumping jack iterations. Masking was applied by connecting all points of the PLS to an additional static external point (drawn in gray for better visibility; within the experiments, all point and lines were drawn in white on a black background). Three variations of the original jumping jack PLS were analyzed experimentally (second to forth column). The original point-light stimuli (first row) are displayed alongside with two corresponding masked versions (second and third row). Adapted from [Layher and Neumann, 2018], CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

displayed point patterns. However, the correlation of the lateral extension and the recognition performance could not be reproduced for a PLS displaying a human performing a jumping jack. In this case, the reported recognition performance is almost constant at ceiling (100 %).
In [Layher and Neumann, 2018] a novel approach of masking point-light stimuli is proposed with the intention of minimizing the influence of symmetry effects, figure-ground segregation, as well as a priori knowledge about the nature of the motion. The masking is achieved by adding one stationary external point to the stimulus and radially connecting it to all points of the PLS. The connection lines end with a short distance to the PLS points, such that the local dot motion patterns are unaltered compared to the original PLS. Figure 2.1 displays one frame of an unmodified jumping jack PLS (second column; first row) alongside with two masked PLS of the same frame using different external points (second column; second and third row).
Following the experimental routine described in Thirkettle et al., 2009, subjects were asked to distingush 50 ms subsequences of point-light stimuli from phase-scrambled controls for four different stimulus configurations. After the experiment, roughly two thirds of the subjects were not able to name the kind of motion which was contained in the masked stimuli, while the remaining third identified the motion as *biological* or *human*.
For an unmasked jumping jack PLS, the ceiling effect reported in [Thirkettle et al., 2009] could successfully be reproduced. In contrast, the results show a strong correlation of the human recognition performance and the lateral extension extension of a point-light pattern for

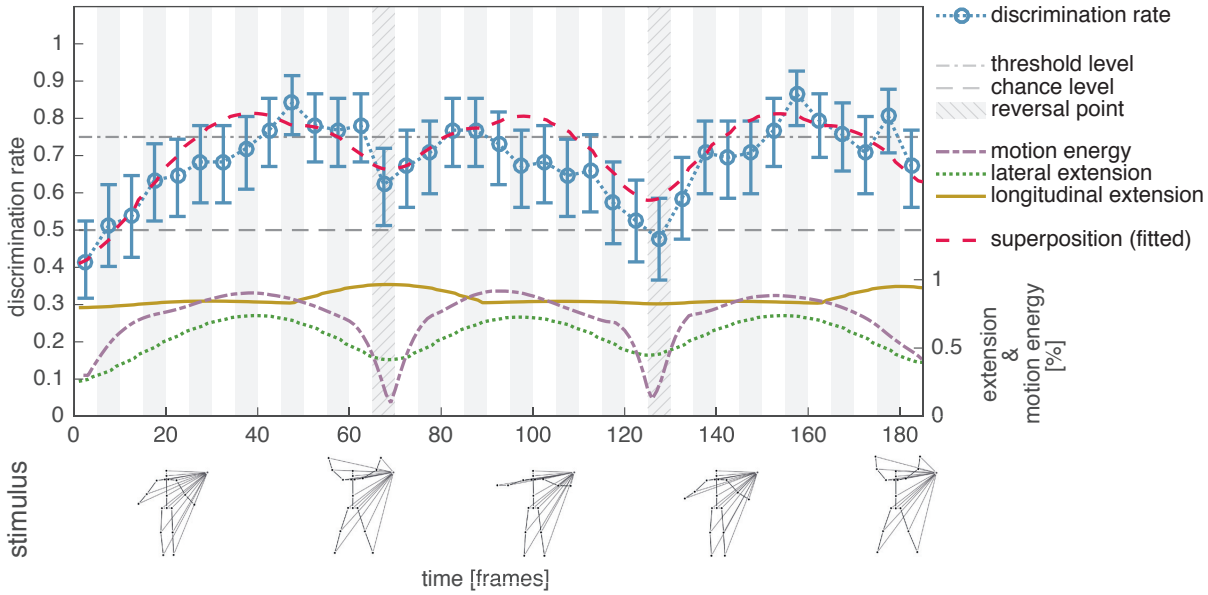


Figure 2.2: Correlation of the average human recognition performance and the extension of a point-light pattern. A masked jumping jack PLS sequence of 187 frames length was divided into 37 subsequences of 5 frames length and 41 subjects were asked to discriminate the subsequences from phase-scrambled controls. The average discrimination rate of 41 subjects is plotted (dotted blue) together with the corresponding 95% confidence intervals. The normalized lateral (dotted green) and longitudinal (solid yellow) extensions are plotted alongside with the overall motion energy (dashed purple). The superposition of the two extensions strongly correlates to the human discrimination rate ($r = 0.83,\ p < 0.001$). Adapted from [Layher and Neumann, 2018], CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

the masked jumping jack PLS. Further, an increase of the recognition performance with an increasing longitudinal extension of the stimulus can be observed. The strongest correlation was obtained for a measure combining the lateral and the longitudinal extension (see "superposition" in Figure 2.2). The results of the four experiments conducted within the study support the conclusion, that local temporal maxima in (potentially a variety of) feature channels form the basis for an increase in the perceptual salience of a biological motion (sub-) sequence. Thus, the (lateral and longitudinal) body extension is one candidates which qualifies as feature channel for the selection of key poses within biological motion sequences. Note that due to the nature of human kinetics, local extrema in the motion energy co-occur with extrema in the lateral and longitudinal extension of the body pose.

In short, the key contributions presented in [Layher and Neumann, 2018] comprehend:

- A novel approach of masking point-light stimuli to reduce the influence of symmetry effects, figure-ground segregation and a priori knowledge about the kind of motion on biological motion interpretation or recognition.
- Empirical evidence for the dependence of human performance in recognizing biological point-light stimuli on the lateral and longitudinal extension of a point-light pattern.
- Empirical evidence, that temporal maxima in spatial feature channels increase the perceptual salience of biological motion (sub-) sequences and are candidates to trigger the establishment of stationary snapshot representations to encode biological motions.

3 Learning Representations of Articulated Motion Sequences - A Neural Model

Within this chapter, the contributions presented in [Layher et al., 2014a] to the overall context of this thesis are briefly summarized. In Sections 3.1 and 3.2, text passages and figures extracted from [Layher et al., 2017b] are adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Articulated Motion for Social Signal Processing by Georg Layher, Michael Glodek, Heiko Neumann ©2017 [Layher et al., 2017b].

3.1 Introduction and Motivation

The early processing stages of visual cortical processing of biological motion patterns is performed along two major processing streams, namely the ventral *"what"* and the dorsal *"where"* pathway (see Figure 3.1). The two pathways later converge into fused representations at the level of the superior temporal sulcus (STS; [Boussaoud et al., 1990; Felleman and Van Essen, 1991]). Functional imaging studies [Grossman et al., 2000], as well as single-cell recordings [Oram and Perrett, 1994] indicate the existence of specific mechanisms for the processing of biological motion within this area STS. Further evidence suggests that cells in STS integrate form and motion representations of biological objects [Oram and Perrett, 1996b] and selectively respond to face, limb and whole body motion [Puce and Perrett, 2003].

In [Layher et al., 2014a], a model is presented which proposes functional mechanisms and processing principles underlying the properties observed in neurophysiological, as well as psychophysical experiments. In particular, the model is in line with the perceptual phenomena observed in the psychophysical study presented in [Layher and Neumann, 2018] and summarized in Chapter 2. In the following, the key properties of the proposed model and learning mechanisms are briefly summarized. A detailed description is included into this thesis on page 67 [Layher et al., 2014a], as well as [Layher et al., 2017b].

3.2 Main Contributions

In [Layher et al., 2014a] a biologically inspired model for the learning of articulated motion representations is presented. The model extends the work of Giese and Poggio, 2003 by

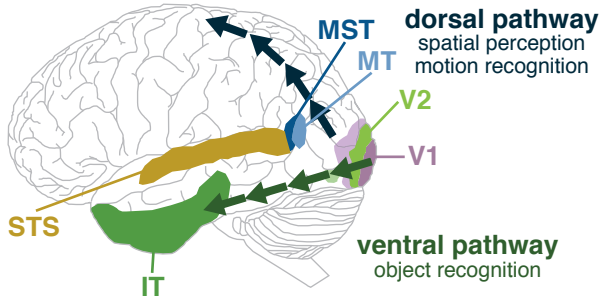


Figure 3.1: Both, form and motion information both contribute to the neural processing of biological motion. The ventral (form) and the dorsal (motion) pathways converge to the superior temporal sulcus complex, where there is evidence that biological motion specific representations are built. Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Articulated Motion for Social Signal Processing by Georg Layher, Michael Glodek, Heiko Neumann ©2017 [Layher et al., 2017b].

combining it with mechanisms for the hierarchical feedforward and feedback processing of motion and form information along the dorsal and ventral pathway [Bayerl and Neumann, 2004; Weidenbacher and Neumann, 2009]. In Figure 3.2, the overall structure of the model is depicted. Characteristic prototypes of intermediate level form representations in the inferior temporal gyrus (model area IT) and motion patterns in the medial superior temporal area (model area MST) are learned using a Hebbian learning scheme. In addition, a motion dependent reinforcement mechanism is used to regulate the learning of prototypical form representations. This prototypical learning scheme is inspired by empirical findings. Psychophysical and physiological experiments suggest that body poses which show a high degree of articulation allow a higher performance in the recognition of human actions [Layher and Neumann, 2018; Thirkettle et al., 2009; Barraclough et al., 2006] and evoke the percept of implied motion as well [Kourtzi and Kanwisher, 2000]. Note the correlation between the experimentally observed human recognition performance and the (lateral and longitudinal) body extension reported in Chapter 2, Figure 2.2 and compare the static snapshots displayed in Chapter 1, Figure 1.1. In [Layher et al., 2014a], a simplistic but generic approach that selects such characteristic poses out of continuous streams of video sequences in an unsupervised fashion is proposed. In case of such selections, learning is enabled to build representations which preferentially focus on poses with a high degree of articulation. Both pathways converge in model complex STS, where sequence-selective representations are built using a combined Hebbian bottom-up and top-down learning [instar/outstar; Carpenter, 1989] alongside with a short term memory trace, realizing a temporal associative memory. 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 optical flow patterns (areas MT and MST). The membrane potential of individual model neurons is calculated by conductance-based mechanisms of feed-forward integration of excitatory and inhibitory feeding input and a passive leakage. The potential can be enhanced by a gating mechanism to amplify the efficacy of the current potential by a matching top-down feedback signal. The membrane potential is finally regulated by a gain control mechanism that leads to activity normalization for a pool of neurons through mutual divisive inhibition (shunting inhibition, compare [Carandini et al., 1999]). These mechanisms are summarized in

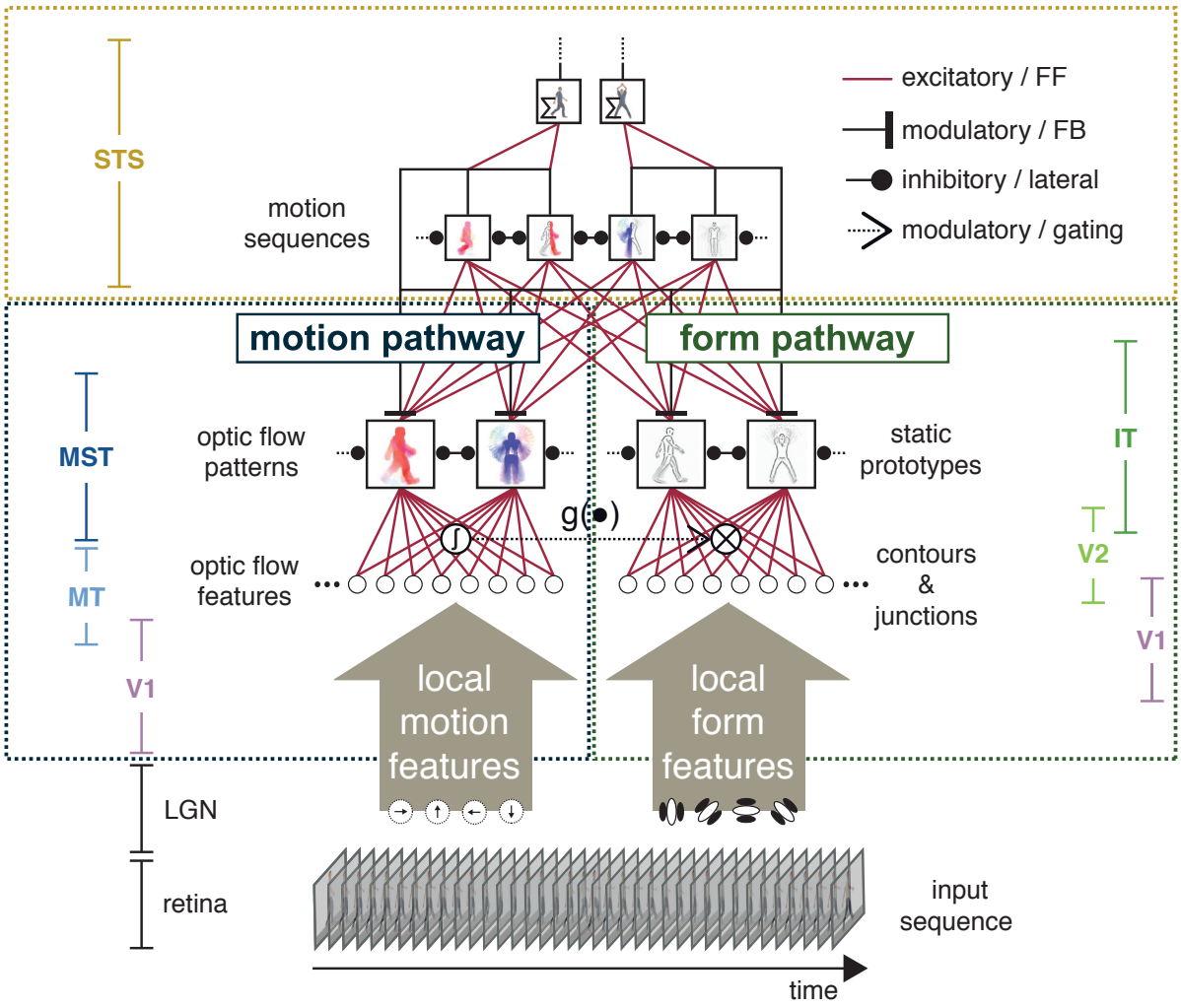


Figure 3.2: Model overview. The model consists of two separate processing pathways, which both converge in model complex STS. Separate motion and form pattern representations are established in the (dorsal) motion and the (ventral) form pathway. At the level of STS, combined representations of MST and IT activities are learned via an instar/outstar learning scheme in combination with a temporal trace rule. The weights of afferent connections are used to recognize the closest matching form and motion pattern. The weights of efferent connections are fed back to preceding stages realizing a prediction mechanism. In addition, in [Layher et al., 2014a] an interaction between the pathways at the level of MST/IT is suggested, which enables the selective learning of key poses in the form pathway. Adapted by kind permission of John Wiley and Sons and the Cognitive Science Society from [Layher et al., 2014a].

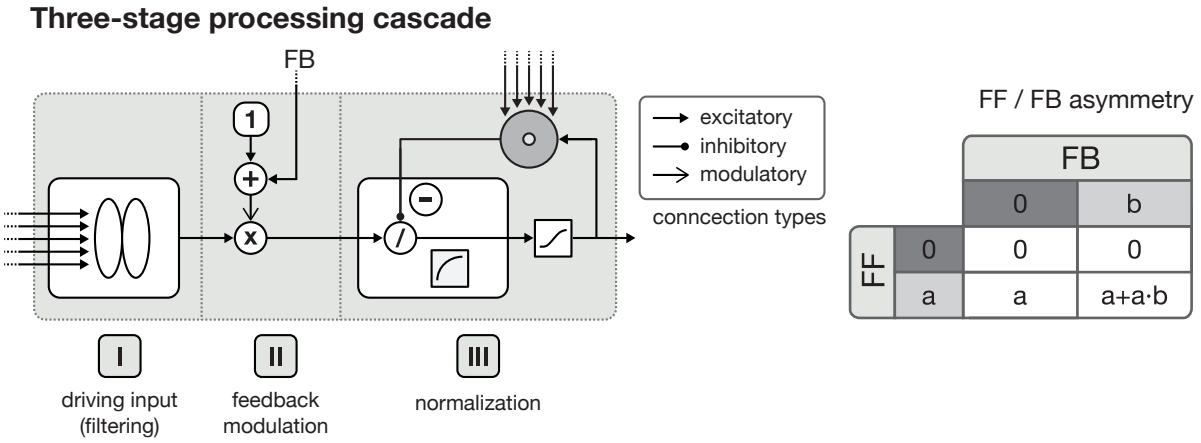


Figure 3.3: Three-stage model cascade. The cascade comprises an initial input filtering (stage I), the modulation of the activity(stage II) and a final pool normalization (stage III). Re-entrant feedback from higher level areas is incorporated in stage II where the current activity is modulated by $(1 + \text{net}_{FB})$. This kind of feedback integration is essential, since it results in an asymmetry of the roles the feedforward and the feedback signals play in the signal processing. As illustrated in the table on the right-hand side, without the presence of a feedforward signal, a feedback signal cannot evoke any activity. Adapted from [Layher et al., 2014b], CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

a three-stage hierarchy of processing that includes input filtering, modulatory feedback, and pool normalization (see Figure 3.3). The output of a cell is defined by a signal function which converts the membrane potential into a firing rate, or activity. Such model cells are grouped into layers which form abstract models of cortical areas.The computational properties of the model neurons and the three-stage cascade have been reported in, e.g., [Hansen and Neumann, 2008; Bayerl and Neumann, 2004; Bouecke et al., 2011; Raudies et al., 2011]. The representations that are learned in the segregated form and motion pathways define the input activities for model STS cells. Based on [Jellema et al., 2004] we suggest that form and motion activities converge to drive cells in categorial representations of temporal movement selectivity.

The Giese-Poggio model [Giese and Poggio, 2003] 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, in [Layher et al., 2014a] a mechanism is suggested that automatically selects snapshots that correspond to strongly articulated poses. Such snapshot representations are learned in the form pathway (model area IT) by utilizing a gating reinforcement signal which is driven by the complementary representation of motion in the motion pathway (model areas MT/MST, see the dashed arrow in Figure 3.2). The motion energy signal itself is a function of time which is used to steer the instar learning in the form pathway. In the proposed learning scheme, snapshot poses correspond to *highly* articulated postures with signatures of maximum limb spreading, which share the characteristic of a large lateral and longitudinal extension as described in Chapter 2. The overall motion energy at the limbs drops during phases of high articulation when their apparent direction of motion reverses. This, in turn leads to local minima in the overall motion energy of an articulating actor which coincide with the above-mentioned events of articulated

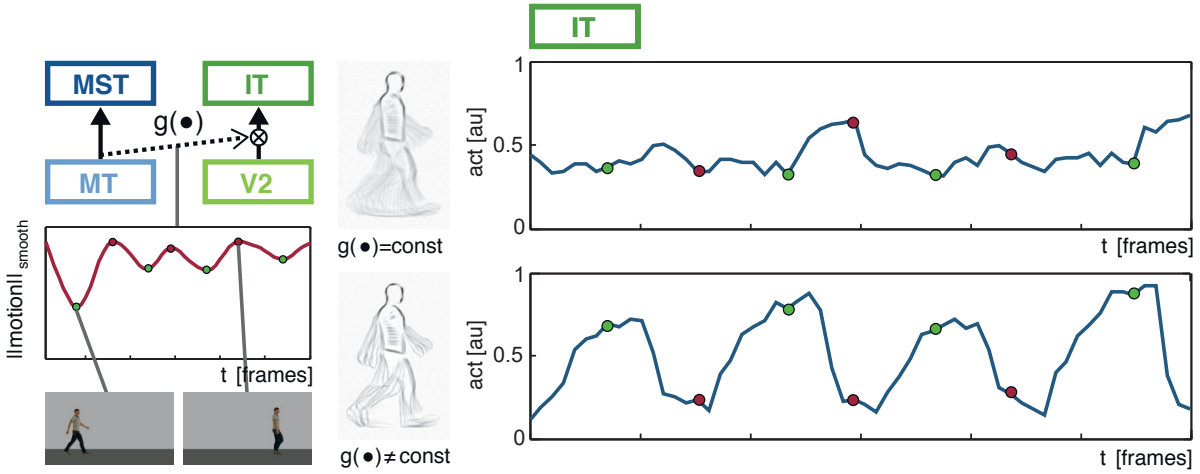


Figure 3.4: Unsupervised learning of key body postures. The motion energy derived within the motion pathway is used to gate the learning of the form prototypes. Local minima in the motion energy signal correspond to body postures containing a high degree of articulation (see plot on the left and compare to Figure 2.2 in Chapter 2). The weight of those postures is amplified during the learning process. Without incorporating the gating signal $g(\bullet)$, the statistical mean of the whole sequence is learned (on the top right). If $g(\bullet)$ is applied on the learning as a weighting function, a prototype representation selectively responding to highly articulated poses is established (bottom right). Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Articulated Motion for Social Signal Processing by Georg Layher, Michael Glodek, Heiko Neumann ©2017 [Layher et al., 2017b].

key poses. Due to the nature of human kinetics, local extrema in the motion energy correspond to local extrema in the lateral/longitudinal extension of a body pose. Thus, the gating signal is capable selectively reinforcing postures with a high perceptual salience, such as reported in e.g. [Layher and Neumann, 2018; Thirkettle et al., 2009]. Figure 3.4 displays the activity profiles of prototypical form representations in model area IT with (second row) and without (first row) using the motion energy signal to gate their establishment. In [Layher et al., 2012; Layher and Neumann, 2013; Layher et al., 2014a; Layher et al., 2017b], several experiments were performed to highlight the contributions of the individual components in the model architecture. The input configurations included a sequence containing both form and motion data (original), a point light sequence of a walker (PLW), an articulated and a not articulated static posture, as well as a sequence showing a person performing a jumping jack (different). Figure 3.5 shows the activations of the form and motion prototypes at model areas MST and IT, as well as the sequence-selective cells (model area STS) after learning. The sequence selective representations at STS respond at a maximum level to the input they were trained on (original). In contrast, they respond at their minimum when probed with a different motion sequence (different), showing that the model is capable of learning and distinguishing different articulated motions. In line with experimental evidence, the sequence-selective cells at model STS show increased activities either if just motion (point-light-walker; see, e.g., [Beauchamp et al., 2003]) or solely form (articulated pose) is presented to the network (compare [Barraclough et al., 2006; Jellema and

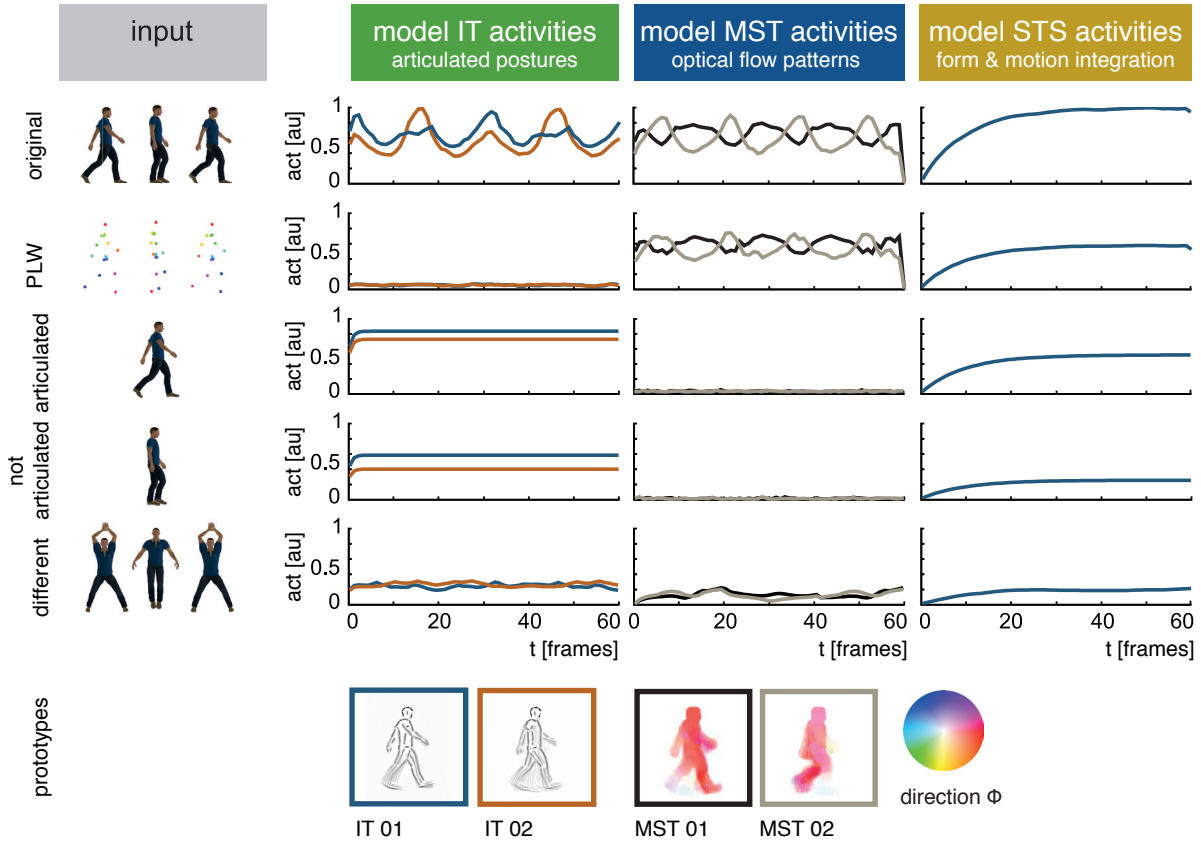


Figure 3.5: Model Simulations. Artificially generated image sequences are used to demonstrate the network responses to different input configurations (first column). The last row shows two form and two motion representations (at model areas IT and MST) which were established during the training. First, the training data was used as input, resulting in activities at their maximum in all model areas (original). Note that the activity of model IT (second column) selectively increases when a high degree of articulation is present. Second, the network was tested with a point-light walker sequence, generated on basis of the same articulated motion as in the original case (PLW). As expected, the activities of model IT cells almost drop to zero. The activities of model MST (third column) cells, on the contrary, stay, at a level comparable to the original case and in combination with the feedback trigger a medium level excitation of the model STS cell. The third and the fourth column, shows the results of single pictures being continuously presented to the network as a static input. Two different static images were used, one containing a highly articulated pose and the other barely showing any articulation. The activity of the model IT, as well as STS cells is clearly increased if the sequence shows an articulated posture. The fifth row shows how the activities drop throughout all model areas when a different motion seqeunces is used as input and thus, the models capability of distinguishing different articulated motion sequences. Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Articulated Motion for Social Signal Processing by Georg Layher, Michael Glodek, Heiko Neumann ©2017 [Layher et al., 2017b].

Perrett, 2003]). In line with [Layher and Neumann, 2018; Thirkettle et al., 2009], model STS cells respond at a higher level for input pictures showing an articulated posture compared to input images containing a not articulated pose. In short, the contributions and propositions presented in [Layher et al., 2014a] include:

- A model architecture for the processing of biological motion sequences consisting of two separate processing pathways for form and motion information which both converge in model area STS.
- An unsupervised extended Hebbian learning scheme for the establishment of form and motion representations in model areas IT and MST.
- A mechanism for the automatic selection of highly articulated poses within a biological motion sequence through the extension of the basic Hebbian learning into a 3-factor-learning rule.
- An interaction between the motion and the form pathway to reinforce the learning of highly articulated pose representations in model area IT defining the third factor in the extended learning rule.
- Simulations showing that the applied processing and learning principles are capable of reproducing data obtained from psychophysical and physiological studies.
- The combination of an unsupervised instar/outstar learning scheme with a temporal trace-rule and a feedback mechanism for the establishment of sequence-selective representations in model area STS.

4 Biologically Inspired Action Recognition from Key Poses

In this chapter, the relation of the biologically inspired action recognition approach presented in [Layher et al., 2017a] to the overall context of this thesis is briefly summarized. In Sections 4.1 and 4.2, text passages and figures extracted from [Layher et al., 2017a] are included.

4.1 Introduction and Motivation

Recognizing actions in the surrounding area is an integral part of interpreting the own situative context and environment, and thus is in particular crucial for future technical systems which may find themselves embedded in a variety of different situations. In computer vision, the term vision-based action recognition summarizes approaches to assign an action label to each frame or a collection of frames of an image sequence. The key pose based action recognition approach proposed in [Layher et al., 2017a] is functionally motivated and inspired by the evidences about the learning mechanisms and representations involved in the cortical processing of articulated motion sequences presented in [Layher and Neumann, 2018] and [Layher et al., 2014a]. Also, recent hardware and software developments from various fields of visual sciences further motivate the presented contribution from a technical perspective. In the following, the key characteristics of the proposed action recognition approach are summarized. A detailed description is included into this thesis on page 83 [Layher et al., 2017a].

4.2 Main Contributions

In [Layher et al., 2017a] a feature-driven key pose selection method is proposed, which is driven by combining two features in the biological motion input, namely the extrema in the temporal motion energy signal and the relative extension of the observed subject's pose. Such temporally defined features (from the motion stream) help to automatically select key pose representations. The use of these dynamic features has been motivated by the psychophysical investigations summarized in Chapter 2 and presented in [Layher and Neumann, 2018], which demonstrate that humans select specific poses in a continuous sequence of video input based on such criteria. In Figure 4.1, the relative lateral and longitudinal extension of a human performing a jumping jack are displayed alongside with the relative motion energy (compare Figure 2.2 and Figure 3.4). Key pose candidates are first selected by identifying local extrema in the

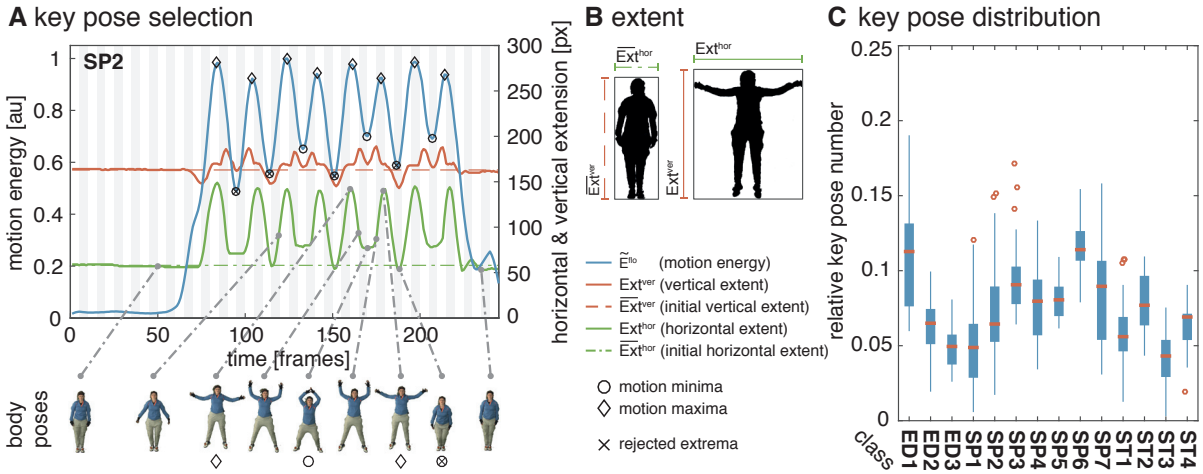
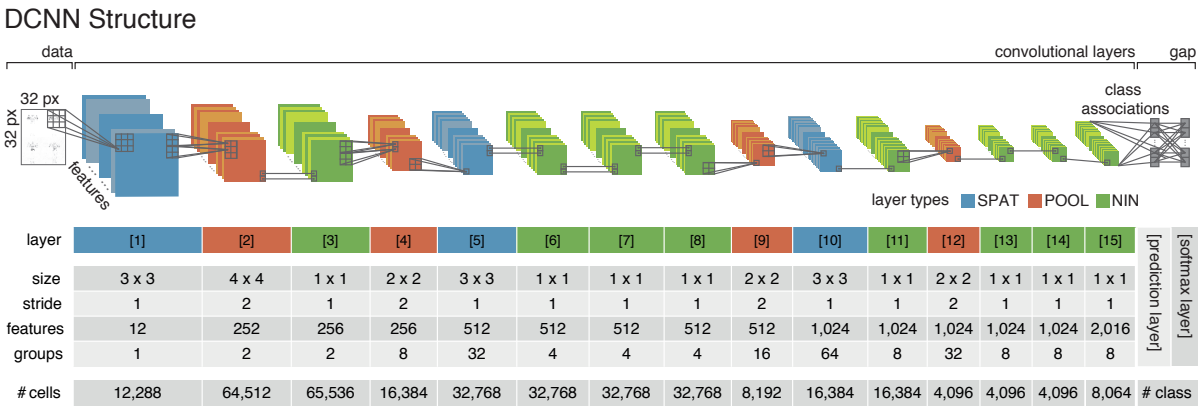


Figure 4.1: Selection of key poses. For each action sequence, key pose frames are selected by identifying extrema in the motion energy which co-occur with a sufficiently increased extent of the body pose. In (A), the smoothed motion energy (blue) together with the horizontal (green) and vertical (red) extent of the body pose are displayed for a subject performing the action of jumping jack. Local minima in the motion energy are marked by a circle, local maxima by a diamond. Extrema which are rejected as a key pose frame because of an insufficient extension are additionally marked with x. (B) shows an example for the horizontal and vertical extension of a neutral and a highly articulated body pose. (C) shows the relative number of identified key poses per action sequence for the uulmMAD dataset [Glodek et al., 2015] used for the simulations. Note that the relative number per action significantly varies. Written informed consent for the publication of the exemplary images was obtained from the displayed subject. Adapted from [Layher et al., 2017a], CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

motion energy and potentially rejected if neither the horizontal nor the vertical extension are above an a priori defined threshold level. The number of selected key poses per action can strongly vary per action class (see Figure 4.1 **C**). It is important to note, that the key pose selection is purely feature-driven and does not not utilize any information about the inter- or intra-class distribution of the key poses (or key pose descriptors). Thus the selection of the key poses is independent of the number and type of other actions within the dataset. Contour and optical flow data extracted from the automatically selected key poses is used as input to learn action specific representations combining form and motion information by training a deep convolutional neural network (DCNN). The preprocessing builds upon coarsely segregated streams of sensory processing along different pathways which separately process form and motion information analogous to the model architecture described in Chapter 3 and proposed in [Layher et al., 2014a].

An energy efficient deep convolutional neural network (Eedn; [Esser et al., 2016]) was employed to realize the key pose learning and classification, which achieves a computationally efficient solution using a sparse and energy efficient implementation based on neuromorphic hardware. This allowed to establish a cascaded hierarchy of representations with an increasing complexity for combined key pose form and motion patterns. After their establishment, key pose



layer	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[prediction layer]	[softmax layer]
size	3 x 3	4 x 4	1 x 1	2 x 2	3 x 3	1 x 1	1 x 1	1 x 1	2 x 2	3 x 3	1 x 1	2 x 2	1 x 1	1 x 1	1 x 1		
stride	1	2	1	2	1	1	1	1	2	1	1	2	1	1	1		
features	12	252	256	256	512	512	512	512	512	1,024	1,024	1,024	1,024	1,024	2,016		
groups	1	2	2	8	32	4	4	4	16	64	8	32	8	8	8		
# cells	12,288	64,512	65,536	16,384	32,768	32,768	32,768	32,768	8,192	16,384	16,384	4,096	4,096	4,096	8,064	# class	

Figure 4.2: Layers of the deep convolutional neural network. The implemented DCNN employs three different convolutional layer types (layers 1–15). Spatial layers (SPAT; blue) perform a linear filtering operation by convolution. Pooling layers (POOL; red) decrease the spatial dimensionality while increasing the invariance and diminishing the chance of overfitting. Network-in-network layers (NIN; green) perform a parametric cross channel integration. The proposed network consists of a data (or input) layer, 15 convolutional layers and a prediction and softmax layer on top. The parameters used for the convolutional layers of the network are given in the central rows of the table. In the last row, the number of artificial cells per layer is listed. Adapted from [Layher et al., 2017a], CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

representations allow an assignment of a given input image to a specific action category. An offline training scheme that utilizes a deep convolutional neural network with 15 convolutional layers was applied. Figure 4.2 gives an overview over the DCNN topology and briefly explains the different implemented layers. The trained network runs on IBM's TrueNorth chip [Merolla et al., 2014; Akopyan et al., 2015]. This solution renders it possible to approach faster than real-time capabilities for processing input streams and classify articulated still images at about 1, 000 frames per second while the computational stages consume only about 70 mW of energy. The proposed network was trained on key poses extracted for the 14 classes of the uulmMAD dataset [Glodek et al., 2015] and evaluated using different temporal integration schemes of single frame classification results. Figure 4.3 **A** displays the results of a framewise classification scheme that allows to recognize an action in an instant when the key pose frame is presented to the network. This kind of immediate decision might be crucial for systems which rely on decisions in real time. In a leave-one-subject-out (LOSO) cross-validation analysis, the proposed network achieved an average recall of 0.887 (std 0.057). In Figure 4.3 **B**, the results of a full sequence majority-voting scheme are summarized. Here, the final classification is performed after evaluating all key pose frames within an action sequence and determining the class with the most frequent occurrence of key poses. The average recall obtained by the LOSO cross-validation procedure increases to 0.967 (std 0.028). Finally, the results of accumulating single key pose decisions within sliding windows of increasing size is summarized in 4.3 **C**. In addition, the performance of the network was analyzed on the widely used Weizmann dataset to allow a comparison to other approaches. A cross-dataset evaluation between the overlapping classes of the the two datasets still showed a promising performance of the network.

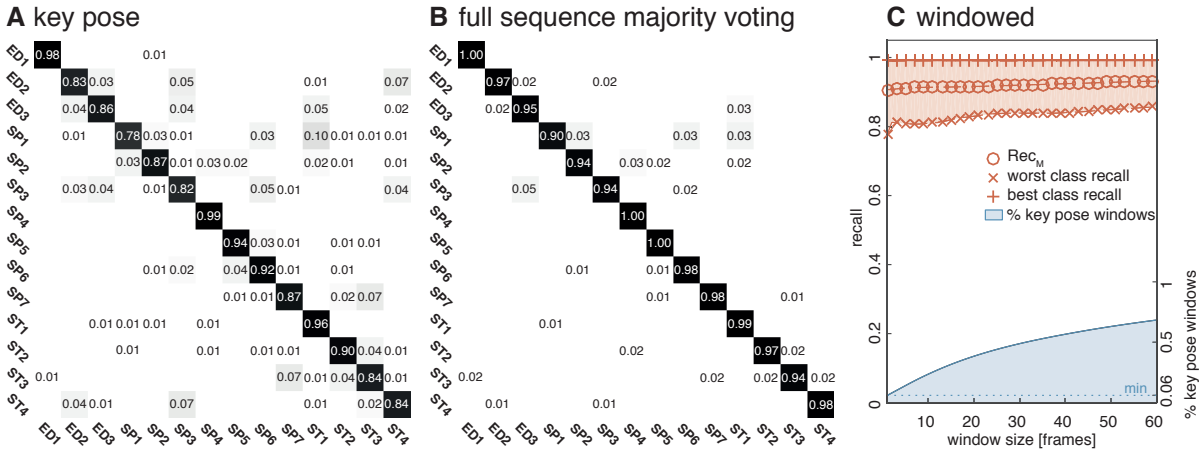


Figure 4.3: Key pose based action recognition performance classification rates. The proposed network was trained on the key pose frames extracted from the uulmMAD action recognition dataset. (A) Shows the per class classification rates obtained by single key pose frame classification. This allows the recognition of an action in the instant a key pose frame emerges in the input sequence. Average classwise recall (on the diagonal) ranges from 0.78 to 0.98. In (B) sequence level majority voting was applied. The resulting average recall rates range from 0.94 to 1.00. A sliding window based classification scheme was evaluated in (C). The best and worst per class average recall values together with the average recall value of Rec_M are displayed for temporal window sizes from 1 to 60 frames. In addition, the percentage of windows containing one or more key pose frames (and thus allow a classification of the action) is shown (blue line). Adapted from [Layher et al., 2017a], CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

In short, the key contributions presented in [Layher et al., 2017a] are:

- A feature-driven mechanism of automatically selecting key poses by evaluating the lateral and longitudinal extension of an observed body around local extrema in the motion energy. The proposed selection principle is in line with the feature characteristics identified in the perceptual study presented in [Layher and Neumann, 2018] and reflects the interaction between form and motion processing pathways described in [Layher et al., 2014a].
- An action recognition system inspired by the modeling efforts presented in [Layher et al., 2014a], combining form and motion information to learn hierarchical representations of key pose frames.
- An application of a deep convolutional neural network (DCNN) for the learning of action classes based on such key pose representations.
- The mapping of the trained network onto the IBM Neurosynaptic System platform, which enables a computationally and energy efficient execution and its final implementation and online execution on the TrueNorth neuromorphic chip.
- The obtained results show that classifying actions is possible solely based on key poses and thus using a minimal amount of temporal information.

5 Learning of Category & Subcategory Representations - A Neural Model

In the following, the biologically inspired model for the learning of category and subcategory representations presented in [Layher et al., 2014b] is related to the overall context of this thesis. In Sections 5.1 and 5.2, text passages and figures extracted from [Layher et al., 2014b] are included.

5.1 Introduction and Motivation

The unsupervised learning of categories of visual input stimuli has been addressed by numerous approaches in machine learning as well as in computational neuroscience. However, the question of what kind of mechanisms might be involved in the process of subcategory learning, or category refinement, remains a topic of active investigation. In [Layher et al., 2014b], a recurrent computational network architecture for the unsupervised learning of categorial and subcategorial representations is proposed. During learning, the connection strengths of bottom-up weights from input to higher-level category representations are adapted according to the input activity distribution. In a similar manner, top-down weights learn to encode the characteristics of a specific stimulus category. Feedforward and feedback learning in combination realize an associative memory mechanism, enabling the selective top-down propagation of a category's feedback weight distribution. The difference between the expected input encoded in the projective field of a category node and the current input pattern controls the amplification of feedforward-driven representations. Large enough differences trigger the recruitment of new representational resources and the establishment of additional (sub-) category representations. The proposed learning scheme is highly inspired by the mechanisms presented in [Layher et al., 2014a] for the unsupervised learning of sequence selective representations of biological motions. Both modeling efforts realize an associative memory by combining an instar with an outstar learning rule (compare [Carpenter, 1989]) and integrate the established outgoing distributions in preceding stages by applying a modulative feedback mechanism. In [Layher et al., 2014a], temporal associations between prototypical form and motion patterns are established to enable sequence-selective representations of biological motions. In [Layher et al., 2014b], on the other hand, an input is assigned to the best matching category representation and the difference between the input and the associated feedback distribution is used to establish a subcategory representation. In the following section, the central properties of the proposed model for the learning of category and subcategory representations are summarized. For a

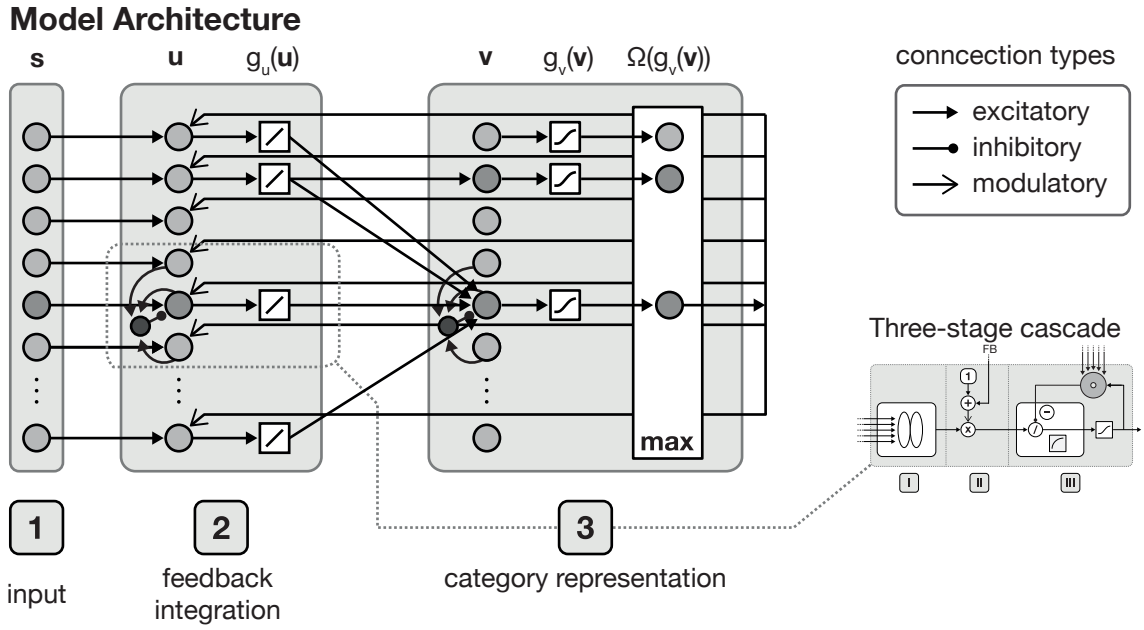


Figure 5.1: Three layer processing architecture. Category and subcategory representations of visual input stimuli are established in a three layer processing architecture. In layer 2, activities propagated from input layer 1 are combined with a residual signal emitted from layer 3. The residual signal encodes the differences between the current input and the best matching internal representation of the input. The feedback modulation and activity normalization in layer 2 emphasize this difference in the current activity and potentially trigger the establishment of a subcategory representation. Adapted from [Layher et al., 2014b], CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

detailed description please refer to the original publication [Layher et al., 2014b] included into this thesis on page 107.

5.2 Main Contributions

The proposed model is composed of interconnected cortical columns, subdivided into a compartmental structure of three processing layers see Figure 5.1. Layer 1 propagates input activities to the second layer, where they are combined with a residual signal derived from feedback activities emitted from layer 3 and the current activity. After normalization, layer 3 category cells perform a correlation of the current layer 2 activities and their respective synaptic input weights. The cell with the strongest activation is then selected by a winner-take all (WTA) mechanism for weight adaptation and activity propagation. In principle all of the model layers consist of a three-stage processing cascade as illustrated in Chapter 3, Figure 3.3.

Model layer 3 cells establish categorial and subcategorial representations using a Hebbian learning rule in combination with a modulatory feedback mechanism similar to the learning scheme presented in [Layher et al., 2014a] but without incorporating a temporal trace rule. An associative memory is realized by coupling an instar and outstar learning rule. The afferent

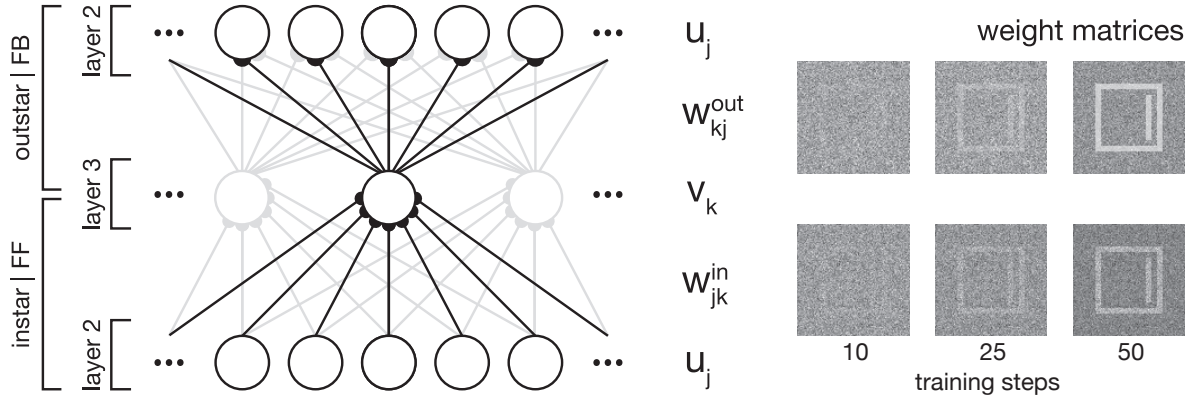


Figure 5.2: Instar and Outstar learning. An associative memory is realized by combining an instar with an outstar learning rule. Feedback weights associated to the best matching input distribution are used to derive a residual signal, which is modulated on the layer 2 activities. On the right-hand side exemplary input and output weight representations for simple visual stimuli are displayed. Adapted from [Layher et al., 2014b], CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

connections weights are used to select the best matching representation to a given input. The weights projecting away from the cell are incorporated in the top-down feedback to layer 2 cells. For a given stimulus, only the cell with the highest activation is selected for weight adaptation. Figure 5.2 displays the schematic of the proposed learning mechanism, together with exemplary weight matrices after several training steps.

The capability of the proposed model of establishing category and subcategory representations are demonstrated in several simulations. Figure 5.3 shows how one category and four subcategory representations are learned in model layer 3 cells. In the beginning all inputs are learned into one category cell. After about 200 training steps, the effect of the feedback is large enough to trigger the learning of a new subcategory representation. This process repeats several times, until each subcategory is represented by an own category cell.

In short, the contributions published in [Layher et al., 2014b] include:

- The combination of an instar and outstar learning rule to associate input activities to their best matching internal representations.
- A feedback mechanism enabling the amplification of the differences to the best matching internal representations in the input and, as such, combining predictive coding and modulatory feedback.
- A three-layered processing architecture for the unsupervised learning of categorial and subcategorial representations of visual input stimuli.

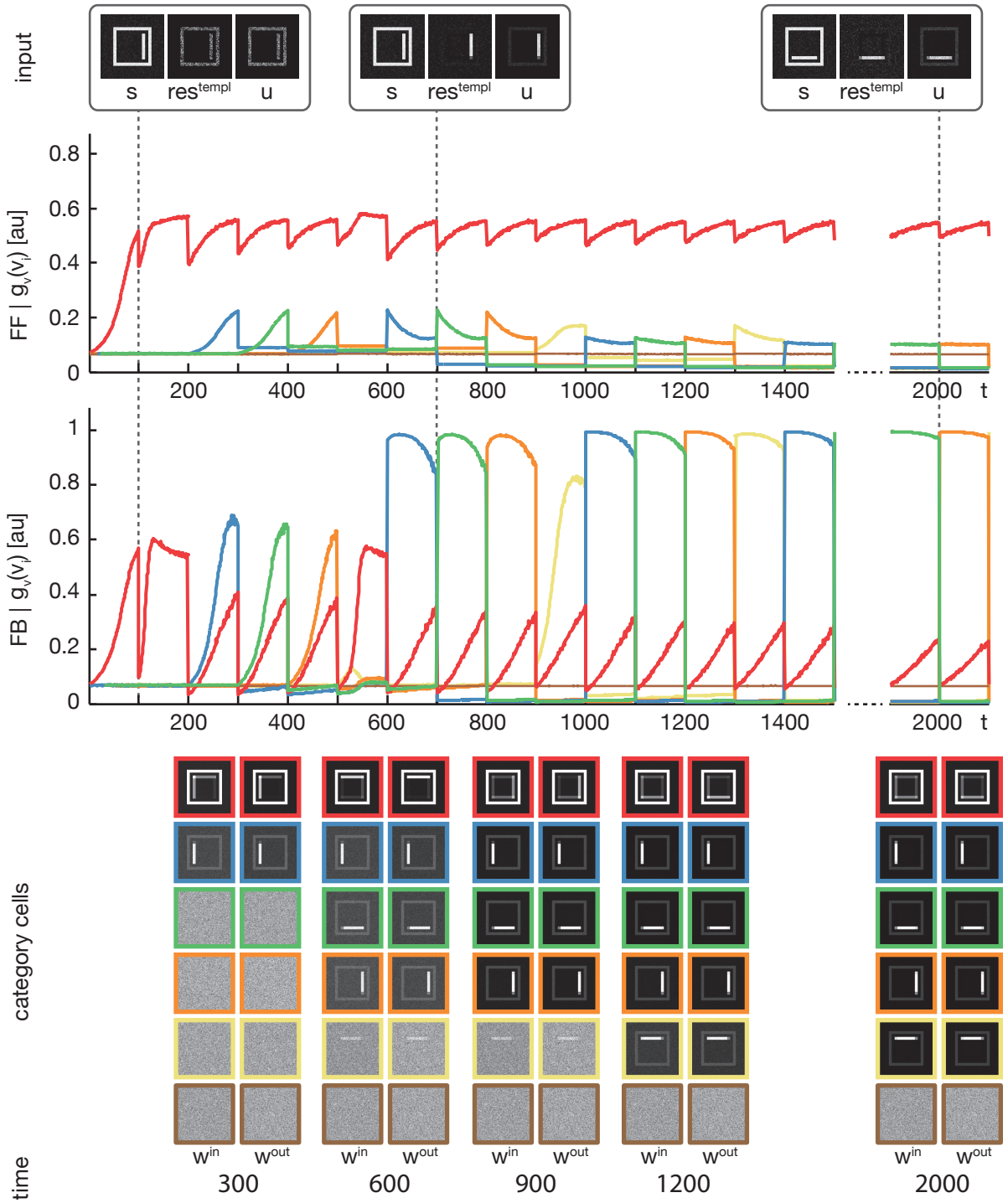


Figure 5.3: Learning of category and subcategory representations. The proposed model was trained using four simple rectangular shapes as input stimuli. The first row shows exemplary input configurations s, along with the corresponding residual signal $\text{res}^{\text{templ}}$ and the input signal u after feedback modulation. In the second row, the activities of the six category cells before the feedback sweep are shown. Before the feedback is effective on the input, only one cell (encoded in red) responds to all input configurations. This cell represents the overall category cell. The third row shows the activities after the feedback sweep. In the last row, the category cell weights are displayed framed by colors according to the activity plots. Adapted from [Layher et al., 2014b], CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

6 Summary & Conclusion

The articles summarized in Chapters 2-4 approach the processing principles and mechanisms underlying biological motion recognition from a perceptual, a modeling, as well as a pattern recognition perspective. Further, Chapter 5 briefly recapitulates how parts of the proposed mechanisms combining an associative learning scheme with a modulatory feedback rule successfully can be adapted to model different aspects of visual cortical processing, such as the learning of (sub-)category representations. In the following conclusion, text extracts from [Layher et al., 2014a; Layher et al., 2014b; Layher et al., 2017a; Layher and Neumann, 2018] are included.
Combining form and motion information for the task of recognizing biological motions or human actions in image sequences is a widely applied approach in the development of computer vision action recognition systems. The majority of approaches rely on analyzing spatio-temporal representations which equally integrate all frames, or regularly sample key poses (or snapshots) within a motion (sub-) sequence without considering their different expressiveness. Only a few number of approaches aim at systematically selecting highly expressive key poses for the recognition of an action and even less identify such poses without exploiting intra-/inter-class based statistics or optimizing the classification performance. In contrast, purely feature-driven key pose selection methods are independent of the number of different actions within a dataset. Thus, retraining is not necessary if, e.g., a new action is added the classification task and the sharing of key poses among different actions is in principle possible.
In [Layher and Neumann, 2018], a perceptual study using a novel masking technique to analyze the characteristics of visual features increasing the salience of an observed biological motion (sub-) sequence is presented. The reported results show a correlation between the combined lateral and longitudinal extension of a displayed biological motion point-light stimulus and the human recognition performance. While occurrences of a high lateral or longitudinal extension seem to have an impact on the perceptual salience of a biological motion subsequence, it is the temporal relationship which renders this presence of characteristic form information special and distinctive. This observation supports the hypothesis that local temporal extrema in (potentially a variety of) feature channels form the basis for an increase in the perceptual salience of a biological motion (sub-) sequence. In other words, the conjecture advocated in [Layher and Neumann, 2018] is that first order mechanisms operate upon the feature representations to register local temporal contrasts in the respective feature channels (or combinations of these). In turn, these generate event-triggers to serve as a marker, e.g., to steer the learning of form and motion pattern characteristics. It is worth to note that due to the nature of human kinetics, temporal local maxima in the lateral and longitudinal extension of a body pose co-occur with local extrema in the energy of whole body or limb motions (see Figure 2.2). Thus, motion energy in turn might be a candidate to identify local extrema of form/motion based feature characteristics.
In [Layher et al., 2014a] a neural model is presented, which proposes potential functional mech-

anisms which might account for the generation of such event-triggers around local temporal extrema in the motion energy and gate the establishment of biological motion representations of combined form and motion information. The model builds upon neurophysiological evidence about the cortical sites and specific neuronal representations which contribute to articulated motion and implied motion perception. The model suggests how prototypical form and motion representations can be established by unsupervised learning in the ventral (form) and dorsal (motion) processing pathways of the visual cortex. The applied learning mechanisms are based on modified Hebbian schemes which are further stabilized through a temporal trace mechanism. Sequence-selective representations in model area STS are learned by using the same learning mechanisms, but combining the responses of intermediate-level representations in the form and motion pathways. The learning of highly articulated poses (key poses) is controlled by a reinforcement mechanism that enables Hebbian learning in the form pathway through cross-pathway motion-form interaction. According to the principles observed in [Layher and Neumann, 2018; Thirkettle et al., 2009], key poses are automatically selected corresponding to strong body pose extensions/articulations by gating the learning of form prototypes based on the energy in the motion pathway. The sequence-selective cells in model area STS project to their respective input representations in the form and motion pathways. Such feedback connections are again learned by a Hebbian mechanism. The feedforward and the feedback interactions in combination establish a loop of recurrent processing that stabilizes the patterns of form, motion, and sequence representation. Via feedback, model STS cells generate a predictive signal through the backward connections' weights to encode the expected matching input that is suitable to match the currently activated sequence pattern. Together with the proposed feedback mechanism, the model is able to account for various experimental findings, in particular, the ability to infer and predict future motion sequence development from articulated postures. Importantly, cells in STS are responsive to both motion as well as static form [Oram and Perrett, 1996a]. The proposed model hypothesizes how the presentation of static articulated poses yields the emergence of predictive motion perception and enhanced neural activations in the motion pathway [implied motion; Kourtzi and Kanwisher, 2000; Senior et al., 2000]. Further, the model makes the prediction that the snapshot frames suggested by Giese and Poggio, 2003 and tested experimentally by Singer and Sheinberg, 2010; Vangeneugden et al., 2009 are not the result of a regular sampling process, but are established by selectively learning representations of highly articulated poses based on the suggested reinforcement mechanism.

In [Layher et al., 2017a], an action recognition system is presented. A feature-driven key pose selection mechanism is proposed, which builds upon the evidences about human action perception reported in, e.g., [Layher and Neumann, 2018; Thirkettle et al., 2009]. In contrast to the majority of key pose based action recognition approaches reported in the literature (see [Layher et al., 2017a] for a detailed literature review), the selection mechanism does not utilize any information about the inter- or intra-class distribution of the key poses to optimize the classification accuracy. Key poses are identified here purely feature-driven by selecting local extrema in the motion energy and evaluating the lateral and longitudinal body extension around these events. It is demonstrated, that the classification accuracy is on par with state-of-the-art key pose based action recognition approaches, although only motion and form related feature characteristics are used to select a key pose frame. Further, a biologically inspired architecture combining form and motion information to learn hierarchical representations of the selected key pose frames is proposed. Such hierarchical feature representations are expected to make the

recognition more robust against clutter and partial occlusions, in comparison to the holistic shape representations of the full body configurations used in the modeling approach in [Layher et al., 2014a]. Form and motion pattern representations are established employing a neuromorphic deep convolutional neural network. The trained network is mapped onto the IBM Neurosynaptic System platform [Merolla et al., 2014; Esser et al., 2016], which enables a computationally and energy efficient execution of the algorithm. In a cross-dataset evaluation, the proposed action recognition system showed a promising recognition performance. This generalization capability might partially be attributed to the purely feature-driven key pose selection mechanism, which does not rely on optimizing the inter-class distinctiveness or classification performance of the selected key poses and thus is independent of the type and number of action classes within a dataset.
In [Layher et al., 2014b], the learning mechanisms proposed in [Layher et al., 2014a] are adapted and applied to learn representations of visual (sub-) categories. As for the sequence-selective cells in [Layher et al., 2014a], the learning mechanisms occur in the feedforward as well as in the feedback connections. Feedforward connections will learn the specific configuration of an (average) appearance of an input pattern that the learned category is selectively tuned to. The feedback connections, on the other hand, adjust their weights in order to improve the predicted input pattern that maximally excites the category representation. The top-down feedback learning mechanism combines the modulatory feedback [Girard and Bullier, 1989; De Pasquale and Sherman, 2013] with the concept of top-down predictors that tend to minimize the residual error between feedforward sensory signals and the top-down pattern [Rao and Ballard, 1999; Bastos et al., 2012]. The idea behind this concept is that weights will be increased when the predicted pattern and the current input differ. The amount of this gain increase depends on the residual difference between these two patterns. If the difference large enough, a new category or subcategory representation is established. The principles proposed in [Layher et al., 2014b] allow to further develop the understanding of how complex distributed representations can be learned and how average categories are learned together with subcategories for components that deviated significantly from the average category representation.

Taken together, in the present article thesis, the visual appearance of human biological motion is analyzed following an interdisciplinary research approach from perceptual studies to computational modeling and computer vision algorithms. It is demonstrated how the results of an empirical study and the modeling of the potential corresponding processing principles in the visual cortex complement each other to contribute to a better understanding of the principles and mechanisms behind the phenomenon of biological motion perception. It is shown how the development of a technical system for the recognition of actions in video sequences benefits from such insights gained in experimental and modeling efforts. In particular, the example demonstrates how parts of the neural modeling approach could be transferred to a challenging application task. Further, the mechanisms proposed for the learning of biological motion representations are partially adapted to realize an unsupervised learning of (sub-) category representations of simple visual patterns. The latter makes a new contribution how coarse grained average representations might be further elaborated and detailed automatically when more information becomes available to a learning system.

Bibliography

Akopyan, F., Sawada, J., Cassidy, A., Alvarez-Icaza, R., Arthur, J., Merolla, P., et al. (2015). "TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip". In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 34.10, pp. 1537–1557 (cit. on p. 24).

Barraclough, N. E., Xiao, D., Oram, M. W., and Perrett, D. I. (2006). "The sensitivity of primate STS neurons to walking sequences and to the degree of articulation in static images". In: *Progress in Brain Research* 154, pp. 135–148 (cit. on pp. 16, 19).

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). "Canonical microcircuits for predictive coding". In: *Neuron* 76.4, pp. 695–711 (cit. on p. 32).

Bayerl, P. and Neumann, H. (2004). "Disambiguating visual motion through contextual feedback modulation". In: *Neural Computation* 16.10, pp. 2041–2066 (cit. on pp. 16, 18).

Beauchamp, M. S., Lee, K. E., Haxby, J. V., and Martin, A. (2003). "FMRI responses to video and point-light displays of moving humans and manipulable objects". In: *Journal of Cognitive Neuroscience* 15.7, pp. 991–1001 (cit. on p. 19).

Bouecke, J. D., Tlapale, E., Kornprobst, P., and Neumann, H. (2011). "Neural mechanisms of motion detection, integration, and segregation: from biology to artificial image processing systems". In: *Eurasip Journal on Advances in Signal Processing*, pp. 1–22 (cit. on p. 18).

Boussaoud, D., Ungerleider, L. G., and Desimone, R. (1990). "Pathways for Motion Analysis: Cortical Connections of the Medial Superior Temporal and Fundus of the Superior Temporal Visual Areas in the Macaque". In: *Journal of Comparative Neurology* 296.3, pp. 462–495 (cit. on p. 15).

Carandini, M., Heeger, D. J., and Movshon, J. A. (1999). "Linearity and gain control in V1 simple cells". In: *Cerebral Cortex* 13, pp. 401–444 (cit. on p. 16).

Carpenter, G. A. (1989). "Neural network models for pattern recognition and associative memory". In: *Neural Networks* 2.4, pp. 243–257 (cit. on pp. 16, 26).

De Pasquale, R. and Sherman, S. M. (2013). "A modulatory effect of the feedback from higher visual areas to V1 in the mouse." In: *Journal of Neurophysiology* 109.10, pp. 2618–2631 (cit. on p. 32).

Esser, S. K., Merolla, P. A., Arthur, J. V., Cassidy, A. S., Appuswamy, R., Andreopoulos, A., et al. (2016). "Convolutional Networks for Fast, Energy-Efficient Neuromorphic Computing". In: *Proceedings of the National Academy of Sciences* 113.41, pp. 11441–11446 (cit. on pp. 23, 32).

Felleman, D. J. and Van Essen, D. C. (1991). "Distributed Hierarchical Processing in the Primate Cerebral Cortex". In: *Cerebral cortex* 1.1, pp. 1–47 (cit. on p. 15).

Garcia, J. O. and Grossman, E. D. (2008). "Necessary but not Sufficient: Motion Perception is Required for Perceiving Biological Motion". In: *Vision Research* 48.9, pp. 1144–1149 (cit. on p. 11).

Giese, M. A. and Poggio, T. (2003). "Neural Mechanisms for the Recognition of Biological Movements". In: *Nature Reviews Neuroscience* 4.3, pp. 179–192 (cit. on pp. 15, 18, 31).

Girard, P. and Bullier, J. (1989). "Visual activity in area V2 during reversible inactivation of area 17 in the macaque monkey". In: *Journal of Neurophysiology* 62.6, pp. 1287–1302 (cit. on p. 32).

Glodek, M., Layher, G., Heilemann, F., Gawrilowicz, F., Palm, G., Schwenker, F., et al. (2015). "uulmMAD – A Human Action Recognition Dataset for Ground-Truth Evaluation and Investigation of View Invariances". In: *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*. Lecture Notes in Computer Science. Springer, pp. 77–91 (cit. on pp. 23, 24).

Grossman, E. D., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., et al. (2000). "Brain Areas involved in Perception of Biological Motion". In: *Journal of Cognitive Neuroscience* 12.5, pp. 711–720 (cit. on p. 15).

Hansen, T. and Neumann, H. (2008). "A recurrent model of contour integration in primary visual cortex". In: *Journal of Vision* 8.8, pp. 1–25 (cit. on p. 18).

Hiris, E. (2007). "Detection of Biological and Nonbiological Motion". In: *Journal of Vision* 7.12, p. 4. DOI: 10.1167/7.12.4 (cit. on p. 11).

Jellema, T., Maassen, G., and Perrett, D. I. (2004). "Single cell integration of animate form, motion and location in the superior temporal cortex of the macaque monkey". In: *Cerebral Cortex* 14.7, pp. 781–790 (cit. on p. 18).

Jellema, T. and Perrett, D. I. (2003). "Cells in monkey STS responsive to articulated body motions and consequent static posture: a case of implied motion?" In: *Neuropsychologia* 41.13, pp. 1728–1737 (cit. on p. 19).

Johansson, G. (1976). "Spatio-Temporal Differentiation and Integration in Visual Motion Perception". In: *Psychological Research* 38.4, pp. 379–393 (cit. on p. 11).

Johansson, G. (1973). "Visual Perception of Biological Motion and a Model for its Analysis". In: *Perception & Psychophysics* 14.2, pp. 201–211 (cit. on pp. 7, 11).

Kourtzi, Z. and Kanwisher, N. (2000). "Activation in human MT/MST by static images with implied motion". In: *Journal of Cognitive Neuroscience* 12.1, pp. 48–55 (cit. on pp. 9, 16, 31).

Lange, J. and Lappe, M. (2006). "A Model of Biological Motion Perception from Configural Form Cues". In: *Journal of Neuroscience* 26.11, pp. 2894–2906. ISSN: 0270-6474 (cit. on p. 11).

Layher, G., Brosch, T., and Neumann, H. (2017a). "Real-time Biologically Inspired Action Recognition from Key Poses using a Neuromorphic Architecture". In: *Frontiers in Neurorobotics* 11.13, p. 19. ISSN: 1662-5218. DOI: 10.3389/fnbot.2017.00013 (cit. on pp. 9, 22–25, 30, 31, 49, 81).

Layher, G., Giese, M. A., and Neumann, H. (2012). "Learning Representations for Animated Motion Sequence and Implied Motion Recognition". In: *Proceedings of the 22nd International Conference on Artificial Neural Networks (ICANN), 2012*. Ed. by A. E. P. Villa, W. Duch, P. Érdi, F. Masulli, and G. Palm. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 288–295. ISBN: 978-3-642-33269-2. DOI: 10.1007/978-3-642-33269-2_37 (cit. on p. 19).

Layher, G., Giese, M. A., and Neumann, H. (2014a). "Learning Representations of Animated Motion Sequences — A Neural Model". In: *Topics in Cognitive Science* 6.1, pp. 170–182.

ISSN: 1756-8765. DOI: 10.1111/tops.12075 (cit. on pp. 9, 10, 15–19, 21–23, 25–27, 30, 32, 49, 65).

Layher, G., Glodek, M., and Neumann, H. (2017b). "Analysis of Articulated Motion for Social Signal Processing". In: *Companion Technology: A Paradigm Shift in Human-Technology Interaction*. Ed. by S. Biundo and A. Wendemuth. Springer International Publishing, pp. 345–364. ISBN: 978-3-319-43665-4. DOI: 10.1007/978-3-319-43665-4_17 (cit. on pp. 8, 15, 16, 19, 20).

Layher, G. and Neumann, H. (2013). "Combining Form and Motion - An Integrated Approach for Learning Biological Motion Representations". In: *Journal of Vision (VSS Abstracts)* 13.9, p. 194. DOI: 10.1167/13.9.194 (cit. on p. 19).

Layher, G. and Neumann, H. (2018). "Points and Stripes: A Novel Technique for Masking Biological Motion Point-Light Stimuli". In: *Frontiers in Psychology* 9, p. 12. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2018.01455 (cit. on pp. 9, 11–16, 19, 21, 22, 25, 30, 31, 49, 51).

Layher, G., Schrodt, F., Butz, M. V., and Neumann, H. (2014b). "Adaptive Learning in a Compartmental Model of Visual Cortex - How Feedback Enables Stable Category Learning and Refinement". In: *Frontiers in Psychology* 5.1287, p. 19. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2014.01287 (cit. on pp. 10, 18, 26–30, 32, 49, 105).

Mather, G. and Murdoch, L. (1994). "Gender Discrimination in Biological Motion Displays Based on Dynamic Cues". In: *Proceedings of the Royal Society of London B: Biological Sciences* 258.1353, pp. 273–279 (cit. on p. 11).

Merolla, P. A., Arthur, J. V., Alvarez-Icaza, R., Cassidy, A. S., Sawada, J., Akopyan, F., et al. (2014). "A Million Spiking-Neuron Integrated Circuit with a Scalable Communication Network and Interface". In: *Science* 345.6197, pp. 668–673 (cit. on pp. 24, 32).

Oram, M. W. and Perrett, D. I. (1994). "Responses of Anterior Superior Temporal Polysensory (STPa) Neurons to "Biological Motion" Stimuli". In: *Journal of Cognitive Neuroscience* 6.2, pp. 99–116 (cit. on p. 15).

Oram, M. W. and Perrett, D. I. (1996a). "Integration of Form and Motion in the Anterior Superior Temporal Polysensory Area (STPa) of the Macaque Monkey". In: *Journal of Neurophysiology* 76.1, pp. 109–129. ISSN: 0022-3077 (cit. on p. 31).

Oram, M. W. and Perrett, D. I. (1996b). "Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey". In: *Journal of Neurophysiology* 76.1, pp. 109–129 (cit. on p. 15).

Puce, A. and Perrett, D. I. (2003). "Electrophysiology and Brain Imaging of Biological Motion". In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 358.1431, pp. 435–445 (cit. on p. 15).

Rao, R. P. and Ballard, D. H. (1999). "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects". In: *Nature Neuroscience* 2.1, pp. 79–87 (cit. on p. 32).

Raudies, F., Mingolla, E., and Neumann, H. (2011). "A model of motion transparency processing with local center-surround interactions and feedback". In: *Neural Computation*, pp. 1–45 (cit. on p. 18).

Senior, C., Barnes, J., Giampietroc, V., Simmons, A., Bullmore, E. T., Brammer, M., et al. (2000). "The functional neuroanatomy of implicit-motion perception or 'representational momentum'". In: *Current Biology* 10.1, pp. 16–22 (cit. on p. 31).

Singer, J. M. and Sheinberg, D. L. (2010). "Temporal cortex neurons encode articulated actions as slow sequences of integrated poses". In: *Journal of Neuroscience* 30.8, pp. 3133–3145 (cit. on p. 31).

Thirkettle, M., Benton, C. P., and Scott-Samuel, N. E. (2009). "Contributions of Form, Motion and Task to Biological Motion Perception". In: *Journal of Vision* 9.3, p. 28 (cit. on pp. 9, 11, 12, 16, 19, 21, 31).

Vangeneugden, J., Pollick, F., and Vogels, R. (2009). "Functional differentiation of macaque visual temporal cortical neurons using a parametric action space". In: *Cerebral Cortex* 19.3, pp. 593–611 (cit. on p. 31).

Weidenbacher, U. and Neumann, H. (2009). "Extraction of surface-related features in a recurrent model of V1-V2 interactions". In: *PloS ONE* 4.6, e5909 (cit. on p. 16).

I Individual Contributions & Publications

Individual Contributions

The central contributions presented within the thesis are published in four journal articles, parts of which are recapitulated in Chapters 1-6. On the following pages, the individual contributions of the respective authors of the four articles are detailed.

[Layher and Neumann, 2018] G. Layher and H. Neumann (2018). "Points and Stripes: A Novel Technique for Masking Biological Motion Point-Light Stimuli". In: *Frontiers in Psychology* 9, p. 12. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2018.01455

In Chapter 2, the results of the psychophysical study presented in [Layher and Neumann, 2018; included page 53 of the thesis] are summarized and embedded in the overall context of analyzing the principles behind the learning of biological motion representations. Based upon a novel technique for masking *point-light stimuli*, the conducted experiments investigate how specific spatio-temporal features, such as the longitudinal and lateral extension, affect the perceptual salience of an observed human motion subsequence. The presented results further support evidences that articulated pose configurations increase the human performance in discriminating biological motion sequences. Here, local extrema in form and/or motion based feature channels act as a marker, e.g., to steer the learning of biological motion representations.

Georg Layher proposed the novel technique for masking *point-light stimuli* without modifying local dot motion patterns, as well as designed the stimulus variations intended to isolate different feature characteristics of the *point-light stimuli* under investigation. With the support of two students, Georg Layher generated the stimuli on basis of the dataset presented in [Glodek et al., 2015] and implemented/designed the experiments. Furthermore, Georg Layher designed and conducted the presented evaluations and performed the statistical analysis, as well as wrote the manuscript.

Prof. Dr. Heiko Neumann developed the presented novel masking technique and the overall design of experiments together with Georg Layher, participated in the design and the conduction of the experiments and in the interpretation of the results. Prof. Neumann improved the presentation of the manuscript by contributing to the abstract, the introduction and the discussion of the results in light of the current state of research.

[Layher et al., 2014a] G. Layher, M. A. Giese, and H. Neumann (2014a). "Learning Representations of Animated Motion Sequences — A Neural Model". In: *Topics in Cognitive Science* 6.1, pp. 170–182. ISSN: 1756-8765. DOI: 10.1111/tops.12075

Chapter 3 recapitulates the structure, functional mechanisms and simulation results of the neural model for the learning of form and motion based biological motion representations presented in [Layher et al., 2014a; included on page 67 of the thesis]. The model is in line with evidences from psychophysical and biophysiological experiments and gives a potential explanation for the functional mechanisms underlying *implied motion* perception, as well as the learning of form and motion based representations of biological motion sequences.

Georg Layher extended the model proposed in [Giese and Poggio, 2003] by combining the models for the hierarchical feedforward and feedback processing along the ventral (form) and dorsal (motion) pathways presented in [Bayerl and Neumann, 2004; Weidenbacher and Neumann, 2009], designed/implemented the unsupervised Hebbian learning mechanism used on top of both processing pathways to establish static form (model area IT) and motion (model area MST) prototypes, as well as proposed the gating mechanism, which enables the unsupervised learning of key pose representations by modulating the learning in the ventral pathway with a signal derived in the dorsal pathway. Furthermore, Georg Layher suggested the combination of the instar/outstar learning schemes in model area STS with a temporal trace rule to realize the temporal association learning, implemented the overall model, conducted the simulations and the statistical analysis, as well as wrote the manuscript.

Prof. Dr. Martin A. Giese inspired the overall design of the model, made suggestions concerning the questions addressed in the performed simulations and significantly improved the presented results by providing his excellent expertise in the field of biological motion in numerous discussions. Furthermore Prof. Giese improved the readability of the article during the writing of the manuscript.

Prof. Dr. Heiko Neumann developed the overall design of the model together with Georg Layher, participated in the design and the conduction of the experiments and in the interpretation of the simulation results. Prof. Neumann improved the presentation of the manuscript by contributing to the discussion of the results in the context of the current state of research.

[Layher et al., 2017a] G. Layher, T. Brosch, and H. Neumann (2017a). "Real-time Biologically Inspired Action Recognition from Key Poses using a Neuromorphic Architecture". In: *Frontiers in Neurorobotics* 11.13, p. 19. ISSN: 1662-5218. DOI: 10.3389/fnbot.2017.00013

The real-time biologically inspired action recognition system presented in [Layher et al., 2017a; included on page 83 of the thesis] is recapitulated in Chapter 4. The proposed architecture builds upon the key principles behind the learning of biological motion representations identified and summarized in Chapter 2 & 3 and is on par with the recognition performance reported for other key pose based action recognition in the literature. For the assignment of key poses to action classes, a deep convolutional neural network was trained and mapped onto the *IBM Neurosynaptic System* platform, which enables a computationally and energy efficient execution of the trained model.

Georg Layher designed/implemented the unsupervised key pose selection mechanism based on the experimental findings described in [Layher and Neumann, 2018; see Chapter 2] and applied in [Layher et al., 2014a; see Chapter 3], implemented the preprocessing of the input images, as well as performed the training and the evaluation of the neuromorphic deep convolutional network using the *energy-efficient deep neuromorphic networks* (*Eedn*) framework [Esser et al., 2016]. Furthermore, Georg Layher designed and conducted the presented evaluations, performed the statistical analysis, and wrote the manuscript.

Dr. Tobias Brosch took part in the overall design of the model, made suggestions concerning the applied topology of the deep convolutional neural network and implemented the normalization of the input as part of the preprocessing. Furthermore, Dr. Brosch improved the presentation of the manuscript by thorough proofreading.

Prof. Dr. Heiko Neumann developed the overall design of the model in collaboration with Georg Layher, participated in the design and the conduction of the experiments and in the interpretation of the simulation results. Prof. Neumann improved the presentation of the manuscript by contributing to the discussion of the results in the context of the current state of research and by thorough proofreading.

[Layher et al., 2014c] G. Layher, F. Schrodt, M. V. Butz, and H. Neumann (2014c). "Adaptive Learning in a Compartmental Model of Visual Cortex - How Feedback Enables Stable Category Learning and Refinement". In: *Frontiers in Psychology* 5.1287, p. 19. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2014.01287

Chapter 5 summarizes the model for the adaptive learning of category and subcategory representations presented in [Layher et al., 2014c; included on page 107 of the thesis]. Combining an instar and an outstar learning rule with a feedback mechanism, the applied learning scheme is strongly inspired by the adaptation mechanism proposed in [Layher et al., 2014a]. In a three-layered processing architecture, a residual signal is derived from feedback activities reflecting already established categorial representations and the current input activity. The residual signal is then modulated on the current input, normalized and propagated to the successive layer, where categorial representations are learned. Simulations show how the model establishes categorial representations of artificial visual stimuli and successfully further refines them into subcategorial representations.

Georg Layher proposed the three-layer processing architecture for the establishment of category and subcategory representations, suggested the learning scheme combining instar and outsar learning rules with a feedback mechanism, as well as proposed the integration of the residual signal derived from the feedback activities and the current input. Furthermore, Georg Layher implemented the overall model, designed and conducted the simulations, performed the evaluation of the presented results, and wrote the manuscript.

Dr. Fabian Schrodt took part in the overall design of the model, made suggestions concerning the applied learning mechanisms and contributed on the interpretation of the presented results. Furthermore, Dr. Schrodt improved the presentation of the manuscript by thorough proofreading.

Prof. Dr. Martin V. Butz took part in the overall design of the model. In various discussions, Prof. Butz made suggestions concerning the applied learning mechanisms and provided his excellent expertise in the field of cognitive and learning systems. Furthermore, Prof. Butz contributed on the interpretation of the results and improved the presentation of the manuscript by thorough proofreading.

Prof. Dr. Heiko Neumann developed the overall design of the model together with Georg Layher, participated in the design and the conduction of the simulations and in the interpretation of the presented results. Prof. Neumann improved the presentation of the manuscript by contributing to the discussion of the results in the context of the current state of research, as well as by thorough proofreading.

List of Publications

Journal Articles

† **Layher**, **G.** and Neumann, H. (2018). "Points and Stripes: A Novel Technique for Masking Biological Motion Point-Light Stimuli". In: *Frontiers in Psychology* 9, p. 12. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2018.01455.

† **Layher**, **G.**, Brosch, T., and Neumann, H. (2017a). "Real-time Biologically Inspired Action Recognition from Key Poses using a Neuromorphic Architecture". In: *Frontiers in Neurorobotics* 11.13, p. 19. ISSN: 1662-5218. DOI: 10.3389/fnbot.2017.00013.

Riemer, V., Frommel, J., **Layher**, **G.**, Neumann, H., and Schrader, C. (2017). "Identifying Features of Bodily Expression As Indicators of Emotional Experience during Multimedia Learning". In: *Frontiers in Psychology* 8, p. 1303. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2017.01303.

Schrodt, F., **Layher**, **G.**, Neumann, H., and Butz, M. V. (2015). "Embodied Learning of a Generative Neural Model for Biological Motion Perception and Inference". In: *Frontiers in Computational Neuroscience* 9.79, p. 20. ISSN: 1662-5188. DOI: 10.3389/fncom.2015.00079.

† **Layher**, **G.**, Giese, M. A., and Neumann, H. (2014a). "Learning Representations of Animated Motion Sequences — A Neural Model". In: *Topics in Cognitive Science* 6.1, pp. 170–182. ISSN: 1756-8765. DOI: 10.1111/tops.12075.

† **Layher**, **G.**, Schrodt, F., Butz, M. V., and Neumann, H. (2014b). "Adaptive Learning in a Compartmental Model of Visual Cortex - How Feedback Enables Stable Category Learning and Refinement". In: *Frontiers in Psychology* 5.1287, p. 19. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2014.01287.

Scherer, S., Glodek, M., **Layher**, **G.**, Schels, M., Schmidt, M., Brosch, T., et al. (2012a). "A Generic Framework for the Inference of User States in Human Computer Interaction". In: *Journal on Multimodal User Interfaces* 6.3, pp. 117–141. ISSN: 1783-8738. DOI: 10.1007/s12193-012-0093-9.

Book Chapters

Layher, **G.**, Glodek, M., and Neumann, H. (2017b). "Analysis of Articulated Motion for Social Signal Processing". In: *Companion Technology: A Paradigm Shift in Human-Technology Interaction*. Ed. by S. Biundo and A. Wendemuth. Springer International Publishing, pp. 345–364. ISBN: 978-3-319-43665-4. DOI: 10.1007/978-3-319-43665-4_17.

Siegert, I., Schüssel, F., Schmidt, M., Reuter, S., Meudt, S., **Layher**, **G.**, et al. (2017). "Multimodal Information Processing inCompanion-Systems: A Ticket Purchase System". In: *Companion Technology: A Paradigm Shift in Human-Technology Interaction*. Ed. by S. Biundo and A. Wendemuth. Springer International Publishing, pp. 493–500. ISBN: 978-3-319-43665-4. DOI: 10.1007/978-3-319-43665-4_25.

† Part of the article thesis.

Schels, M., Glodek, M., Meudt, S., Scherer, S., Schmidt, M., **Layher**, **G.**, et al. (2013). "Multi-modal classifier-fusion for the recognition of emotions". In: *Coverbal Synchrony in Human-Machine Interaction*. Ed. by M. Rojc and N. Campbell. CRC Press. Chap. 4, pp. 73–98.

Conference Proceedings

Zimoch, M., Pryss, R., **Layher**, **G.**, Neumann, H., Probst, T., Schlee, W., et al. (2018). "Utilizing the Capabilities Offered by Eye-Tracking to Foster Novices' Comprehension of Business Process Models". In: *Proceedings of the Second International Conference on Cognitive Computing (ICCC), 2018.* Ed. by J. Xiao, Z.-H. Mao, T. Suzumura, and L.-J. Zhang. Springer International Publishing, pp. 155–163. ISBN: 978-3-319-94307-7. DOI: 10.1007/978-3-319-94307-7_12.

König, D., Adam, M., Jarvers, C., **Layher**, **G.**, Neumann, H., and Teutsch, M. (2017). "Fully Convolutional Region Proposal Networks for Multispectral Person Detection". In: *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 243–250. DOI: 10.1109/CVPRW.2017.36.

Zhang, Y., **Layher**, **G.**, and Neumann, H. (2017a). "Continuous activity understanding based on accumulative pose-context visual patterns". In: *Proceedings of the Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), 2017.* Pp. 1–6. ISBN: 978-1-5386-1842-4. DOI: 10.1109/IPTA.2017.8310114.

Zhang, Y., **Layher**, **G.**, Walter, S., Kessler, V., and Neumann, H. (2017b). "Visual Confusion Recognition in Movement Patterns from Walking Path and Motion Energy". In: *Proceedings of the 15th International Conference on Enhanced Quality of Life and Smart Living (ICOST), 2017.* Ed. by M. Mokhtari, B. Abdulrazak, and H. Aloulou. Springer International Publishing, pp. 124–135. ISBN: 978-3-319-66188-9. DOI: 10.1007/978-3-319-66188-9_11.

Geier, T., Glodek, M., **Layher**, **G.**, Neumann, H., Biundo, S., and Palm, G. (2016). "On Stacking Probabilistic Temporal Models with Bidirectional Information Flow". In: *Proceedings of the Eighth International Conference on Probabilistic Graphical Models (PGM), 2016.* Ed. by A. Antonucci, G. Corani, and C. P. de Campos. Vol. 52. JMLR, pp. 195–206. ISBN: 1938-7228.

Layher, **G.**, Brosch, T., and Neumann, H. (2016). "Towards a Mesoscopic-Level Canonical Circuit Definition for Visual Cortical Processing." In: *The First International Workshop on Computational Models of the Visual Cortex: Hierarchies, Layers, Sparsity, Saliency and Attention*. Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering (ICST), pp. 543–550. DOI: 10.4108/eai.3-12-2015.2262448.

Siegert, I., Reuter, S., Schüssel, F., **Layher**, **G.**, Hoernle, T., Meudt, S., et al. (2016). "Multimodal Information Processing: The Ticket Purchase - a Demonstration Scenario of the SFB/TRR-62." In: *Proceedings of the 27. Konferenz Elektronische Sprachsignalverarbeitung (ESSV), 2016.* Ed. by O. Jokisch. Vol. 81. Studientexte zur Sprachkommunikation. TUDpress, pp. 111–118. ISBN: 978-3-95908-040-8.

Velana, M., Gruss, S., **Layher**, **G.**, Thiam, P., Zhang, Y., Schork, D., et al. (2016). "The SenseEmotion Database: A Multimodal Database for the Development and Systematic Validation of an Automatic Pain- and Emotion-Recognition System". In: *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*. Ed. by F. Schwenker and S.

Scherer. Lecture Notes in Computer Science. Springer International Publishing, pp. 127–139. ISBN: 978-3-319-59259-6. DOI: 10.1007/978-3-319-59259-6_11.

Glodek, M., **Layher**, **G.**, Heilemann, F., Gawrilowicz, F., Palm, G., Schwenker, F., et al. (2015). "uulmMAD – A Human Action Recognition Dataset for Ground-Truth Evaluation and Investigation of View Invariances". In: *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*. Ed. by F. Schwenker, S. Scherer, and L.-P. Morency. Lecture Notes in Computer Science. Springer International Publishing, pp. 77–91. ISBN: 978-3-319-14898-4. DOI: 10.1007/978-3-319-14899-1_8.

Layher, **G.**, Tschechne, S., Niese, R., Al-Hamadi, A., and Neumann, H. (2015). "Towards the Separation of Rigid and Non-rigid Motions for Facial Expression Analysis". In: *Proceedings of the 11th International Conference on Intelligent Environments (IE), 2015*. IEEE, pp. 176–179. ISBN: 978-1-4673-6654-0. DOI: 10.1109/IE.2015.38.

Schrodt, F., **Layher**, **G.**, Neumann, H., and Butz, M. V. (2014a). "Modeling Perspective-Taking Upon Observation of 3D biological Motion." In: *Proceedings of the Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-Epirob), 2014*. IEEE, pp. 305–310. ISBN: 978-1-4799-7540-2. DOI: 10.1109/DEVLRN.2014.6982998.

Schrodt, F., **Layher**, **G.**, Neumann, H., and Butz, M. (2014b). "Modeling perspective-taking by correlating visual and proprioceptive dynamics." In: *Proceedings of the 36th Annual Meeting of the Cognitive Science Society (CogSci), 2014*. Ed. by B. Bello, M. Guarini, M. McShane, and B. Scassellati. Cognitive Science Society, pp. 1383–1388. ISBN: 978-0-9911967-0-8.

Layher, **G.**, Giese, M. A., and Neumann, H. (2013). "Learning Representations of Animated Motion Sequences—A Neural Model". In: *Proceedings of the 35th Annual Meeting of the Cognitive Science Society (CogSci), 2013*. Cognitive Science Society, pp. 870–875. ISBN: 978-0-9768318-9-1.

Best Computational Modeling Paper Award (Action / Perception).

Tschechne, S., **Layher**, **G.**, and Neumann, H. (2013). "A Biologically Inspired Model for the Detection of External and Internal Head Motions". In: *Proceedings of the 23rd International Conference on Artificial Neural Networks (ICANN), 2013*. Ed. by V. Mladenov, P. Koprinkova-Hristova, G. Palm, A. E. P. Villa, B. Appollini, and N. Kasabov. Springer Berlin Heidelberg, pp. 232–239. ISBN: 978-3-642-40728-4. DOI: 10.1007/978-3-642-40728-4_29.

Glodek, M., **Layher**, **G.**, Schwenker, F., and Palm, G. (2012). "Recognizing Human Activities Using a Layered Markov Architecture". In: *Proceedings of the 22nd International Conference on Artificial Neural Networks (ICANN), 2012*. Ed. by A. E. P. Villa, W. Duch, P. Érdi, F. Masulli, and G. Palm. Springer Berlin Heidelberg, pp. 677–684. ISBN: 978-3-642-33269-2. DOI: 10.1007/978-3-642-33269-2_85.

Layher, **G.**, Giese, M. A., and Neumann, H. (2012). "Learning Representations for Animated Motion Sequence and Implied Motion Recognition". In: *Proceedings of the 22nd International Conference on Artificial Neural Networks (ICANN), 2012*. Ed. by A. E. P. Villa, W. Duch, P. Érdi, F. Masulli, and G. Palm. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 288–295. ISBN: 978-3-642-33269-2. DOI: 10.1007/978-3-642-33269-2_37.

Scherer, S., **Layher**, **G.**, Kane, J., Neumann, H., and Campbell, N. (2012b). "An audiovisual political speech analysis incorporating eye-tracking and perception data". In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), 2012*. Ed. by N. C. (Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, et al.

European Language Resources Association (ELRA), pp. 1114–1120. ISBN: 978-2-9517408-7-7.

Glodek, M., Tschechne, S., **Layher**, **G.**, Schels, M., Brosch, T., Scherer, S., et al. (2011). "Multiple Classifier Systems for the Classification of Audio-Visual Emotional States". In: *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction (ACII), 2011*. Ed. by S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin. Springer Berlin Heidelberg, pp. 359–368. DOI: 10.1007/978-3-642-24571-8_47.

Layher, **G.**, Liebau, H., Niese, R., Al-Hamadi, A., Michaelis, B., and Neumann, H. (2011a). "Robust Stereoscopic Head Pose Estimation in Human-Computer Interaction and a Unified Evaluation Framework". In: *Proceedings of the 16th International Conference on Image Analysis and Processing (ICIAP), 2011*. Ed. by G. Maino and G. L. Foresti. Springer Berlin Heidelberg, pp. 227–236. DOI: 10.1007/978-3-642-24085-0_24.

Layher, **G.**, Neumann, H., Scherer, S., Tschechne, S., Brosch, T., and Curio, C. (2011b). "Social Signal Processing in Companion Systems: Challenges Ahead". In: *Proceedigns of the Workshop "Companion-Systeme und Mensch-Companion-Interaktion" (41st Jahrestagung der Gesellschaft für Informatik), 2011*. Vol. P-192. Lecture Notes in Informatics. Gesellschaft für Informatik, pp. 239–239. ISBN: 978-3-88579-286-4.

Strauß, P.-M., Scherer, S., **Layher**, **G.**, and Hoffmann, H. (2010). "Evaluation of the PIT Corpus Or What a Difference a Face Makes?" In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC), 2010*. European Language Resources Association (ELRA), pp. 3470–3474. ISBN: 2-9517408-6-7.

Weidenbacher, U., **Layher**, **G.**, Strauss, P.-M., and Neumann, H. (2007). "A Comprehensive Head Pose and Gaze Database". In: *Proceedings of the 3rd IET International Conference on Intelligent Environments (IE), 2007*. Institution of Engineering and Technology, pp. 455–458. DOI: 10.1049/cp:20070407.

Weidenbacher, U., **Layher**, **G.**, Bayerl, P., and Neumann, H. (2006). "Detection of Head Pose and Gaze Direction for Human-Computer Interaction". In: *Proceedings of the Conference on Perception and Interactive Technologies (PIT), 2006*. Ed. by E. André, L. Dybkjær, W. Minker, H. Neumann, and M. Weber. Vol. 4021. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 9–19. ISBN: 978-3-540-34743-9. DOI: 10.1007/11768029_2.

Conference Abstracts

Zimoch, M., Pryss, R., Probst, T., Schlee, W., **Layher**, **G.**, Neumann, H., et al. (2017). "Evaluating the Comprehensibility of Graphical Business Process Models? An Eye Tracking Study". In: *Journal of Eye Movement Research (ECEM'17 Abstracts)* 10.6, p. 299. ISSN: 1995-8692. DOI: 10.16910/jemr.10.6.

Layher, **G.** and Neumann, H. (2013). "Combining Form and Motion - An Integrated Approach for Learning Biological Motion Representations". In: *Journal of Vision (VSS Abstracts)* 13.9, p. 194. DOI: 10.1167/13.9.194.

Layher, **G.** and Neumann, H. (2012). "Neural mechanisms of action recognition and implied motion." In: *Journal of Vision (VSS Abstracts)* 12.9, p. 142. DOI: 10.1167/12.9.142.

Layher, **G.** and Neumann, H. (2011). "Learning action recognition and implied motion - a neural model." In: *Perception (ECVP Abstract Supplements)* 40.1, p. 176. DOI: 10.1177/03010066110400S102.

II Articles of the Thesis

G. Layher and H. Neumann (2018). "Points and Stripes: A Novel Technique for Masking Biological Motion Point-Light Stimuli". In: *Frontiers in Psychology* 9, p. 12. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2018.01455 53

G. Layher, M. A. Giese, and H. Neumann (2014a). "Learning Representations of Animated Motion Sequences — A Neural Model". In: *Topics in Cognitive Science* 6.1, pp. 170–182. ISSN: 1756-8765. DOI: 10.1111/tops.12075 67

G. Layher, T. Brosch, and H. Neumann (2017a). "Real-time Biologically Inspired Action Recognition from Key Poses using a Neuromorphic Architecture". In: *Frontiers in Neurorobotics* 11.13, p. 19. ISSN: 1662-5218. DOI: 10.3389/fnbot.2017.00013 83

G. Layher, F. Schrodt, M. V. Butz, and H. Neumann (2014b). "Adaptive Learning in a Compartmental Model of Visual Cortex - How Feedback Enables Stable Category Learning and Refinement". In: *Frontiers in Psychology* 5.1287, p. 19. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2014.01287 107

[Layher and Neumann, 2018]

G. Layher and H. Neumann (2018). "Points and Stripes: A Novel Technique for Masking Biological Motion Point-Light Stimuli". In: *Frontiers in Psychology* 9, p. 12. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2018.01455

©2018 G. Layher and H. Neumann.
This is an open-access article distributed under the terms of the Creative Commons Attribution License CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

frontiers in Psychology

ORIGINAL RESEARCH
published: 28 August 2018
doi: 10.3389/fpsyg.2018.01455

Points and Stripes: A Novel Technique for Masking Biological Motion Point-Light Stimuli

Georg Layher and Heiko Neumann**

Institute of Neural Information Processing, Ulm University, Ulm, Germany

Human articulated motion can be readily recognized robustly even from impoverished so-called point-light displays. Such sequence information is processed by separate visual processing channels recruiting different stages at low and intermediate levels of the cortical visual processing hierarchy. The different contributions that motion and form information make to form articulated, or biological, motion perception are still under investigation. Here we investigate experimentally whether and how specific spatio-temporal features, such as extrema in the motion energy or maximum limb expansion, indicated by the lateral and longitudinal extension, constrain the formation of the representations of articulated body motion. In order to isolate the relevant stimulus properties we suggest a novel masking technique, which allows to selectively impair the ankle information of the body configuration while keeping the motion of the point-light locations intact. Our results provide evidence that maxima in feature channel representations, e.g., the lateral or longitudinal extension, define elemental features to specify key poses of biological motion patterns. These findings provide support for models which aim at automatically building visual representations for the cortical processing of articulated motion by identifying temporally localized events in a continuous input stream.

Keywords: biological motion, articulated motion, point-light stimuli, perceptual salience, form, motion

OPEN ACCESS

Edited by:
Haluk Ogmen,
University of Denver, United States

Reviewed by:
Nuala Brady,
University College Dublin, Ireland
Emily D. Grossman,
University of California, Irvine, United States

****Correspondence:***
Georg Layher
georg.l@yher.de
Heiko Neumann
heiko.neumann@uni-ulm.de

Specialty section:
This article was submitted to Perception Science, a section of the journal Frontiers in Psychology

Received: *06 January 2018*
Accepted: *24 July 2018*
Published: *28 August 2018*

Citation:
Layher G and Neumann H (2018) Points and Stripes: A Novel Technique for Masking Biological Motion Point-Light Stimuli. Front. Psychol. 9:1455. doi: 10.3389/fpsyg.2018.01455

1. INTRODUCTION

The investigation of human biological or articulated motion is still a major focus of scientific research. Deeper insights about the key structural features in space-time as well as the underlying computational principles have been revealed since the pioneering work of Johansson (1973). In his investigations of perceptual reference-systems in spatio-temporal cognition he proposed a method of using impoverished displays of biological motion stimuli. There, a set of lights move in coherence with the joints of a subject's skeleton (Johansson, 1973). Observers viewing animated sequences of such *point-light stimuli* (PLS) over a temporal period of 200 ms or longer are capable of reliably recognizing biological motions. Subjects are even able to discriminate different types of motions, such as e.g., *walking to the left*, *walking to the right* or *jumping jack* presented as PLS with stimulus onset intervals (SOI) of larger than 400 ms (Johansson, 1976). Over the last decades of more detailed analysis of the perceptual discriminative capabilities of human observers, the influences of global configurational and more local (or intermediate) level features have been investigated (Mather and Murdoch, 1994; Lange and Lappe, 2006; Garcia and Grossman, 2008). Apart from the impoverishment of the display, PLS contain *both* motion and—at least configurational—form

information. This led to a continuing discussion about the relative contributions of these two visual input sources. The early processing stages of visual cortical processing of biological motion patterns is performed along two major processing streams, namely the ventral *"what"* and the dorsal *"where"* pathway. The two pathways later converge into fused representations at the level of the superior temporal sulcus (STS; Boussaoud et al., 1990; Felleman and Van Essen, 1991). Functional imaging studies (Grossman et al., 2000), as well as single-cell recordings (Oram and Perrett, 1994) indicate the existence of specific mechanisms for the processing of biological motion within this area STS. Further evidence suggests that cells in STS integrate form and motion representations of biological objects (Oram and Perrett, 1996) and selectively respond to face, limb, and whole body motion (Puce and Perrett, 2003).

These findings support model building about the functionality of segregated and interacting visual cortical transformations in biological motion perception. In particular, the modeling approach of Giese and Poggio (2003) suggests that the two parallel motion and form pathways process visual input largely independently over a hierarchy of stages incorporating homologous filtering and integration mechanisms in both pathways. As such, experimental as well as modeling investigations furthered the discussion about the contributions of form and motion information to the observed discrimination performance (Garcia and Grossman, 2008). This debate is still ongoing. Over the last decades, evidences from experimental studies indicated a strong influence of motion based features (Mather and Murdoch, 1994) and highlighted the importance of structural form information as well (Beintema and Lappe, 2002; Lange and Lappe, 2006; Hiris, 2007).

The experimental analysis of the contributions of form and motion information on the perception of biological motion sequences requires different visual parameters of motion and form, or shape, configurations to be isolated from each other. Taking the enormous variety of both motion and form based feature candidates into account, various studies identify different characteristics of low and high level visual cues which enable a reliable and robust recognition of biological motion sequences (Beintema and Lappe, 2002; Casile and Giese, 2005; Hiris, 2007; Thurman and Grossman, 2008; Thirkettle et al., 2009). Complementing a statistical analysis of form and motion contributions on biological motion recognition with a perceptual study, Casile and Giese (2005) conclude that patterns of local opponent motion (LOM) increase the perceptual salience and enable the robust recognition of a PLS. The investigations of Thurman and Grossman (2008) result in a similar conclusion highlighting the importance of LOM patterns by using "temporal bubbles" of point-light sequences embedded in additional noise dot motions. Beintema and Lappe (2002), on the other hand, show that PLS sequences can be robustly recognized without the presence of local image motion information. Further supporting the influence of spatio-configurational biological motion features, Hiris (2007) experimentally show that augmenting non-biological motion with structural information results in perceptual phenomena as reported similarly for biological motion stimuli. The work of Thirkettle et al. (2009) finally shows that the lateral extension of a biological motion PLS pattern correlates with the average human performance in discriminating such a pattern. Although the lateral extension of a PLS is solely defined by the spatial positions of the dots within a PLS pattern at a particular point in time (i.e., one articulated pose), it is is not free of temporal information when being related to the degree of lateral extension in the remaining sequence. Regarding biological motion as *deformation over time*, the maxima in form based feature channels might act as an anchor for characteristic form configurations (e.g., articulated poses). The distictiveness of such canonical representations, or key frames, might allow the reliable and robust recognition of the underlying biological motion.

Due to the properties of human kinetics, such maxima often co-occur with extrema in the motion energy of a displayed motion sequence. Thus, extrema in the motion energy might as well be candidates to act as a mediator indicating occurrences of form configurations with a high perceptual salience.

A key element in analyzing the relative contributions of form or motion based features on the perception of biological motion is the masking technique applied to isolate the feature under investigation and to suppress the immediate recognition of the displayed biological motion. One established way is to embed the spatio-temporal PLS pattern in a field of additional dots, either performing random motions or temporally incoherent motion patterns extracted from the biological motion PLS (e.g., Thurman and Grossman, 2008). The introduced dot motion noise aims at reducing the chance of perceiving the underlying skeleton structure as such and isolating relevant features at threshold level. A common critique of this masking paradigm argues that—particularly in case of random dot motion fields—the extraction of the embedded PLS is in fact facilitated by enabling figure-ground segregation. Focusing on the analysis of the contribution of local dot motion trajectories of a PLS, another class of masking techniques spatially scrambles the dot motions in PLS sequences (Grossman et al., 2000; Kim et al., 2015). This results in a dissolution of the spatial arrangement of the dot patterns and thus in a suppression of configurational form cues. Further PLS masking schemes aim at isolating specific stimulus characteristics by inverting the spatial arrangement (Sumi, 1984; Pavlova and Sokolov, 2000; Troje and Westhoff, 2006) or modifying the number / locations of visible dots anchored on the underlying body skeleton (Dittrich, 1993; Beintema and Lappe, 2002).

In the present contribution we propose a different set of visual stimuli which capture biological motion in PLS but utilize complex patterns of form as overlays radiating from an external reference point that serves as center location for the dynamic form pattern. The main aim of the investigation is to justify the hypothesis derived from a previous modeling investigation (Layher et al., 2014), that articulated motion sequences are segmented by key poses on the basis of specific form defined cues (high degree of articulation). There, events of distinguishing shape characteristics are identified by analyzing simple motion based features (extrema in the motion energy). Thus, the motion information mediates the salience of the form pattern with respect to the biological motion sequence. It has been demonstrated that such key poses capture highly informative

characteristics about poses and actions that allow to already classify activity components by combined form and motion information (Layher et al., 2017). We conjecture that points of maximum limb expansion identify such key poses within biological motion sequences and are correlated to extrema in basic feature parameters, e.g., local maxima in the overall motion energy. In order to isolate the parameters under discussion and degrade additional supporting features, such as a priori knowledge about the type of motion, the recognition of the underlying skeleton, or symmetry effects, we propose a novel masking technique, applying an overlay pattern on the PLS. The temporally varying center point of the dynamic pattern is randomly defined in a spatial neighborhood of the center of gravity of the PLS, while the different lines which radiate from the center point are connected to each of the point-lights in the target stimulus. Together, the dynamic radial line pattern defines a new masking structure which inherits part of the articulated motion but transfers this into the overlay instead of the PLS. Details of the stimulus generation and its structure are discussed in sections 2.2, 2.2.2 and in **Figure 2**. The main motivation of using this new kind of stimuli is that the motion and configural information of the PLS is still intact but the form encoded by the ankle configurations is severely impaired in its visibility through the dominating dynamic radial line patterns that define the mask. Opposed to varying masking techniques, the proposed stimuli prevent the visual system from potentially segregating the PLS from the masking noise stimuli in a figure-ground separation step. Using such stimuli we ran four different experiments to judge whether subjects can rapidly recognize a masked *jumping jack* PLS sequence and to analyze the influence of the lateral and longitudinal extension on the measured human recognition performance. In addition, we analyzed manipulated versions of the masked PLS, which are generated by removing the turning in the motion of the left and the right hands and by temporally shifting the motions of the left body parts.

Our results demonstrate and confirm that characteristic properties of form features define elemental features to specify the occurrence of key configurations, or poses, of biological motion patterns as they occur in natural articulation sequences. Such maxima go synchronously with local extrema in motion energy, as represented in the dorsal pathway of the visual system. Our findings indicate that maxima in form based feature channels (e.g., the lateral and longitudinal extension) do not solely determine the identification of key poses but serve as tags to be further utilized in conjunction with localized temporal events. The identification of such multidimensional feature selection mechanisms provides support for modeling investigations to automatically build visual representations based on key poses in the cortical processing hierarchy of articulated motion.

The correlation between the lateral extension and the perceptual saliency of a biological motion subsequence shown by Thirkettle et al. (2009) for a PLS displaying a walking gait could successfully be reproduced for a jumping jack sequence when further reducing the visual cues in the PLS through the proposed masking technique and even when the human observers are not able to recognize the motion underlying the dot motion pattern.

The rest of the paper is structured as follows. Section 2 describes the acquisition of the motion capture data for the generation of the PLS and further details the creation of the PLS variations, the masking technique, as well as the experimental design and setup. In section 3, the the results of the conducted experiments are presented and interpredted. In the last section 4, the implications of the conclusions drawn from the observed measurements are discussed.

2. METHODS

2.1. Motion Capture Data

Stimuli were generated using data obtained from the *uulm multiperspective action recognition dataset* (uulmMAD[1]) which contains several sets of synchronized motion capture (MoCap) and video sequences. The dataset consists of 31 subjects performing 14 actions from activities of fitness/stretching and everyday life. The subjects' skeletons were recorded using an inertial motion capturing system (XSens MVN Biomech) operating at a sampling rate of 120 Hz (Glodek et al., 2015). Out of the 23 captured skeleton points, a total number of 18 were selected to constitute the PLS (displayed in solid black in the first column of **Figure 2**). The motion capture data of 20 subjects, each performing four repetitions (cycles) of a jumping jack action in three separate recordings, was first normalized in speed (resulting in a final duration of one jumping jack cycle of 150 samples/1.25 s), orientation (0° hip rotation) and height (with a longitudinal distance between the head and the left toe of 180 cm in a neutral pose) per person and then averaged over all subjects. The final stimulus consisted of 1.5 jumping jack cycles (225 samples/1.875 s) and was subsampled to match a temporal resolution of 100 Hz using shape-preserving piecewise cubic interpolation (resulting in 187 frames). The number of cycles was chosen such that the final sequence not only contains a full cycle of the jumping jack but additionally covers all moments of direction reversal in their temporal context.

2.2. Generations of Point-Light Stimulus Variations

2.2.1. Sequences With Phase Variations

Three variations of the jumping jack PLS were generated for the different experimental configurations. Under all conditions, the 3D motion capture data was first projected onto the image plane by orthographic projection. Motion sequences were generated for three different temporal pattern configurations, namely, a sequence displaying the unmodified PLS jumping jack (*unmodified*), a modified version that contains loopy articulated limb extensions (*looped*), and a shifted counter-balanced articulation sequence (*shifted*; see **Figure 2**).

The *unmodified* version directly uses the sequence of point positions derived from the motion capture data as described in section 2.1.

In the *looped* configuration, the lateral component of the hands' motion direction is inverted whenever one of the hands

[1]The dataset is made publicly available upon request under http://uulm.de/imagedb

changes the direction of motion. Thus, the resulting motion pattern of the hands does not contain any reversal points and the lateral extension is reduced in the mid part of the sequence. The transformation is realized by mirroring the position of the left and right hand around the axis through the left, respectively the right elbow, parallel to the ordinate of the image plane. As a result of the transformation, the hands perform two counter-rotating circular motion patterns (loops). The key rationale for introducing such variations in the unmodified PL displays is to selectively manipulate appearances in the overall body movements. Thus the signature of upper body and arm movements is significantly varied: (a) changes in the limb movement direction are eliminated (*looped*) and (b) the symmetry of the limbs' movements is eliminated (*shifted*). Equal to the *unmodified* sequence, the *looped* sequence consists of 187 frames.

The *shifted* variation aims at reducing the variance in the lateral extension of the stimulus without affecting the longitudinal extension or the motion patterns of the individual points. This is achieved by temporally shifting the motions of the left body parts such that changes in the longitudinal extension caused by motion of right body parts are counterbalanced. The maximum reduction was obtained for a temporal shift of 30 frames (0.3 s), decreasing the standard variation of the observed lateral extension by a factor of 3.8. Compared to the *unmodified* and *looped* sequences, the final number of the frames in the *shifted* configurations is reduced to 157 frames. **Figure 1** visualizes the trajectories of the point-lights for the three stimulus configurations using color encoding. In contrast to the *umodified* sequence, the hands show a circular trajectory for the *looped* variant, while the *shifted* sequence possesses an asymmetrical motion pattern.

2.2.2. Masking

The displays utilized here each consist of PLS combined with dynamic radial line patterns such that the stimulus appearance leads to masked spatio-temporal motion sequences (see **Figure 2**). In a "marionette-like" manner, the spatial patterns are centered to a randomly drawn fixed reference point around the central part of the display and spread toward the points of the PLS dot pattern. Our goal is to identify the relative importance of specific poses during a displayed PLS articulation sequence. We argue that in previous stimulus configurations which have used overlaid spatio-temporal dots as noise patterns one cannot exclude the possible conclusion that an observer is able to segregate the moving PLS from the noise pattern and then solve the given task of, e.g., deciding the walking direction or classifying the articulated motion pattern. The stimuli used here, on the other hand, should be immune to mechanisms of figure-ground segregation. The stimulus strength of the apparent form of the line pattern dominates the global configural appearance of the articulation. Since the external point that serves as the center for the radial line bundle pattern is randomly generated, subjects in the experiment cannot adapt to a spatially symmetric dynamic form for improving the judgment of the point-light motion pattern.

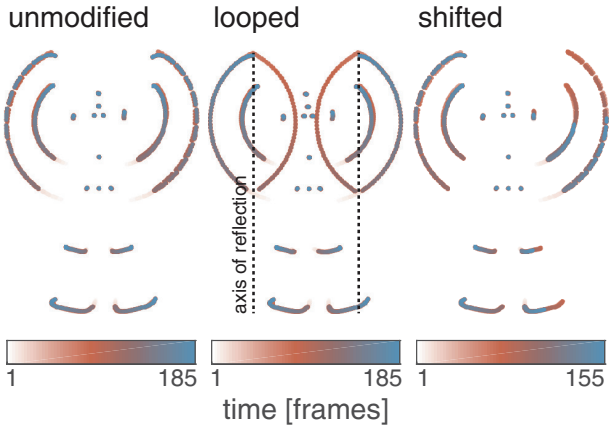


FIGURE 1 | Simulus composition. Three different variations of point-light stimuli were derived from an average normalized motion capture sequence recorded for 18 subjects performing the action of *jumping jack*. In the *unmodified* version, the projections of the 3D motion capture points were directly used to build the stimulus. In the *looped* variant, the horizontal component of the hand positions are transformed such that the hands perform two counter-rotating circular motion patterns. The *shifted* variation was obtained by temporally shifting the point positions of the left body half by 30 frames and results in a minimal variance of the observed lateral extension of the dot pattern. Time is displayed color-encoded from white (first frame) through red to blue (last frame). Note the asymmetry of the left and the right body half for the *shifted* version.

2.3. Stimulus Display

Stimuli were rendered using Matlab 2015b in conjunction with the Psychtoolbox 3 (Brainard, 1997; Kleiner et al., 2007)[2] on an Acer XB270HU monitor with a display diagonal of 27 inches, a resolution of $2,560 \times 1,440$ px and a maximum refresh rate of 144 Hz. A constant distance of 60 cm between the subjects and the screen was ensured using a chin rest. Each of the point-light dots was rendered at a size of 12 arcmin. In case of the masked stimuli, the position of the additional static point was randomly drawn from the range of the maximum bounding box enclosing all original point positions of the original stimulus (approximately 7.82° lateral and 10.60° longitudinal viewing angle) using a uniform distribution. The width of the line interconnecting a point-light dot with the external static dot was set to 2 px. The lines do not fully connect a pair of two points, but end when a 5 px distance to the contour of a point is reached. The point-light stimuli were always presented with the hip point fixed at the center of the display and the frontal plane orthogonal to the viewpoint of the observer. **Figure 2** shows examples of the generated stimuli for both, the original PLS display of the jumping jack (point-light), as well as the configuration with the PLS points connected to two different exemplary static external points (masked) for frame 100. Note how the two different external points change the appearance of the stimulus, although the positions of the remaining point-lights are the equivalent.

[2]http://psychtoolbox.org

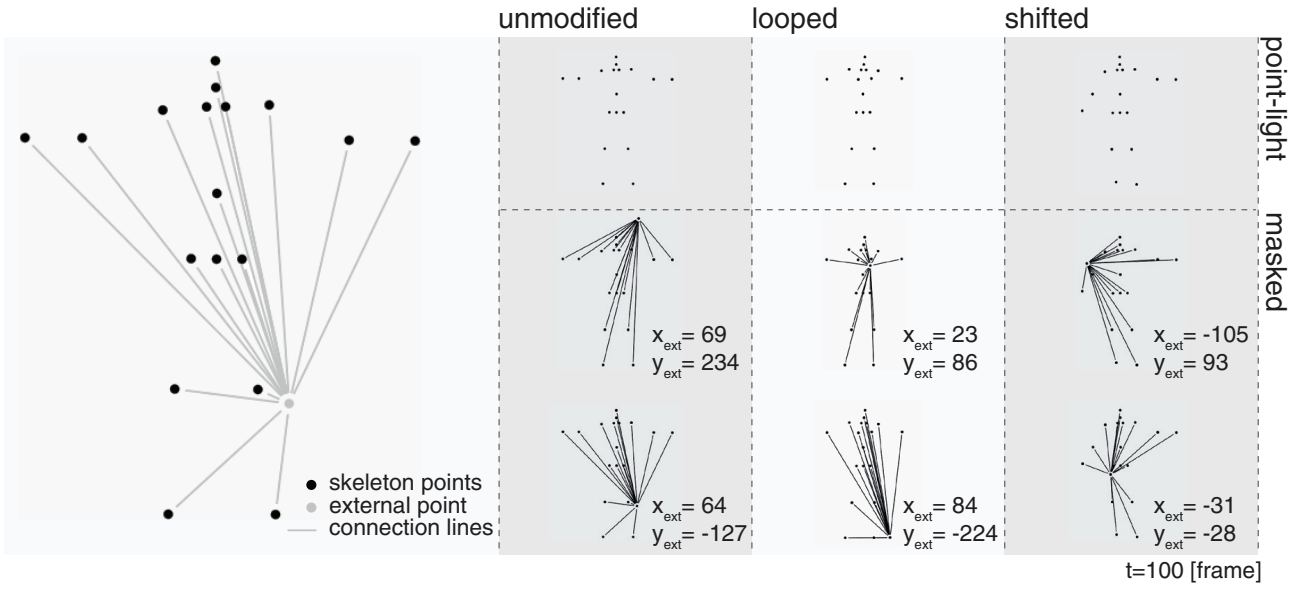


FIGURE 2 | Masking of biological motion point-light simuli. In order to mask the underlying PLS sequence, each of the skeleton points was connected to a randomly drawn static external point. On the left, the skeleton points (black) in one frame are shown alongside with an exemplary external point and the connected radial lines (displayed in gray, the final stimulus was rendered with all points and lines in white on a black background). From the second to the forth column, static snapshots of the unmasked point-light stimuli (first row) and their masked counterparts for two different external points (second and third row) are displayed for the three variations of the jumping jack PLS sequence.

The lateral and longitudinal extension of a point-light pattern is determined by calculating the width and height of the bounding box enclosing all point-lights within one frame. A superposition of the extension in the two directions is calculated by directly summing the two values, logically acting as an "*or*" operation. The resulting values are normalized by dividing through the maximum extension (lateral and longitudinal) over the whole PLS sequence. Equally normalizing the integrated L2-Norm of the point-light motion vectors between two adjacent frames finally gives the relative motion energy.

2.4. Procedure

Since the main focus of the presented study was to analyze the effectiveness of the proposed masking scheme and to allow for a direct comparison, we closely followed the experimental paradigm described in Thirkettle et al. (2009) by employing the same temporal two-alternative forced choice (2AFC) task and adopting the reported threshold duration of 50 ms. The three PLS sequence variations were subdivided in non-overlapping temporal windows of 50 ms length (5 frames), resulting in 37 subsequences for the *unmodified* and the *looped* sequence, as well as 31 for the *shifted* stimulus variation. For each stimulus presentation, a different external static point was selected randomly.

A second sequence of the same length was derived by temporally phase-scrambling the motion patterns of the skeletal points represented in the PLS. This was achieved by randomly drawing an individual temporal starting point for each of the points (from a uniform distribution) and merging the respective 50 ms sequences into one (now incoherent) motion pattern. The same external point was used for the masking of the phase-intact and the phase-scrambled incoherent dot motion patterns. The two patterns were presented in a random temporal order. For each stimulus pair, participants had to decide which of the two 50 ms subsequences was extracted from the original sequence. After the experiment, the subjects' performance in separating the phase-scrambled control from the phase-intact pattern is used as an indicator for the perceptual salience of the particular subsequence.

Since the masking of the dot patterns through the connecting lines results in a different visual appearance of each stimulus, the complete phase-intact masked PLS sequence was presented prior to the display of the 50 ms phase-intact subsequence and the phase-scrambled control. A fixation cross was displayed for a duration of 1 s before and after the presentation of the full sequence and a blank screen was shown for 1 s between the phase-intact test and the phase-scrambled control pattern. The resulting total presentation time per stimulus condition was 4,950 ms for the *unmodified* and the *looped* variations, and 4,650 ms for the *shifted* PLS sequence. All stimuli configurations were presented twice per subject and in randomized order. Examples of the four stimulus variations are included as video sequences in the supplement to this contribution (100 fps playback speed; **Supplementary Videos**).

The experimental procedure was introduced to the participants by providing an explanation of the display setup and the task. However, no details were given in any explanation about the background of the experimental question and in particular terms like "biological motion," "human," or "action" were avoided. Subjects were solely instructed to decide whether the

first or the second of the two presented dot motion patterns was a subsequence of the PLS sequence displayed before. After the completion of an experimental session, each subject was asked to provide a description of the nature of the dot motion patterns they observed during the experiment using a questionnaire.

2.5. Participants

Forty-one healthy naïve subjects (average age 27.46, std 4.40) with normal or corrected to normal vision participated in the experiment. The *unmodified* stimulus configurations were presented in random order together with the *looped* variations (see **Figure 2**) to a group of 21 subjects. A second group of 20 subjects performed the experiment with the same *unmodified* stimulus configurations, but in addition with the *shifted* variations presented in random order as well. After a short break, ten out of the 20 subjects additionally performed the same experiment but with unmasked point-light stimuli (see **Figure 2**). In doing so, we aimed at reproducing the results reported for the unmasked PLS display of a jumping jack in Thirkettle et al. (2009). The presented study was conducted according to the ethical guidelines set out in the WMA Declaration of Helsinki (ethical committee approval was granted: 196/10-UBB/bal) with written consent from all subjects. The study protocol was approved by the ethics committee of the University of Ulm[3].

[3]Helmholtzstrae 20, 89081 Ulm, Germany.

3. RESULTS

3.1. Experiment I

Thirkettle et al. (2009) observed a strong correlation between the lateral extension of a PLS pattern and the average human performance in discriminating 50 ms phase-intact sequences from phase-scrambled controls using a PLS display of a human walking gait. In a variation of the experiment, Thirkettle et al. (2009) replaced the walker with a PLS display of a jumping jack sequence. Contrary to the results obtained for the walker, no correlation between the lateral extension of the PLS and the detection performance was observed. A potential explanation for this degradation is given by the strong symmetry of the point arrangement in the motion pattern which is corrupted in the temporally scrambled stimulus and thus an easy recognition of the target stimulus is possible.

In the first experiment of the present study, we aim at replicating the results reported in Thirkettle et al. (2009) by using an unmasked unmodified jumping jack point-light sequence following the described experimental paradigm as close as possible. **Figure 3** shows the average discrimination rates of ten naïve subjects for the unmodified point-light configuration without masking. Independent of the temporal window the discrimination rate is almost constant at ceiling level around 0.98 (mean: 0.98, std: 0.04). As reported in Thirkettle et al. (2009), the discrimination of the phase-intact motion pattern from the phase-scrambled control is no difficulty for unmasked PLS and

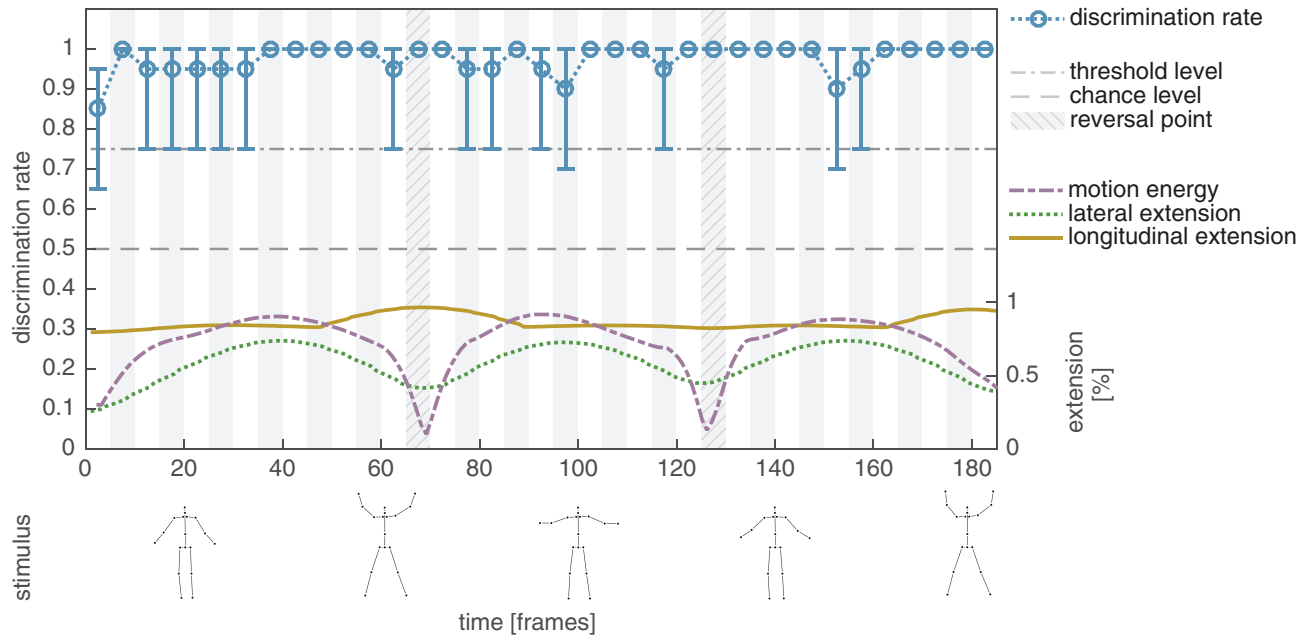


FIGURE 3 | Experiment I, average discrimination performance for the unmodified point-light jumping jack sequence. The average discrimination rate of ten subjects is plotted together with the 95% confidence interval derived by applying bias corrected and accelerated percentile bootstrapping with 100,000 bootstrapping samples (blue). In addition, the normalized lateral (green) and longitudinal (yellow) extent of the minimal bounding box enclosing all points of the PLS per frame. In the bottom row, exemplary snapshots of the stimuli are shown. For a better traceability, point-lights are additionally connected by lines indicating the underlying skeleton. The skeleton lines are not shown in the displays of the experiments. As reported in Thirkettle et al. (2009), the average discrimination rates show no correlation to the lateral extension of the unmasked PLS for a jumping jack sequence.

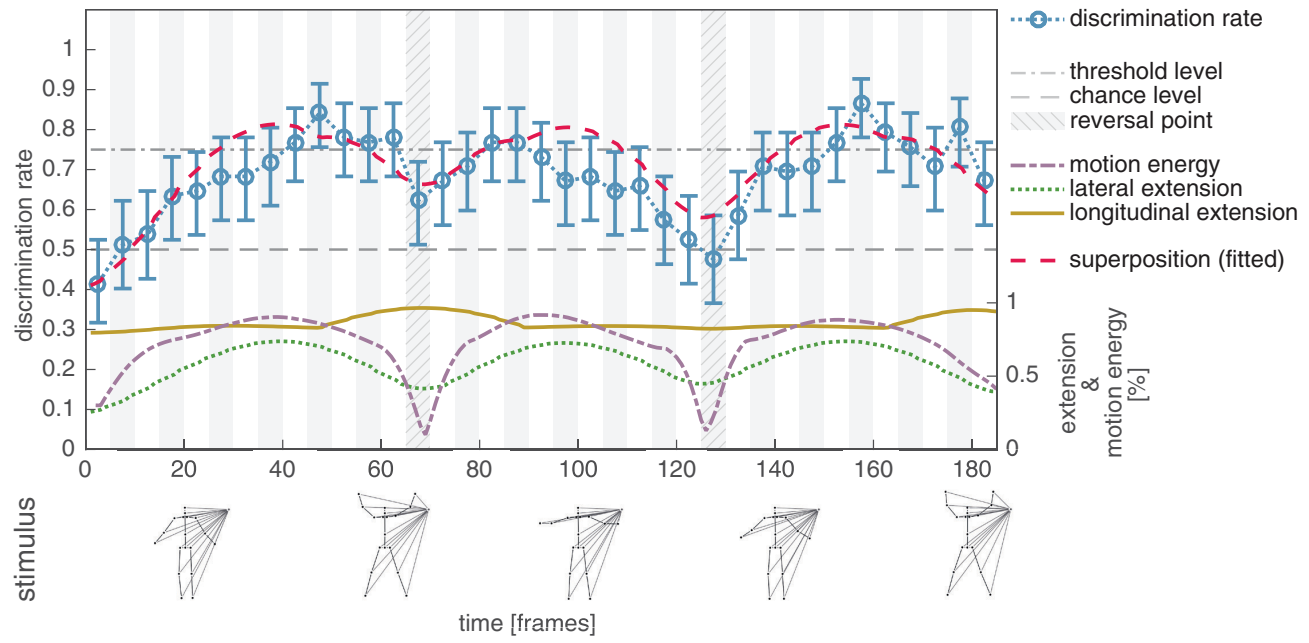


FIGURE 4 | Experiment II, average discrimination rate for the unmodified masked jumping jack sequence. The average discrimination rate of 41 subjects is plotted together with the bootstrapped 95% confidence interval (blue, compare **Figure 3**). In contrast to the unmodified PLS sequence in Experiment I, the average discrimination rate now shows a correlation to the lateral extension (dotted green) of the masked sequence ($r = 0.67, p < 0.001$). The discrimination rate for the subsequence covering frames 66–70 is clearly increased compared to frames 126–130, although the lateral extension is at the same level. This might indicate, that next to the lateral extent, the longitudinal extent (solid yellow) also contributes to the perceptual salience of the stimulus subsequence. This hypothesis is further supported by a strong correlation between the superposition of the lateral and longitudinal extent (dashed red) with the average detection performance ($r = 0.83, p < 0.001$).

thus no correlation to the normalized lateral extension of the stimuli can be revealed ($r = 0.10, p > 0.5$).

After the conduction of the experiment all participants were able to directly identify the observed unmasked dot motion pattern as a display of a human performing the action *jumping jack*.

3.2. Experiment II

In the second experiment the dot motion patterns from Experiment I were masked by lines connecting the points in the PLS to an additional external point (see section 2.2.2 for a detailed description of the stimulus configuration and the logic behind). As expected, this further reduces the visual cues in the PLS and results in a decreased average discrimination rate (mean: 0.68, std: 0.10). However, in contrast to Experiment I, it is now possible to observe a correlation between the lateral extension of the PLS dot pattern and the discrimination performance ($r = 0.67, p < 0.001$). **Figure 4** shows the average discrimination performance, alongside with the relative lateral and longitudinal extension of the PLS display, as well as the relative amount of motion energy.

The displayed results indicate that, next to the lateral extension, the longitudinal extension also influences the recognizability of a PLS subsequence (thus implicating an increase of the perceptual salience). In particular, the two subsequences covering frames 66–70 and 126–130 are almost equal in their normalized lateral extension (averages of 0.42 and 0.46), but differ in the average of their normalized longitudinal spread (0.96 vs. 0.82). The measured human discrimination performances reflect this difference between the two subsequences (mean: 0.62, std: 0.13 for starting frame 66; mean: 0.48, std: 0.06 for starting frame 126). Correlating a measure combining the lateral and longitudinal extent by superposition ($r = 0.83, p < 0.001$) further supports the conjecture that, alongside the lateral extension, the longitudinal extension might increase the perceptual salience of a subsequence.

After the experiment, about two thirds (68.29%) of the participants were not able to name the kind of motion underlying the presented stimuli in the subsequent questionnaire. The remaining third (31.71%) characterized the sequences as biological or human motion. Note that the described effects were observed although all of the stimuli had a different visual appearance and the majority of the participants did not recognize the motions as biological. Compared to unmasked PLS stimuli, or an embedding in random dot motion patterns, the proposed masking thus reduces the chance of an influence on the discrimination rate of high level processes, such as skeleton fitting or the utilization of a priori knowledge about the kind of action displayed in the PLS sequence. Together, these results indicate, that by applying the described masking technique, the effect reported in Thirkettle et al. (2009) can also be observed for

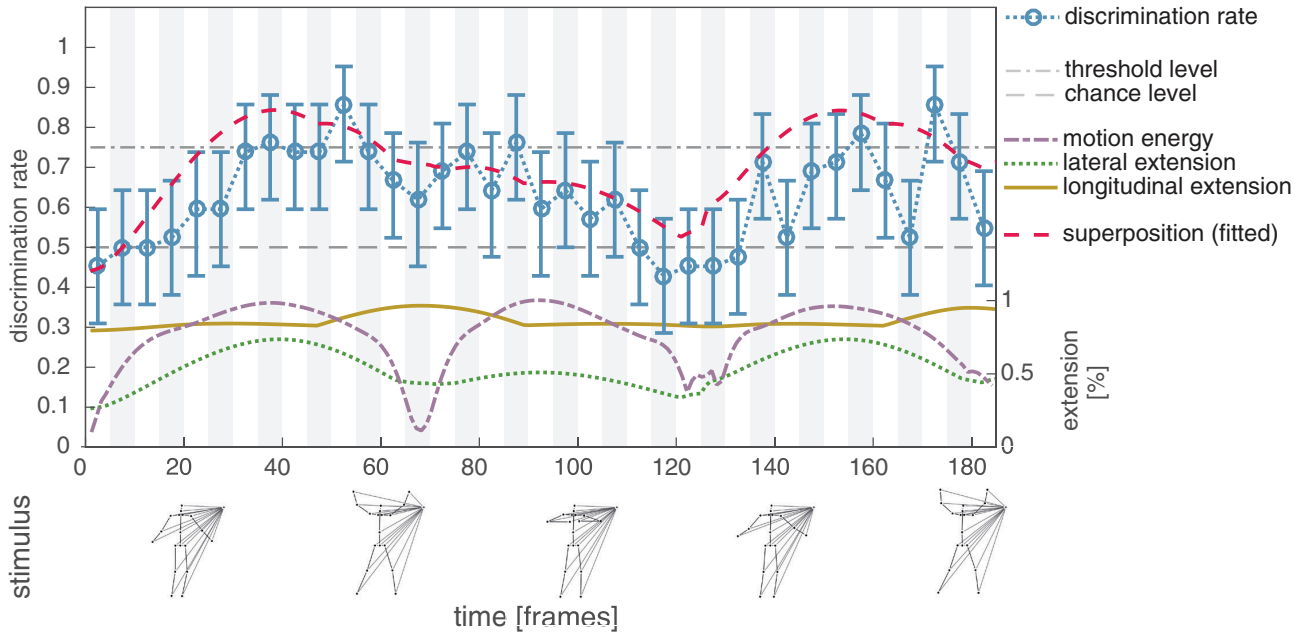


FIGURE 5 | Experiment III, average discrimination rate for the looped masked jumping jack sequence. The average discrimination rate of 21 subjects is plotted together with the bootstrapped 95% confidence interval (blue, compare **Figure 4**). As in Experiment II, the discrimination rate is strongly correlated with the combined (dashed red) lateral (dotted green) and longitudinal (solid yellow) extension ($r = 0.71, p < 0.001$). Removing the turning points of the hand does not have an effect on the measured discrimination rates, whereas the decreased lateral extension slightly reduces the discrimination performance.

a jumping jack sequence after reducing visual cues and that the increased perceptual salience of point patterns with a large spread is not dependent on the direct perception of a biological motion PLS as such.

3.3. Experiment III

In the third experiment we modified the PLS sequence such that the hands as part of the most extendable body configuration follow a counter-rotating circular movement, while the remaining body parts perform the unaltered jumping jack (see **Figure 1**; for details refer to section 2.2). This modification has two major effects on the resulting stimuli. First, the lateral extension is reduced during the middle part of the sequence. Second, due to the counter-rotating motion pattern of the hands the occurrence of LOM patterns is increased within this region. For the *looped* stimuli, participants achieved an average discrimination rate similar to Experiment II (mean: 0.63, std: 0.12). As in Experiment II, a correlation between the human discrimination performance and the combined lateral and longitudinal extension can be observed for the *looped* stimuli ($r = 0.71, p < 0.001$; see **Figure 5**). The discrimination rate for the modified parts of the sequence (subsequences with starting frames $\in [66, ..., 126]$), however, is slightly decreased (mean: 0.60, std: 0.13) when compared to the *unmodified* stimuli (mean: 0.65, std: 0.10). A direct effect on the discrimination performance measured for subsequences with starting frame 66 (mean: 0.62, std: 0.13 for *unmodified*; mean: 0.62, std: 0.08 for *looped*) and 126 (mean: 0.48, std: 0.06 for *unmodified*; mean: 0.45, std: 0.08 for *looped*) caused by the missing turn in direction of the hands cannot be observed.

Notably, the discrimination rate in between starting frames 66 and 131 decreases only marginally compared to the performance measured in Experiment II although the lateral extension within the subsequence is clearly decreased. Possibly, the characteristic local (opponent) motion patterns introduced by the counter-rotating circular motion of the hands support the discrimination of the phase-intact target and phase-scrambled control. Again, this might be an indication, that the perceptual salience of a subsequence is not driven by characteristic occurrences (e.g., local maxima) in one single, but a variety of feature channels.

3.4. Experiment IV

Experiment IV aimed at systematically reducing the variance of the lateral extension over the whole jumping jack point-light sequence without modifying the local motion patterns of the individual dots (*shifted* stimulus configuration; see section 2.2 for details). In addition, this modification implicitly prevents the occurrence of LOM patterns caused by the hands. By temporally shifting the motions of the left body parts, the standard deviation of the lateral extension was reduced by a factor of 3.8 from 59.71 px to 15.66 px. This results in a decreased average discrimination rate (mean: 0.57, std: 0.12) when compared to the measurements for the *unmodified* stimulus configuration (mean: 0.68, std: 0.10) and thus supports the assumed impact of the lateral extension on the human discrimination performance. The second hypothesis underlying Experiment IV was that decreasing the variance of the

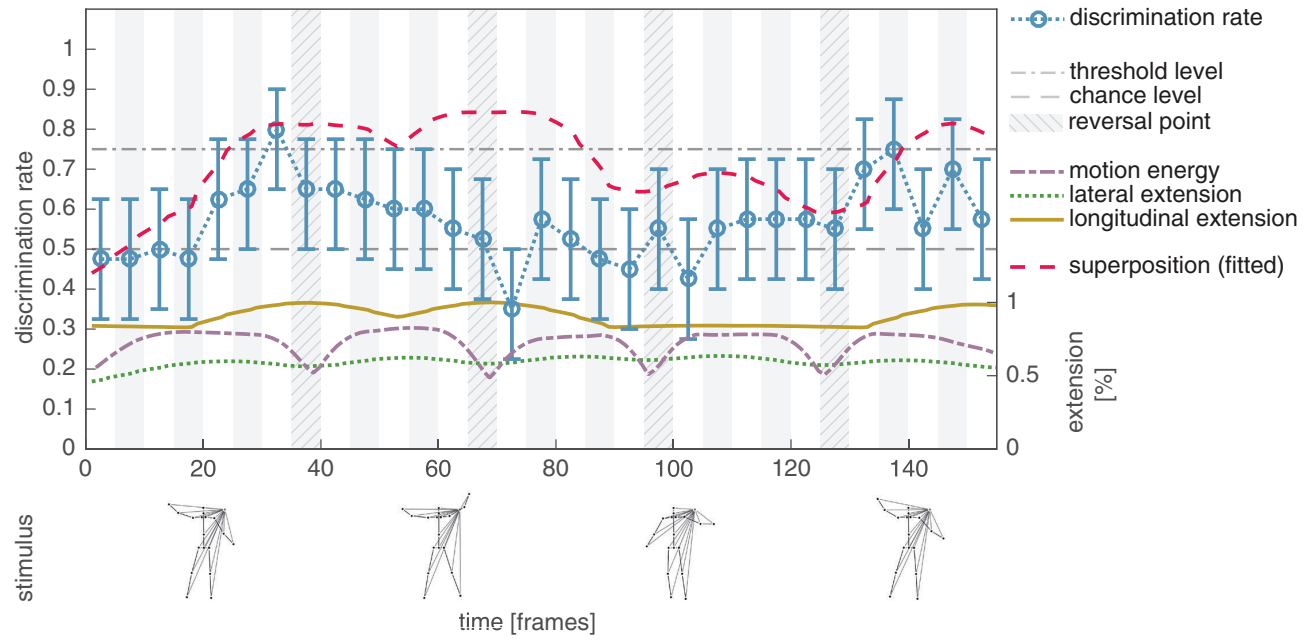


FIGURE 6 | Experiment IV, average discrimination rate for the shifted masked jumping jack sequence. The average discrimination rate of 20 subjects is plotted together with the bootstrapped 95% confidence interval (blue, compare **Figure 4**). Compared to the results of Experiment II, the average discrimination rate is decreased (mean: 0.57, std: 0.12). A high discrimination performance tends to co-occur with an increased longitudinal extension (solid yellow), with exception of the subsequences starting with frames 61, 65, and 71. This results in a decreased correlation between the combined extension feature (dashed red) and the discrimination performance, which might be explained by the statistics of the asymmetric dot patterns within the PLS sequence (see text for details). Nevertheless the results indicate that although there is just little variance in the lateral extension (dotted green) of the PLS, there are distinct events with an increased perceptual salience.

lateral extension results in a clearly observable—since isolated—influence of the longitudinal extension on the perceptual salience of a subsequence, as well as a correlation of the combined extension values to the discrimination performance similar to Experiment II and III. Correlating the summed lateral and longitudinal extension to the discrimination performance ($r = 0.31$, $p = 0.08$), which is dominantly driven by the contribution of the longitudinal extension ($r = 0.33$, $p = 0.07$), does not directly support this hypothesis. In **Figure 6**, a tendency of the discrimination performance to increase with the longitudinal extension of the stimulus can be observed, but with the exception of the subsequences with starting frames $\in [61, ..., 76]$.

Noticeably, the discrimination performance clearly falls below chance level (mean: 0.35, std: 0.08) for the subsequence with starting frame 71, indicating a systematic articulated motion discrimination. In all, these mistake by the subjects in choosing the non phase-scrambled point-light pattern. A possible explanation of this effect might be the asymmetric nature of the point configurations and the statistics of the asymmetric patterns within the whole PLS sequence. Subjects might perceive the configuration displayed from frame 71 to 75 as a deviation from the subsequence starting with frame 41 and as a consequence tend to select the phase-scrambled PLS pattern. This assumption—if applicable—might explain the only fragmentary correlation between the longitudinal extension and the discrimination performance displayed in **Figure 6**, but is a topic that needs to be further analyzed experimentally.

Nevertheless, the results obtained in Experiment IV show that even though the lateral extension is almost constant over the displayed PLS sequence, there are distinct moments in time (events) with an increased discrimination rate (indicating a higher perceptual salience). These events are partially co-located with increasing values of the longitudinal extension but given the obtained results we cannot exclude that another feature might be causally responsible for this effect.

4. DISCUSSION

4.1. Articulated Motion, Perceived Pattern, and Its Masking

The perception of articulated motion, or biological motion, is still a target of detailed investigation. Perceptual degradations have been studied to reveal the contributions of different representations and dynamic mechanisms in visual cortex which contribute to such coherent form-in-motion percepts. In this contribution we introduced a novel PLS masking approach, which allows to selectively impair the ankle information of the body configuration while keeping the motion of point-light locations intact. The mask is centered at a (randomly selected) anker point with radial line connections from this point to each joint represented in the PLS. The line endings leave a small distance to the individual point-light in order to leave the original PLS items unimpaired. The rationale to employ this new configuration is to counteract possible

spatio-temporal groupings in the PL walker configurations. Coherently moving dots can give rise to contour grouping effects, following the Gestalt laws of proximity and common fate (Metzger, 1936; Stevens, 1978; Metzger, 2006; Thirkettle et al., 2009; Wertheimer, 2012). The neural mechanisms underlying such grouping phenomena have been explored to a certain degree but remain elusive when complex dynamic item configurations are displayed. Inspired by Ullman (1984), Roelfsema distinguished different grouping mechanisms, namely, base-grouping and incremental grouping, respectively, the first operating automatically and fast utilizing primary features, while the latter dynamically links multiple features flexibly taking more time (Roelfsema, 2006). Experimental evidence suggests that feature-based base-groupings are encoded in the feedforward sweep of cortical processing while incremental encodings are mediated by horizontal and feedback processing utilizing knowledge and contextual information to modulate lower-level neural representations (compare Phillips et al., 2015).

The radial line mask pattern proposed in this study comprises a dynamic physical luminance pattern that strongly overcasts any dynamic shape configurations imposed by the dynamic point patterns. In order to leave the configuration of the PLS itself unimpaired we left the line endings disconnected to the individual dots. Our reasoning is that any groupings which might be dynamically established to segregate spatio-temporal figural patterns from background (Sporns et al., 1991) will operate incrementally and, thus, on a slower time-scale. Such line mask pattern will be immediately visible and the spatio-temporal effects created by such masks are mainly determined by base-groupings. Based on such segregation of different temporal phases of neuronal processing, we believe that the individual studies conducted in the reported experiments allow us to dissect the contribution of different stimulus elements in the perceptual judgement of biological motion. While we are investigating the selective manipulation of articulated limb trajectories - represented in the still intact PLS—the radial line mask reflects the primary pattern but suppresses secondary grouping operating on slow temporal scales.

4.2. Experimental Investigation of Articulation Patterns and Their Temporal Structure

Following the experimental methodology and paradigm employed in Thirkettle et al. (2009), we studied the effect of three masked PLS sequence variations on the measured performance in discriminating a phase-intact target PLS subsequence and a phase-scrambled control. In line with (Thirkettle et al., 2009), the results show that the lateral extension of a PLS pattern correlates with the average discrimination performance and thus is a candidate to act as a predictor of the perceptual salience of specific moments in the course of articulated motion sequences. In addition, the reported results show indications, that the longitudinal extension of a point-light pattern might as well support the perceptual salience of a PLS subsequence (*unmodified*). Decreasing the variance of the lateral extension by modifying the motion trajectory of the hands alone only resulted in a slight decrease of the discrimination rate, thus opening the question whether the implicitly introduced local (*opponent*) motion patterns might be responsible for an increased perceptual salience (*looped*). In a further experiment, the PLS sequence was modified, such that the lateral extension was almost constant and the occurrence of LOM pattern was excluded (*shifted*). The measured discrimination performance was clearly decreased, which suggests less importance of flow contrast detection as specific feature for articulated motion discrimination.

In all, these results indicate, that first order changes in the representations along characteristic feature channels, such as e.g., extrema in the lateral extension of the body configuration during an articulation, might act as triggers or mediators for the establishment of key representations used to identify PLS sequences (Nothdurft, 2000a,b).

In Experiment I, we first reproduced the results reported in Thirkettle et al. (2009) for separating 50 ms subsequences of an unmasked PLS displaying a jumping jack from phase-scrambled control patterns in a temporal 2AFC task (Experiment I; section 3.1). In both displays, the discrimination performance saturated almost at ceiling. In contrast, the results obtained in Thirkettle et al. (2009) for a PLS sequence displaying a human walking gait showed a direct correlation between the measured task performance and the lateral extension of the point-light pattern. Thirkettle et al. (2009) propose that the constant task performance found for the jumping jack PLS sequence might possibly be explained by the exploitation of symmetry properties of the point-light configurations and the detection of violations thereof. Furthermore, the authors argued that the measured performance is strongly dependent on the displayed action, as well as the experimental task and methodology (e.g., the kind of noise, which is applied on the stimulus).

Accordingly, Thurman and Grossman (2008) obtained different results for experimentally analyzing the perceptual salience of subsequences of a jumping jack PLS using a different methodology. There, subsequences with different starting frames were embedded in a field of noise dot motions at randomized spatial positions with trajectories drawn from the original sequence but with individual starting frames. The target dot motion patterns emerged as "temporal bubbles" within random dot motion sequences and subjects were asked to discriminate sequences containing a target pattern from scrambled motion controls in a threshold task. Experiments were conducted to identify subsequences containing key features which drive an increased perceptual salience. For a jumping jack, as well as walking gait PLS sequence, the measured task performance clearly indicated that patterns of opponent motion might constitute such a key feature (supporting the role of opponent motion as a feature to perceive biological motion; Casile and Giese, 2005). Opponent motion is most pronounced when the relative distance between the arms and the legs is smallest. This result is—in case of the walking gait PLS—in conflict with the results reported in Thirkettle et al. (2009), where subsequences containing highly articulated postures with a large

lateral extension were identified as perceptually more salient. As discussed in Thirkettle et al. (2009), these contradictory results might be explained by low-level effects caused either by the density of the noise dots used for the masking in case of a threshold task, or, on the contrary, by the modified spread of dots in discriminating a target pattern from a phase-scrambled control.

In our study in Experiment II, we applied the masking technique as described in section 2.2.2. As a result of reducing the visual cues we were able to show a correlation between the extent of a jumping jack PLS pattern and the performance in discriminating it from a phase-scrambled control. The same relationship was reported in Thirkettle et al. (2009) for an unmasked walking PLS, but could not be replicated for an unmasked jumping jack PLS. Although we cannot guarantee that the proposed new masking method excludes the influence of low-level cues introduced by the phase-scrambling, we argue that this effect is at least weakened. This is likely since the visual appearance of the biological motion pattern is altered for each individual stimulus by the randomly positioned additional external point. Introducing the mask pattern reduces the chances that fast base-groupings are established which provide the input to higher-level processes, such as e.g., skeleton fitting. Even more, the potential utilization of prior knowledge about the kind of action pattern that guides incremental groupings is diminished. The semantics of the motion signature is extinguished as indicated by the separate questionnaire to identify the pattern (after the experiment only one third (31.71%) of the subjects was able to identify the observed masked PLS as a biological or human motion). Local motion sub-patterns, such as LOM, positioned at locations internal to the figural pattern are suppressed as well. Temporal maxima in the whole body appearance here co-occur with maxima in feature channel activations detected by observers. Our experimental procedure thus enables segregating the perceptual influences of extremal global feature properties in global configural appearances from local salient features such as LOMs.

The other two experiments were designed to manipulate the spatio-temporal signature of the jumping jack articulation sequence under the masking, particularly the spatio-temporal features leading to extrema in global shape appearance. In Experiment III, the PLS pattern with the adapted mask are manipulated so that the hands follow circular movements. As a consequence, the lateral extension is reduced, however, the occurrence of LOMs is increased. In other words, the strength of the salient extremal features for global discriminating shape features is diminished. At the same time the frequency and integrated strength of LOM is increased again. We expected that now with the elimination of the global pattern extrema the descrimination rate is significantly reduced. The observed effect was less strong than expected. We conclude that the discriminability is somewhat restored by the increase in LOM patterns. We should not, however, draw any conclusions about the relative weights of the perceptual contribution of the local and the global features.

Experiment IV was finally designed to systematically reduce the variance of the pattern of lateral body pose extension. By shifting the temporal movement phases of the left body parts over the whole motion sequence the lateral extension of the global appearance pattern remains almost constant. At the same time, though, the occurrence of LOMs is diminished. In the discrimination experiments observers identified distinct temporal events which were characterized by longitudinal extensions of the global appearance pattern. In terms of causal relations between features and the detection of key pose events, we cannot unambiguously identify a conjunction of such properties with the detection of features or extrema in the feature dimensions discussed, such as LOMs or local extrema in motion energy. Such distinct events in an otherwise rather homogeneous spatio-temporal pattern might be identified by extrema in the temporal evolution of configural patterns and their appearance. Here, it would be reasonable to investigate whether such extrema may also be detected and judged as key events for spatio-temporal appearances generated by non-human shapes, such a deformable objects of arbitrary shape.

4.3. Features Indicating Characteristic Articulated Motion Patterns

The observed correlation between the combined lateral and longitudinal extension of the PLS and the human recognition performance might provide an indication on the characteristic properties of salient form features in the context of biological motion perception. While occurrences of a lateral or longitudinal extension seem to have an impact on the perceptual salience of a PLS subsequence, it is the temporal relationship which renders this presence of characteristic form information special and distinctive. We hypothesize that local temporal maxima in (potentially a variety of) feature channels form the basis for an increase in the perceptual salience of a biological motion sequence. In other words, the conjecture advocated here is that first order mechanisms operate upon the feature representations to register local temporal contrasts in the respective feature channels (or combinations of these). They, in turn, generate event-triggers to serve as a marker, e.g., to steer the learning of form and motion pattern characteristics. We do not claim that these feature channels are only to be found in the spatial domain. On the contrary, we are convinced, that dependent on the type of action displayed in the PLS, e.g., patterns of LOM (Casile and Giese, 2005) are candidates to explain an increased perceptual salience just as well and are in line with this hypothesis. It is worth to note that due to the nature of human kinetics, such temporal maxima in feature channel activations co-occur with local extrema in the energy of whole body or limb motions (see **Figure 4**). Thus, motion energy in turn may be a candidate to tag local maxima of form/motion based feature characteristics.

AUTHOR CONTRIBUTIONS

GL developed the presented stimulus design and the overall setup of the experiment. He implemented and conducted the experiments together with students, as well as performed the interpretation and statistical analysis of the experimental data. HN developed the overall design of the experiments in collaboration with GL, participated in the design and the

conduction of the experiments, as well as in the interpretation of the experimental results.

FUNDING

This research has been supported by the Transregional Collaborative Research Centre SFB/TRR 62 A Companion Technology for Cognitive Technical Systems funded by the German Research Foundation (DFG). HN has been supported by the collaborative project SenseEmotion funded by the German Federal Ministry of Education and Research (BMBF).

ACKNOWLEDGMENTS

The authors gratefully thank Sebastian Thoma and Felix Traub, who supported the implementation and the conduction of the experiments presented in this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01455/full#supplementary-material

REFERENCES

Beintema, J. A., and Lappe, M. (2002). Perception of biological motion without local image motion. *Proc. Natl. Acad. Sci. U.S.A.* 99, 5661–5663. doi: 10.1073/pnas.082483699

Boussaoud, D., Ungerleider, L. G., and Desimone, R. (1990). Pathways for motion analysis: cortical connections of the medial superior temporal and fundus of the superior temporal visual areas in the macaque. *J. Comp. Neurol.* 296, 462–495. doi: 10.1002/cne.902960311

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vis.* 10, 433–436. doi: 10.1163/156856897X00357

Casile, A., and Giese, M. A. (2005). Critical features for the recognition of biological motion. *J. Vis.* 5:6. doi: 10.1167/5.4.6

Dittrich, W. H. (1993). Action categories and the perception of biological motion. *Perception* 22, 15–22. doi: 10.1068/p220015

Felleman, D. J., and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1, 1–47. doi: 10.1093/cercor/1.1.1

Garcia, J. O., and Grossman, E. D. (2008). Necessary but not sufficient: motion perception is required for perceiving biological motion. *Vis. Res.* 48, 1144–1149. doi: 10.1016/j.visres.2008.01.027

Giese, M. A., and Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nat. Rev. Neurosci.* 4, 179–192. doi: 10.1038/nrn1057

Glodek, M., Layher, G., Heilemann, F., Gawrilowicz, F., Palm, G., Schwenker, F., et al. (2015). "uulmMAD-a human action recognition dataset for ground-truth evaluation and investigation of view invariances," in *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*, Lecture Notes in Computer Science (Cham: Springer), 77–91. doi: 10.1007/978-3-319-14899-1_8

Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., et al. (2000). Brain areas involved in perception of biological motion. *J. Cogn. Neurosci.* 12, 711–720. doi: 10.1162/089892900562417

Hiris, E. (2007). Detection of biological and nonbiological motion. *J. Vis.* 7, 4.1–4.16. doi: 10.1167/7.12.4

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Percept. Psychophys.* 14, 201–211. doi: 10.3758/BF03212378

Johansson, G. (1976). Spatio-temporal differentiation and integration in visual motion perception. *Psychol. Res.* 38, 379–393. doi: 10.1007/BF00309043

Kim, J., Jung, E. L., Lee, S.-H., and Blake, R. (2015). A new technique for generating disordered point-light animations for the study of biological motion perception. *J. Vis.* 15:13. doi: 10.1167/15.11.13

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., and Broussard, C. (2007). What's new in psychtoolbox-3. *Perception* 36, 1–16. doi: 10.1068/v070821

Lange, J., and Lappe, M. (2006). A model of biological motion perception from configural form cues. *J. Neurosci.* 26, 2894–2906. doi: 10.1523/JNEUROSCI.4915-05.2006

Layher, G., Brosch, T., and Neumann, H. (2017). Real-time biologically inspired action recognition from key poses using a neuromorphic architecture. *Front. Neurorobot.* 11:13. doi: 10.3389/fnbot.2017.00013

Layher, G., Giese, M. A., and Neumann, H. (2014). Learning representations of animated motion sequences–A neural model. *Top. Cogn. Sci.* 6, 170–182. doi: 10.1111/tops.12075

Mather, G., and Murdoch, L. (1994). Gender discrimination in biological motion displays based on dynamic cues. *Proc. R. Soc. Lond. B Biol. Sci.* 258, 273–279. doi: 10.1098/rspb.1994.0173

Metzger, W. (1936). *Gesetze des Sehens.* Frankfurt/Main: Verlag W Kramer & Co.

Metzger, W. (2006). *Laws of Seeing*. (Tansl. L. Spillmann). Cambridge, MA: MIT Press.

Nothdurft, H.-C. (2000a). Salience from feature contrast: temporal properties of saliency mechanisms. *Vis. Res.* 40, 2421–2435. doi: 10.1016/S0042-6989(00)00112-7

Nothdurft, H.-C. (2000b). Salience from feature contrast: variations with texture density. *Vis. Res.* 40, 3181–3200. doi: 10.1016/S0042-6989(00)00168-1

Oram, M., and Perrett, D. (1994). Responses of anterior superior temporal polysensory (STPa) neurons to "Biological Motion" stimuli. *J. Cogn. Neurosci.* 6, 99–116. doi: 10.1162/jocn.1994.6.2.99

Oram, M. W., and Perrett, D. I. (1996). Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey. *J. Neurophysiol.* 76, 109–129. doi: 10.1152/jn.1996.76.1.109

Pavlova, M., and Sokolov, A. (2000). Orientation specificity in biological motion perception. *Percept. Psychophys.* 62, 889–899. doi: 10.3758/BF03212075

Phillips, W., Clark, A., and Silverstein, S. (2015). On the functions, mechanisms, and malfunctions of intracortical contextual modulation. *Neurosci. Biobehav. Rev.* 52, 1–20. doi: 10.1016/j.neubiorev.2015.02.010

Puce, A., and Perrett, D. (2003). Electrophysiology and brain imaging of biological motion. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 358, 435–445. doi: 10.1098/rstb.2002.1221

Roelfsema, P. R. (2006). Cortical algorithms for perceptual grouping. *Annu. Rev. Neurosci.* 29, 203–227. doi: 10.1146/annurev.neuro.29.051605.112939

Sporns, O., Tononi, G., and Edelman, G. M. (1991). Modeling perceptual grouping and figure-ground segregation by means of active reentrant connections. *Proc. Natl. Acad. Sci. U.S.A.* 88, 129–133. doi: 10.1073/pnas.88.1.129

Stevens, K. A. (1978). Computation of locally parallel structure. *Biol. Cybernet.* 29, 19–28. doi: 10.1007/BF00365232

Sumi, S. (1984). Upside-down presentation of the Johansson moving light-spot pattern. *Perception* 13, 283–286. doi: 10.1068/p130283

Thirkettle, M., Benton, C. P., and Scott-Samuel, N. E. (2009). Contributions of form, motion and task to biological motion perception. *J. Vis.* 9, 28.1–28.11. doi: 10.1167/9.3.28

Thurman, S. M., and Grossman, E. D. (2008). Temporal "Bubbles" reveal key features for point-light biological motion perception. *J. Vis.* 8, 28.1–28.11. doi: 10.1167/8.3.28

Troje, N. F. and Westhoff, C. (2006). The inversion effect in biological motion perception: Evidence for a "life detector"? *Curr. Biol.* 16, 821–824. doi: 10.1016/j.cub.2006.03.022

Ullman, S. (1984). Visual routines. *Cognition* 18, 97–159. doi: 10.1016/0010-0277(84)90023-4

Wertheimer, M. (2012). *On Perceived Motion and Figural Organization.* ed L. Spillmann (Cambridge, MA: MIT-Press).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Layher and Neumann. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

[Layher et al., 2014a]

G. Layher, M. A. Giese, and H. Neumann (2014a). “Learning Representations of Animated Motion Sequences — A Neural Model”. In: *Topics in Cognitive Science* 6.1, pp. 170–182. ISSN: 1756-8765. DOI: 10.1111/tops.12075

©2013 Cognitive Science Society, Inc. All rights reserved.
The article has been published at DOI: 10.1111/tops.12075 and may be used for non-commercial purposes in accordance with Wiley and Cognitive Science Society Terms and Conditions for Self-Archiving.

TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 6 (2014) 170–182
Copyright © 2013 Cognitive Science Society, Inc. All rights reserved.
ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12075

Learning Representations of Animated Motion Sequences—A Neural Model

Georg Layher,[a] Martin A. Giese,[b] Heiko Neumann[a]

[a]*Institute of Neural Information Processing, Ulm University*
[b]*Section for Computational Sensomotorics, Department for Cognitive Neurology, HIH & CIN, University Clinic Tübingen*

Received 16 September 2013; accepted 18 September 2013

Abstract

The detection and categorization of animate motions is a crucial task underlying social interaction and perceptual decision making. Neural representations of perceived animate objects are partially located in the primate cortical region STS, which is a region that receives convergent input from intermediate-level form and motion representations. Populations of STS cells exist which are selectively responsive to specific animated motion sequences, such as walkers. It is still unclear how and to what extent form and motion information contribute to the generation of such representations and what kind of mechanisms are involved in the learning processes. The article 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 form prototypes characterized by a high degree of articulation. Such key poses are selectively reinforced during learning through a cross talk between the motion and form processing streams. Furthermore, we show how sequence-selective representations are learned in STS by fusing static form and motion input from the segregated bottom-up driving input streams. Cells in STS, in turn, feed their activities recurrently to their input sites along top-down signal pathways. We show how such learned feedback connections enable predictions about future input as anticipation generated by sequence-selective STS cells. Network simulations demonstrate the computational capacity of the proposed model by reproducing several experimental findings from neurosciences and by accounting for recent behavioral data.

Keywords: Animated motion; Articulation; Implied motion; Motion; Form; Neural model; Unsupervised learning; Feedback

Correspondence should be sent to Georg Layher, Institute of Neural Information Processing, Ulm University, James-Franck Ring, 89069 Ulm, Germany. E-mail: georg.layher@uni-ulm.de

1. Introduction

Animated movements in actions, like walking, turning, etc., can be robustly detected from video sequence input and predictions about future occurrences can be derived from such spatiotemporal patterns. Giese and Poggio (2003) proposed a hierarchical feedforward network architecture that aims at explaining the computational mechanisms underlying the perception of body motion, including recognition from impoverished stimuli, such as point-light walkers. In this article, we propose a new learning-based hierarchical model for analyzing animated motion sequences. Prototypes in the form and motion pathways are established using a modified Hebbian learning scheme. We suggest how snapshot prototypes might be automatically selected from continuous input video streams utilizing features from the motion pathway which are indicative for the occurrence of specific snapshots with strongly articulated configurations that serve as key poses. Sequence-selective representations of articulated motions in cortical STS are driven jointly by input activations from both motion and form prototypes. In addition, feedback connections are learned to enable STS neurons predicting expected input from form selective IT and motion-sensitive MST. We argue that for inputs presenting articulated postures without continuing motion, STS representations are fed by the corresponding snapshot prototype activations (Jellema & Perrett, 2003). In turn, STS will send feedback to stages in the segregated pathways for form as well as motion processing. Stationary images which depict articulated postures consequently generate effects of implied motion, which have been shown in functional magnetic resonance imaging (fMRI) studies (Kourtzi & Kanwisher, 2000). We will argue here that this can be accomplished by the proposed model through the action of fusing bottom-up input, driven by snapshot representation only, and the activated sequence representations sending feedback to both form and motion representations, thus amplifying motion representations even if no direct motion input is present.

Several computer vision approaches have been proposed for performing action recognition using different processing strategies of combining form and motion information. These approaches build upon the hierarchical architecture proposed by Poggio and coworkers which aims at defining a framework for form processing in the cortical ventral pathway (Riesenhuber & Poggio, 1999). Extensions of the form processing model to analyze motion information responses in a separate pathway, like the Giese–Poggio model have been suggested (e.g., Jhuang, Serre, Wolf, & Poggio, 2007; Schindler & Van Gool, 2008). Here, the relative contributions of form and motion features to the classification of actions have been investigated. Details of the motion processing cascade alone have been studied more detail (Escobar & Kornprobst, 2012; Escobar, Masson, Vieville, & Kornprobst, 2009). Here, the authors contributed further evidence that detecting motion contrasts in sequences of animated motion is useful to distinguish action classes. In most of these proposed models, the mechanisms for hierarchical motion (and form) processing are predefined and learning only occurs at the level of a final classifier to distinguish given categories. It still remains unclear, to a large extent, how the motion

and form prototypes (e.g., in cortical areas MST and IT, respectively) and the sequence-selective pattern representations in STS interact and which features are used for learning. How can feature representations be learned automatically from given input streams at different levels of the distributed action sequence representations? Also no top-down influences have been considered so far and how such connectivity patterns may transfer different information between pathways and generate proper predictions concerning future input configurations.

2. Model architecture

The hierarchical model proposed here consists of two separate visual pathways for segregated form and motion processing as inspired by the work of Giese and Poggio (2003) and extends it by combining it with models for the hierarchical feedforward and feedback processing of motion and form along the dorsal and the ventral pathway (Bayerl & Neumann, 2004; Weidenbacher & Neumann, 2009). 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. The two separate hierarchical learning approaches are influenced partly by the work of Rolls and collaborators (Rolls & Milward, 2000), in which the authors have suggested that layered neuronal structures arranged in a hierarchy with increasingly larger connectivity kernels can learn invariant representations of objects and specific motion patterns. Here, we propose how such learning in a hierarchy can be utilized for learning sequence-selective representations of animated movement prototypes from convergent form and motion input. In addition, we suggest how 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 modified Hebbian mechanisms. An overview of the model is shown in Fig. 1. The details are outlined below.

2.1. Form and motion processing

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 optical flow patterns (areas MT and MST). We use single compartment model neurons with gradual activation dynamics. The membrane potential of individual model neurons is calculated by conductance-based mechanisms of feedforward integration of excitatory and inhibitory feeding input and a passive leakage. The potential can be enhanced by a gating mechanism to amplify the efficacy of the current potential by a matching top-down feedback signal. The membrane potential is finally regulated by a gain control mechanism

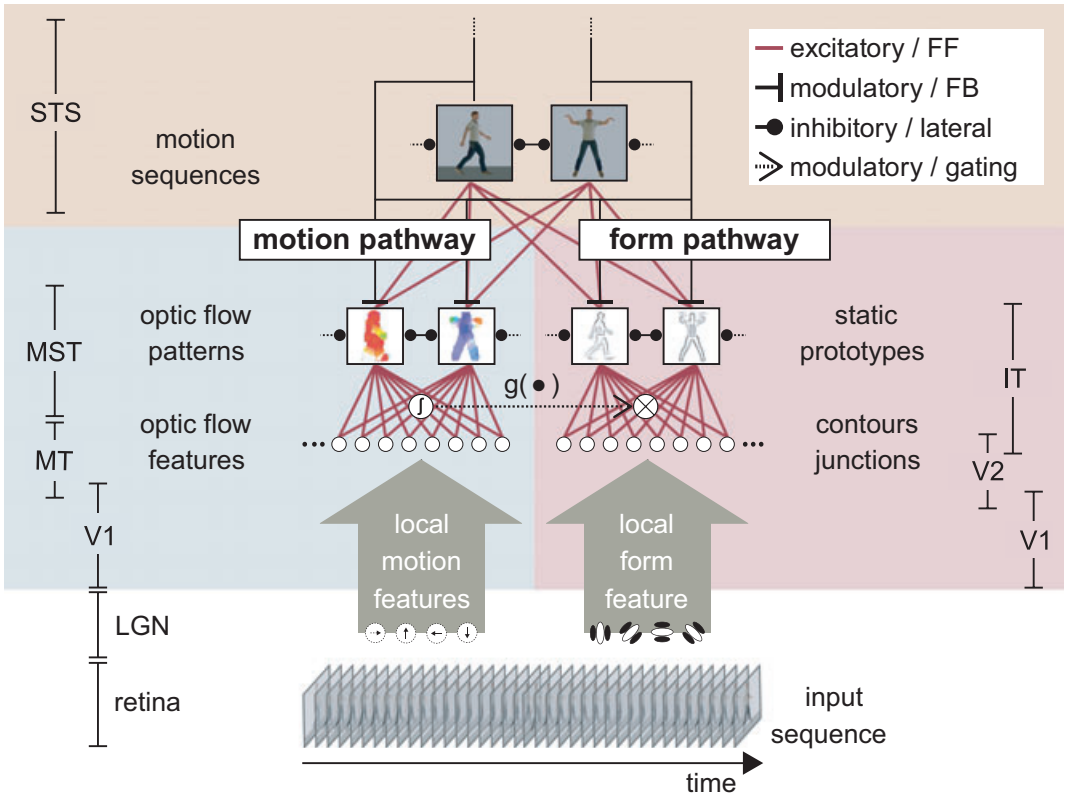


Fig. 1. Overview of the model architecture. The model consists of two separate processing streams, the motion and the form pathway, both converging into model area STS. Static form prototypes in area IT, as well as optic flow patterns in area MST, are learned using an unsupervised Hebbian mechanism. A motion-driven reinforcement signal between the two pathways is used to steer the learning of the IT prototypes. After the suppression of cells with low activities, IT and MST cells propagate into area STS, where sequence-selective cells learn corresponding spatiotemporal activity patterns using a similar Hebbian learning rule. In addition, the sequence-selective cells learn the output weights fed back to the segregated form and motion prototypes, thus stabilizing the input processing and predicting future input configurations.

that leads to activity normalization for a pool of neurons through mutual divisive inhibition (shunting inhibition, compare [Carandini, Heeger, & Movshon, 1999]). These mechanisms are summarized in a three-stage hierarchy of processing that includes input filtering, modulatory feedback, and pool normalization. The output of a cell is defined by a signal function which converts the membrane potential into a firing rate, or activity. Such model cells are grouped into layers which form abstract models of cortical areas which are iconized as boxes in Fig. 2.

The computational properties of the model neurons and the three-stage cascade have been previously reported in (e.g., BayerlBayerl & Neumann, 2004; Bouecke, Tlapale, Kornprobst, & Neumann, 2011; Hansen & Neumann, 2008; Raudies, Mingolla, & Neumann, 2011). We will not go into further detail to explain the computational mechanisms

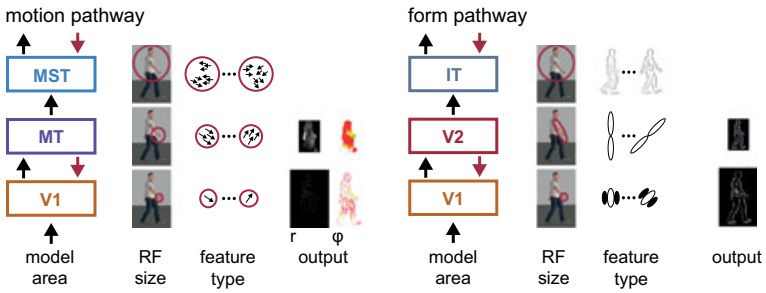


Fig. 2. Structure of hierarchical feedforward and feedback processing along the dorsal (left) and ventral (right) pathway. Each pathway is organized in a homologous fashion, utilizing receptive fields of increasing size over the successive model areas. Feedback between the different areas stabilizes the feature processing from raw input. Representations of prototypical form and motion patterns in IT and MST, respectively, are established by unsupervised Hebbian learning. Processing results are shown in the output columns on the right. Partially modified after Layher et al., (2012).

here. The structure of the processing hierarchy along with characteristic simulation results are shown in Fig. 2.

The representations that were learned in the segregated form and motion pathways define the input activities for model STS cells. On the basis of Jellema, Maassen, and Perrett (2004), we suggest that form and motion activities converge to drive cells in categorial representations of temporal movement selectivity.

2.2. *Learning of form and motion prototypes*

First, we investigated how intermediate-level feature representations can be learned in a biologically plausible fashion by exposing the network architecture with realistic input sequences. To generate feature representations of complex form and motion patterns, we employ an unsupervised learning mechanism based on a modified Hebbian learning scheme. The modification stabilizes the learning such that the growing of weight efficacies is constrained to approach (bounded) activity levels of the input or the output activation. Motivated by the invariance properties observed by Wallis and Rolls (1997), we combined the modified Hebbian learning mechanism with a short-term memory trace of prolonged activity of the pre- or the post-synaptic cells (*trace rule*). The adaptation of weightings is controlled by post-synaptic cells, which, in turn, mutually compete for their ability to adjust their incoming connection weights. The particular details as well as the particular variations of the core architecture are explained below.

2.2.1. *Hebbian learning in the form and motion pathways*

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, respectively. 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; $s \in \{\text{form, motion}\}$ indicates that the same core mechanisms are devoted to the 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, henceforth denoted as activity. The activity $\bar{v}_i$ of the post-synaptic cell is calculated by the temporal trace rule $\bar{v}_i^t = (1-\lambda)\bar{v}_i^{t-1} + \lambda v_i^t$, $0 < \lambda < 1$ (Földiák, 1991). The trace rule (see also Wallis & Rolls, 1997; Rolls & Milward, 2000) has been proposed to incorporate a short-term memory function for the cells to keep their activation over a short temporal window while adapting their weights. The term in brackets on the r.h.s. of learning equation Eq. 1 serves as a biologically plausible mechanism to bound the growth of the cells' input synaptic weights (Oja, 1982). The post-synaptic cells (with activity $\bar{v}_i^{post}$) which gate the learning of their respective input weights are arranged in a competitive layer of neurons competing for the best matching response and their subsequent ability to adapt their kernel of spatial input weights. In a nutshell, the layer of post-synaptic neurons competes to select a winning node for a given input presentation which, in turn, is allowed to automatically adapt their incoming (instar) synaptic weights. The temporal trace (or short-term memory) results in an averaging of input activities over a short temporal interval, thus allowing small pertubations within the input signals.

2.2.2. Reinforcing snapshot learning

The Giese–Poggio model (Giese & Poggio, 2003) 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 suggest a mechanism of how snapshots corresponding to strongly articulated poses can be selected automatically. 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

$$m_e = \int_{\Omega} u_{\phi}(\mathbf{x}) \cdot \Lambda(\mathbf{x}) d\mathbf{x} d\phi \tag{2}$$

with $\Lambda(\bullet)$ denoting a spatial kernel for weighting the relative contribution of motion responses $u_{\phi}(\bullet)$ at spatial locations $\boldsymbol{x}$ in the two-dimensional image plane and with direction selectivity ϕ.[1] The motion energy signal itself is a function of time which is used to steer the instar learning in the form pathway. We suggest that different subpopulations of static form representations, or snapshot, 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 $g(m_e)$ to control a vigilance in snapshot learning to favor form inputs which co-occur with local motion energy minima, that is, when $\partial_t m_e = 0$, given that $\partial_{tt} m_e > 0$. In the weight adaptation, $\Delta w_{ji}^{FF,form}$ in Eq. 1, the learning rate is now gated by the motion-dependent reinforcement, $\eta_{form} \cdot g(m_e)$, which leads to the revised learning rule

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

2.3. 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, feedforward their activations to the stage of STS. 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.

2.3.1. Learning of feedforward connections

Prototypical representations with spatiotemporal sequence selectivity are suggested to exist in the cortical STS complex where both form and motion pathways converge. The selectivities of model STS neurons are learned by again using a modified Hebbian instar learning mechanism similar to the separate learning of form and motion prototypes (Eq. 1),

$$\Delta w_{ji}^{in,FF} = \eta_{seqFF} \cdot \bar{v}_i^{post} \cdot (u_j^{pre} - \bar{v}_i^{post} \cdot w_{ji}^{in,FF}). \tag{4}$$

The weighting kernel $w_{ji}^{in,FF}$ represents convergent IT $\rightarrow$ STS and MST $\rightarrow$ STS bottom-up input to a post-synaptic 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 for the receiving model STS cells are generated by concatenating form and motion output activations, namely $\boldsymbol{u} = \boldsymbol{u}^{IT} \cup \boldsymbol{u}^{MST}$.

2.3.2. Learning feedback connections

An important component is that sequence-selective prototypes in STS in turn learn the output weights back to the segregated form and motion prototype representations, namely STS $\rightarrow$ IT + MST. Unlike the FF learning mechanisms, the learning here is gated by the pre-synaptic cell (in STS) for their top-down weights, which reads

$$\Delta w_{ji}^{out,FB} = \eta_{seqFB} \cdot \bar{v}_i^{pre} \cdot (u_j^{post} - w_{ji}^{out,FB}) \tag{5}$$

with the same components as in the bottom-up learning formalism in Eq. 4. Bottom-up and top-down learning schemes slightly differ in the definition of the competitive terms

(in brackets). In the feedback learning, we employ a difference term between post-synaptic activity and the weighting, $u_j^{post} - w_{ji}^{out,FB}$, omitting the additional weighting of the connectivity strength via the pre-synaptic activity, as in the Oja rule. In steady state, each of the connection strengths emanating from STS cells assumes a value corresponding to the post-synaptic activity distribution, which defines the current input activation. Given an STS cell with attraction $\bar{v}_i^{pre}$, the top-down weight vector approaches $\mathbf{u}^{post} = \mathbf{w}_i^{out,FB}$, thus learning the expected average input. Combined with the temporal trace, this establishes a representation in which each STS sequence-selective prototype encodes and memorizes in its weight pattern the expected driving input activity pattern configuration from the form and motion pathway. Such a top-down weighting pattern can then be used to generate predictions concerning the expected future input given the current maximally activated prototype at the STS level.

3. Results

The model has been tested in various computational experiments. 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. 3 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 ($g(m_e) = 1$,) and when it is enabled ($g(m_e)$ is a monotonically decreasing function of m_e.

We studied the response properties of STS representations and their motion sequence selectivity. There, a prototypical sequence-selective representation is learned for a walker that is traversing from left to right. After training of form, motion, and sequence representations, the network is probed by three different movement scenarios: a forward

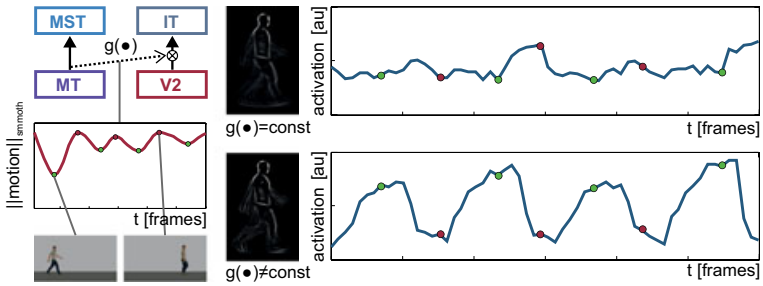


Fig. 3. IT prototypes trained using disabled and enabled reinforcement signal. Minima and maxima in motion energy correspond to articulated and non-articulated postures (bottom left). Continuous learning of IT prototypes leads to activation profiles with low selectivity (top right). Motion-driven reinforcement leads to IT prototypes which signal snapshot poses in synchrony with the gait (bottom right; for details, see text). Partially modified after (Layher et al., 2012).

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, whereas 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, whereas the motion responses remain minimal. As a consequence, the sequence-selective representations respond at an intermediate level (Fig. 4). This evidence is consistent with the experimental findings by Oram and Perrett, (1996) and recent observations by Singer and Sheinberg, (2010).

We further investigated the direction 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°}. Model simulations result in a direction tuning of STS cells with half amplitude of approximately $\pm$ 40 deg (Fig. 5). IT and MST cells, on the other hand, also show a drop in response but have a much larger variability.

In an additional experiment, we selectively lesioned of the model architecture, particularly investigating the effects of extinguishing connections between model areas and the activity flow between learned representations (Fig. 6). 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) 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 Baker, Keysers, Jellema, Wicker, & Perrett, 2001). 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 pre-synaptic activation. We argue that this reflects the induction of increased fMRI

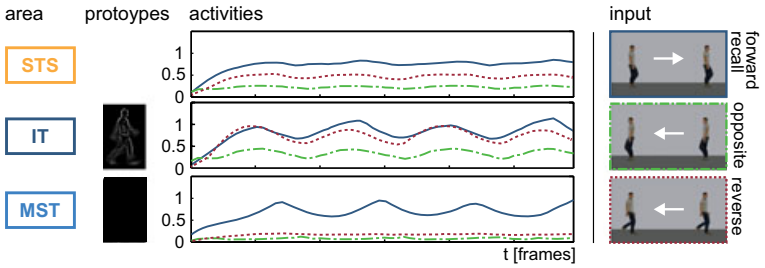


Fig. 4. Response behavior of IT snapshot neurons, MST motion pattern neurons, and sequence-selective STS cells trained by video input for a walker moving from left to right. Activations in the model areas are shown for different input conditions for recall of the training sequence (top), opposite walker movement (middle), and walker displayed in reverse motion (bottom). Line styles and colors encode the input test cases on the right. For details and brief discussion, see text. Partially modified after (Layher et al., 2012).

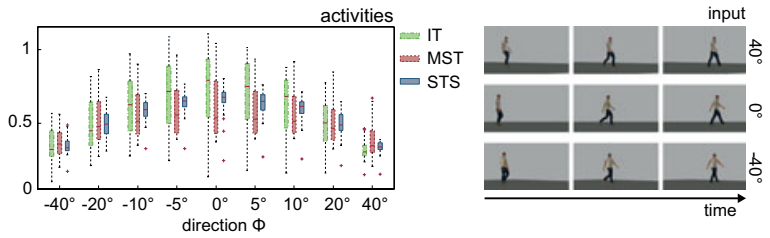


Fig. 5. Response tunings of models cells in area IT (snapshots), MST (motion patterns), and STS (sequence-selective patterns) after training. Category representations have been learned for a walker moving along horizontal direction for $\phi = 0^{\circ}$. Activities of prototypical cells are shown (bottom) which were probed by different inputs with varying movement directions, that is, walkers approaching or receding at different angles with respect to the horizontal reference axis (top). Data have been summarized into box plots showing the response variabilities of models cells as well as the monotonic decline in response for deviations from the target view tuning. The tuning width at half maximum response is around $\pm$ 40. The variance of the MST/IT prototypes decreases toward larger deviations, depicting the loss of response selectivity of prototypes to different parts of a walker's gait.

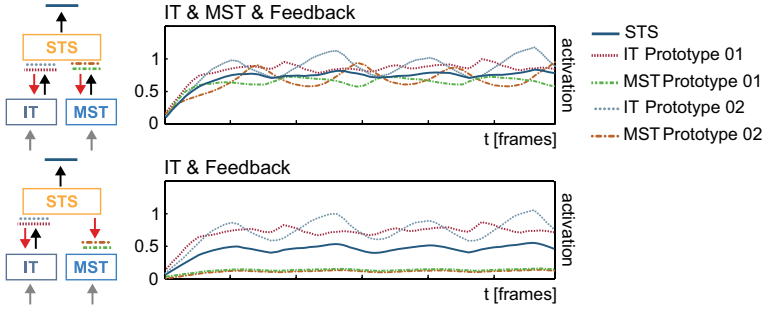


Fig. 6. Selective removal of interconnections (lesioning). The model was trained using the same walking sequence as in the second experiment (see Fig. 4/*forward recall*). The model was left untouched to provide a reference (top). Bottom-up (feedforward) connections between area MST and STS were removed, preventing any motion-related signal being propagated to STS (bottom). The amplitude of the IT prototype activities remains almost the same, whereas the sequence-selective STS cell responds only at about half-magnitude (because of the missing support from the motion pathway). Note the feedback activities propagated from STS to MST optical flow pattern prototypes. We argue that this reflects the induction of increased fMRI BOLD response in human MT + following the presentation of static implied motion stimuli.

BOLD response in human MT + following the presentation of static implied motion stimuli (Kourtzi & Kanwisher, 2000; Senior et al., 2000).

4. Discussion and conclusion

We propose a biologically plausible model for the learning of animated motion sequences. The model builds upon neurophysiological evidence about the cortical sites and specific neuronal representations which contribute to articulated motion and implied motion

perception. There are several new contributions of this article: First, we suggest how prototype representations in the form and motion pathways, namely in model cortical areas IT and MST, can be established by unsupervised learning from sequences containing animated motions. Learning mechanisms are based on modified Hebbian schemes which are stabilized through a trace mechanisms and the incorporation of an objective function taking the weight kernel saturation into account. Second, we suggest that sequence-selective cells in model area STS are learned by using the same learning mechanisms, but now by combining the responses of intermediate-level representations in the form and motion pathways. Third, the learning of articulated poses (snapshots) is controlled by a reinforcement mechanism that enables Hebbian learning in the form pathway through cross-pathway motion-form interaction. Given an animated motion sequence, snapshots are automatically selected as key poses corresponding to strong body pose articulations. Finally, the sequence-selective cells in model STS project to their respective input representations in the form and motion pathways. These feedback connections are again learned by a Hebbian mechanism. Together, the feedforward and the feedback interactions establish a loop of recurrent processing that stabilizes the patterns of form, motion, and sequence representation. Via feedback, model STS cells generate a predictive signal through the backward connections' weights to encode the expected matching input that is suitable to match the currently activated sequence pattern. Together with the newly proposed feedback mechanism, the model is able to account for various experimental findings, in particular, the ability to infer and predict future motion sequence development from articulated postures (implied motion). Importantly, cells in STS are responsive to *both* motion as well as static form (Oram & Perrett, 1996). Experimental findings demonstrate that static images with an increased degree of articulation evoke activation in a population of STS cells. These cells show an increased likelihood of responding to animated walking sequences with matching walking direction as implicated by the articulated pose. A different subpopulation of cells responsive to non-articulated poses, on the other hand, are more likely to respond to sequences with backward-moving walkers (Barraclough, Xiao, Oram, & Perrett, 2006). The model also hypothesizes how the presentation of static articulated poses leads to the emergence of predictive motion perception and enhanced neural activations in the motion pathway (Kourtzi & Kanwisher, 2000; Senior et al., 2000). In accordance to Barraclough, Keith, Xiao, Oram, and Perrett (2009), learned sequence-selective prototype representations in model area STS show a viewpoint-dependent response behavior with direction tunings in response to walkers in the range ± 40 (similar to those reported in Perrett et al., 1989). Once again, the model makes a testable prediction that articulated poses represent the snapshot frames that have been suggested by Giese and Poggio (2003) and that have recently been tested experimentally by Singer and Sheinberg (2010) and Vangeneugden, Pollick, and Vogels (2009)

Acknowledgments

Georg Layher and Heiko Neumann have been supported by the Transregional Collaborative Research Centre SFB/TRR 62 "A Companion Technology for Cognitive

Technical Systems" funded by the German Research Foundation (DFG). A preliminary version of this article has been presented at ICANN 2012 (Layher, Giese, & Neumann, 2012).

Note

1. 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.

References

Baker, C., Keysers, C., Jellema, T., Wicker, B., & Perrett, D. (2001). Neuronal representation of disappearing and hidden objects in temporal cortex of the macaque. *Experimental Brain Research*, *140*(3), 375–381.

Barraclough, N. E., Keith, R. H., Xiao, D., Oram, M. W., & Perrett, D. I. (2009). Visual adaptation to goal-directed hand actions. *Journal of Cognitive Neuroscience*, *21*(9), 1805–1819.

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

Bayerl, P., & Neumann, H. (2004). Disambiguating visual motion through contextual feedback modulation. *Neural Computation*, *16*(10), 2041–2066.

Bouecke, J. D., Tlapale, E., Kornprobst, P., & Neumann, H. (2011). Neural mechanisms of motion detection, integration, and segregation: from biology to artificial image processing systems. *Eurasip Journal on Advances in Signal Processing*, *2011*(1), 781561.

Carandini, M., Heeger, D., & Movshon, J. (1999). Linearity and gain control in V1 simple cells. *Cerebral Cortex*, *13*, 401–444.

Escobar, M., & Kornprobst, P. (2012). Action recognition via bio-inspired features: The richness of center-surround interaction. *Computer Vision and Image Understanding*, *116*(5), 593–605.

Escobar, M., Masson, G., Vieville, T., & Kornprobst, P. (2009). Action recognition using a bio-inspired feedforward spiking network. *International Journal of Computer Vision*, *82*(3), 284–301.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, *3*(2), 194–200.

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

Hansen, T., & Neumann, H. (2008). A recurrent model of contour integration in primary visual cortex. *Journal of Vision*, *8*(8), 1–25.

Jellema, T., Maassen, G., & Perrett, D. (2004). Single cell integration of animate form, motion and location in the superior temporal cortex of the macaque monkey. *Cerebral Cortex*, *14*(7), 781–790.

Jellema, T., & Perrett, D. (2003). Cells in monkey STS responsive to articulated body motions and consequent static posture: a case of implied motion? *Neuropsychologia*, *41*(13), 1728–1737.

Jhuang, H., Serre, T., Wolf, L., & Poggio, T. (2007). A biologically inspired system for action recognition. In (Ed.), *Proceedings of the 11th ieee international conference on computer vision* (pp. 1–8).

Kourtzi, Z., & Kanwisher, N. (2000). Activation in human MT/MST by static images with implied motion. *Journal of Cognitive Neuroscience*, *12*(1), 48–55.

Layher, G., Giese, M. A., & Neumann, H. (2012). Learning representations for animated motion sequence and implied motion recognition. In A. E. P. Villa, W. Duch, P. Érdi, F. Masulli, and G. Palm (Ed.),

Artificial neural networks and machine learning - ICANN 2012 (pp. 288–295). LNCS 7552. Berlin: Springer.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, *15*(3), 267–273.

Oram, M., & Perrett, D. (1996). Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey. *Journal of Neurophysiology*, *76*(1), 109–129.

Perrett, D., Harries, M., Bevan, R., Thomas, S., Benson, P., Mistlin, A., &Ortega, J. (1989). Frameworks of analysis for the neural representation of animate objects and actions. *Journal of Experimental Biology*, *146*(1), 87–113.

Raudies, F., Mingolla, E., & Neumann, H. (2011). A model of motion transparency processing with local center-surround interactions and feedback. *Neural Computation*, *23*(11), 2868–2914.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*, 1019–1025.

Rolls, E. T., & Milward, T. T. (2000). A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Computation*, *12*(11), 2547–2572.

Schindler, K., & Van Gool, L. (2008). Action snippets: How many frames does human action recognition require?. In (Ed.), *Proceedings of the 2008 ieee conference on computer vision and pattern recognition* (pp. 1–8). IEEE Computer Society.

Senior, C., Barnes, J., Giampietroc, V., Simmons, A., Bullmore, E., Brammer, M., & David, A. (2000). The functional neuroanatomy of implicit-motion perception or representational momentum. *Current Biology*, *10*(1), 16–22.

Singer, J., & Sheinberg, D. (2010). Temporal cortex neurons encode articulated actions as slow sequences of integrated poses. *Journal of Neuroscience*, *30*(8), 3133–3145.

Vangeneugden, J., Pollick, F., & Vogels, R. (2009). Functional differentiation of macaque visual temporal cortical neurons using a parametric action space. *Cerebral Cortex*, *19*(3), 593–611.

Wallis, G., & Rolls, E. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, *51*(2), 167–194.

Weidenbacher, U., & Neumann, H. (2009). Extraction of surface-related features in a recurrent model of V1-V2 interactions. *PLoS ONE*, *4*(6), e5909.

[Layher et al., 2017a]

G. Layher, T. Brosch, and H. Neumann (2017a). “Real-time Biologically Inspired Action Recognition from Key Poses using a Neuromorphic Architecture”. In: *Frontiers in Neurorobotics* 11.13, p. 19. ISSN: 1662-5218. DOI: 10.3389/fnbot.2017.00013

©2017 G. Layher, T. Brosch and H. Neumann.
This is an open-access article distributed under the terms of the Creative Commons Attribution License CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 22 March 2017
doi: 10.3389/fnbot.2017.00013

Real-Time Biologically Inspired Action Recognition from Key Poses Using a Neuromorphic Architecture

Georg Layher, Tobias Brosch and Heiko Neumann*

Institute of Neural Information Processing, Ulm University, Ulm, Germany

Intelligent agents, such as robots, have to serve a multitude of autonomous functions. Examples are, e.g., collision avoidance, navigation and route planning, active sensing of its environment, or the interaction and non-verbal communication with people in the extended reach space. Here, we focus on the recognition of the action of a human agent based on a biologically inspired visual architecture of analyzing articulated movements. The proposed processing architecture builds upon coarsely segregated streams of sensory processing along different pathways which separately process form and motion information (Layher et al., 2014). Action recognition is performed in an event-based scheme by identifying representations of characteristic pose configurations (key poses) in an image sequence. In line with perceptual studies, key poses are selected unsupervised utilizing a feature-driven criterion which combines extrema in the motion energy with the horizontal and the vertical extendedness of a body shape. Per class representations of key pose frames are learned using a deep convolutional neural network consisting of 15 convolutional layers. The network is trained using the *energy-efficient deep neuromorphic networks* (*Eedn*) framework (Esser et al., 2016), which realizes the mapping of the trained synaptic weights onto the *IBM Neurosynaptic System* platform (Merolla et al., 2014). After the mapping, the trained network achieves real-time capabilities for processing input streams and classify input images at about 1,000 frames per second while the computational stages only consume about 70 mW of energy (without spike transduction). Particularly regarding mobile robotic systems, a low energy profile might be crucial in a variety of application scenarios. Cross-validation results are reported for two different datasets and compared to state-of-the-art action recognition approaches. The results demonstrate, that (I) the presented approach is on par with other key pose based methods described in the literature, which select key pose frames by optimizing classification accuracy, (II) compared to the training on the full set of frames, representations trained on key pose frames result in a higher confidence in class assignments, and (III) key pose representations show promising generalization capabilities in a cross-dataset evaluation.

Keywords: action recognition, key pose selection, deep learning, neuromorphic architecture, IBM neurosynaptic system

OPEN ACCESS

Edited by:
Poramate Manoonpong, University of Southern Denmark Odense, Denmark

Reviewed by:
Johannes Bill, Heidelberg University, Germany
Quansheng Ren, Peking University, China

****Correspondence:***
Georg Layher
georg.layher@uni-ulm.de

Received: *25 November 2016*
Accepted: *21 February 2017*
Published: *22 March 2017*

Citation:
Layher G, Brosch T and Neumann H (2017) Real-Time Biologically Inspired Action Recognition from Key Poses Using a Neuromorphic Architecture. Front. Neurorobot. 11:13. doi: 10.3389/fnbot.2017.00013

1. INTRODUCTION

Analyzing and understanding the actions of humans is one of the major challenges for future technical systems aiming at visual sensory behavior analysis. Acquiring knowledge about what a person is doing is of importance and sometimes even crucial in a variety of scenarios. In the context of automated surveillance systems, action analysis is an essential ability, allowing to identify potential threads emanating from an individual or a group of persons. In Human-Computer-Interaction (HCI), action analysis helps in understanding the objectives and intentions of a user and increases the potential of a system to adapt to the specific context of an interaction and appropriately support, guide or protect the user. Moreover, recognizing actions in the surrounding area is an integral part of interpreting the own situative context and environment, and thus is in particular crucial for mobile robotic systems which may find themselves embedded in a variety of different situations.

In the presented work, as the first main contribution, a *feature-driven* key pose selection method is proposed, which is driven by combining two features in the biological motion input, namely extrema in the temporal motion energy signal and the relative extent of a subject's pose. Such temporally defined features (from the motion stream) help to automatically select key pose representations. The use of these dynamic features has been motivated by psychophysical investigations (Thirkettle et al., 2009) which demonstrate that humans select specific poses in a continuous sequence of video input based on such criteria. We first show how such key poses define events within articulated motion sequences and how these can be reliably and automatically detected. The proposed processing architecture builds upon coarsely segregated streams of sensory processing along different pathways which separately process form and motion information (Giese and Poggio, 2003). An interaction between the two processing streams enables an automatic selection of characteristic poses during learning (Layher et al., 2014). To use such recognition functionality in an autonomous neurobiologically inspired recognition system various constraints need to be satisfied. Such neurobiological systems need to implement the underlying processes along the processing and recognition cascade which defines the parts of their cognitive functionality.

As the second key contribution, we employ here an *energy efficient deep convolutional neural network* (Eedn; Esser et al., 2016) to realize the key pose learning and classification, which achieves a computationally efficient solution using a sparse and energy efficient implementation based on neuromorphic hardware. This allows us to establish a cascaded hierarchy of representations with an increasing complexity for key pose form and motion patterns. After their establishment, key pose representations allow an assignment of a given input image to a specific action category. We use an offline training scheme that utilizes a deep convolutional neural network with 15 convolutional layers. The trained network runs on IBM's TrueNorth chip (Merolla et al., 2014; Akopyan et al., 2015). This solution renders it possible to approach faster than real-time capabilities for processing input streams and classify articulated still images at about 1,000 frames per second while the computational stages consume only about 70 mW of energy. We present cross-validation results on an action recognition dataset consisting of 14 actions and 22 subjects and about 29,000 key pose frames, which show a recall rate for the presented approach of about 88%, as well as a comparison to state-of-the-art action recognition approaches on a second dataset. To show the generalization capabilities of the proposed key pose based approach, we additionally present the results of a cross-dataset evaluation, where the training and the testing of the network was performed on two completely separate datasets with overlapping classes.

2. RELATED WORK

The proposed key pose based action recognition approach is motivated and inspired by recent evidences about the learning mechanisms and representations involved in the processing of articulated motion sequences, as well as hardware and software developments from various fields of visual sciences. For instance, empirical studies indicate, that special kinds of events within a motion sequence facilitate the recognition of an action. Additional evidences from psychophysics, as well as neurophysiology suggest that both, form and motion information contribute to the representation of an action. Modeling efforts propose functional mechanisms for the processing of biological motion and show how such processing principles can be transfered to technical domains. Deep convolutional networks make it possible to learn hierarchical object representations, which show an impressive recognition performance and enable the implementation of fast and energy efficient classification architectures, particularly in combination with neuromorphic hardware platforms. In the following sections, we will briefly introduce related work and results from different scientific fields, all contributing to a better understanding of action representation and the development of efficient action recognition approaches.

2.1. Articulated and Biological Motion

Starting with the pioneering work of Johansson (1973), perceptual sciences gained more and more insights about how biological motion might be represented in the human brain and what the characteristic properties of an articulated motion sequence are. In psychophysical experiments, humans show a remarkable performance in recognizing biological motions, even when the presented motion is reduced to a set of points moving coherently with body joints (*point light stimuli*; PLS). In a detection task, subjects were capable of recognizing a walking motion within about 200 ms (Johansson, 1976). These stimuli, however, are not free of – at least configurational – form information and the discussion about the contributions of form and motion in biological motion representation is still ongoing (Garcia and Grossman, 2008). Some studies indicate a stronger importance of motion cues (Mather and Murdoch, 1994), others emphasize the role of configurational

form information (Lange and Lappe, 2006). Even less is known about the specific nature and characteristic of the visual cues which facilitate the recognition of a biological motion sequence. In Casile and Giese (2005), a statistical analysis as well as the results of psychophysical experiments indicate that local opponent motion in horizontal direction is one of the critical features for the recognition of PLS. Thurman and Grossman (2008) conclude, that there are specific moments in an action performance which are "more perceptually salient" compared to others. Their results emphasize the importance of dynamic cues in moments when the distance between opposing limbs is the lowest (corresponding to local opponent motion; maxima in the motion energy). On the contrary, more recent findings by Thirkettle et al. (2009) indicate, that moments of a large horizontal body extension (co-occurring with minima in the motion energy) facilitate the recognition of a biological motion in a PLS.

In neurophysiology, functional imaging studies (Grossman et al., 2000), as well as single-cell recordings (Oram and Perrett, 1994) indicate the existence of specialized mechanisms for the processing of biological motion in the *superior temporal sulcus* (STS). STS has been suggested to be a point of convergence of the separate dorsal "where" and the ventral "what" pathways (Boussaoud et al., 1990; Felleman and Van Essen, 1991), containing cells which integrate form and motion information of biological objects (Oram and Perrett, 1996) and selectively respond to, e.g., object manipulation, face, limb and whole body motion (Puce and Perrett, 2003). Besides the evidence that both form and motion information contribute to the registration of biological motion, action specific cells in STS are reported to respond to static images of articulated bodies which in parallel evoke activities in the *medio temporal* (MT) and *medial superior temporal* (MST) areas of the dorsal stream (*implied motion*), although there is no motion present in the input signal (Kourtzi and Kanwisher, 2000; Jellema and Perrett, 2003). In line with the psychophysical studies, these results indicate that poses with a specific feature characteristic (here, articulation) facilitate the recognition of a human motion sequence.

Complementary modeling efforts in the field of computational neuroscience suggest potential mechanisms which might explain the underlying neural processing and learning principles. In Giese and Poggio (2003) a model for the recognition of biological movements is proposed, which processes visual input along two separate form and motion pathways and temporally integrates the responses of prototypical motion and form patterns (snapshots) cells via asymmetric connections in both pathways. Layher et al. (2014) extended this model by incorporating an interaction between the two pathways, realizing the automatic and unsupervised learning of key poses by modulating the learning of the form prototypes using a motion energy based signal derived in the motion pathway. In addition, a feedback mechanism is proposed in this extended model architecture which (I) realizes sequence selectivity by temporal association learning and (II) gives a potential explanation for the activities in MT/MST observed for static images of articulated poses in neurophysiological studies.

2.2. Action Recognition in Image Sequences

In computer vision, the term vision-based action recognition summarizes approaches to assign an action label to each frame or a collection of frames of an image sequence. Over the last decades, numerous vision-based action recognition approaches have been developed and different taxonomies have been proposed to classify them by different aspects of their processing principles. In Poppe (2010), action recognition methods are separated by the nature of the image representation they rely on, as well as the kind of the employed classification scheme. Image representations are divided into *global representations*, which use a holistic representation of the body in the *region of interest* (ROI; most often the bounding box around a body silhouette in the image space), and *local representations*, which describe image and motion characteristics in a spatial or spatio-temporal local neighborhood. Prominent examples for the use of whole body representations are *motion history images* (MHI) (Bobick and Davis, 2001), or the application of *histograms of oriented gradients* (HOG) (Dalal and Triggs, 2005; Thurau and Hlavác, 2008). Local representations are, e.g., employed in Dollar et al. (2005), where motion and form based descriptors are derived in the local neighborhood (cuboids) of spatio-temporal interest points. Classification approaches are separated into *direct classification*, which disregard temporal relationships (e.g., using histograms of prototype descriptors, Dollar et al., 2005) and *temporal state-space models*, which explicitly model temporal transitions between observations (e.g., by employing *Hidden Markov models* (HMMs) Yamato et al., 1992, or *dynamic time warping* (DTW) Chaaraoui et al., 2013). For further taxonomies and an exhaustive overview of computer vision action recognition approaches we refer to the excellent reviews in Gavrila (1999); Aggarwal and Ryoo (2011); Weinland et al. (2011).

The proposed approach uses motion and form based feature properties to extract key pose frames. The identified key pose frames are used to learn class specific key pose representations using a deep convolutional neural network (DCNN). Classification is either performed framewise or by temporal integration through majority voting. Thus, following the taxonomy of Poppe (2010), the approach can be classified as using *global representations* together with a *direct classification* scheme. Key pose frames are considered as temporal events within an action sequence. This kind of action representation and classification is inherently invariant against variations in (recording and execution) speed. We do not argue that modeling temporal relationships between such events is not necessary in general. The very simple temporal integration scheme was chosen to focus on an analysis of the importance of key poses in the context of action representation and recognition. Because of the relevance to the presented approach, we will briefly compare specifically key pose base action recognition approaches in the following.

2.3. Key Pose Based Action Recognition

Key pose based action recognition approaches differ in their understanding of the concept of key poses. Some take a phenomenological perspective and define key poses as events which possess a specific feature characteristic giving rise to their peculiarity. There is no a priori knowledge available about whether, when and how often such *feature-driven* events occur within an observed action sequence, neither during the establishment of the key pose representations during training, nor while trying to recognize an action sequence. Others regard key pose selection as the result of a statistical analysis, favoring poses which are easy to separate among different classes or maximally capture the characteristics of an action sequence. The majority of approaches rely on such statistical properties and either consider the *intra-* or the *inter-class* distribution of image-based pose descriptors to identify key poses in action sequences.

Intra-Class Based Approaches

Approaches which evaluate *intra-class* properties of the feature distributions regard key poses as the most representative poses of an action and measures of centrality are exploited on agglomerations in pose feature spaces to identify the poses which are most common to an action sequence. In Chaaraoui et al. (2013), a contour based descriptor following (Dedeoğlu et al., 2006) is used. Key poses are selected by repetitive *k*-means clustering of the pose descriptors and evaluating the resulting clusters using a compactness metric. A sequence of nearest neighbor key poses is derived for each test sequence and dynamic time warping (DTW) is applied to account for different temporal scales. The class of the closest matching temporal sequence of key poses from the training set is used as the final recognition result. Based on histograms of oriented gradients (HOG) and histograms of weighted optical flow (HOWOF) descriptors, Cao et al. (2012) adapt a local linear embedding (LLE) strategy to establish a manifold model which reduces descriptor dimensionality, while preserving the local relationship between the descriptors. Key poses are identified by interpreting the data points (i.e., descriptors/poses) on the manifold as an adjacent graph and applying a *PageRank* (Brin and Page, 1998) based procedure to determine the vertices of the graph with the highest centrality, or relevance.

In all, key pose selection based on an *intra-class* analysis of the feature distribution has the advantage of capturing the characteristics of one action in isolation, independent of other classes in a dataset. Thus, key poses are not dataset specific and – in principle – can also be shared among different actions. However, most *intra-class* distribution based approaches build upon measures of centrality (i.e., as a part of cluster algorithms) and thus key poses are dominated by frequent poses of an action. Because they are part of transitions between others, frequent poses tend to occur in different classes and thus do not help in separating them. Infrequent poses, on the other hand, are not captured very well, but are intuitively more likely to be discriminative. The authors' are not aware of an *intra-class* distribution based method which tries to identify key poses based on their infrequency or abnormality (e.g., by evaluating cluster sizes and distances).

Inter-Class Based Approaches

Approaches based on *inter-class* distribution, on the other hand, consider highly discriminative poses as key poses to separate different action appearances. Discriminability is here defined as resulting in either the best classification performance or in maximum dissimilarities between the extracted pose descriptors of different classes. To maximize the classification performance, Weinland and Boyer (2008) propose a method of identifying a vocabulary of highly discriminative pose exemplars. In each iteration of the forward selection of key poses, one exemplar at a time is added to the set of key poses by independently evaluating the classification performance of the currently selected set of poses in union with one of the remaining exemplars in the training set. The pose exemplar, which increases classification performance the most is then added to the final key pose set. The procedure is repeated until a predefined number of key poses is reached. Classification is performed based on a distance metric obtained by either silhouette-to-silhouette or silhouette-to-edge matching. Liu et al. (2013) combine the output of the early stages of an HMAX inspired processing architecture (Riesenhuber and Poggio, 1999) with a center-surround feature map obtained by subtracting several layers of a Gaussian pyramid and a wavelet laplacian pyramid feature map into framewise pose descriptors. The linearized feature descriptors are projected into a low-dimensional subspace derived by principal component analysis (PCA). Key poses are selected by employing an adaptive boosting technique (AdaBoost; Freund and Schapire, 1995) to select the most discriminative feature descriptors (i.e., poses). A test action sequence is matched to the thus reduced number of exemplars per action by applying an adapted local naive Bayes nearest neighbor classification scheme (LNBNN; McCann and Lowe, 2012). Each descriptor of a test sequence is assigned to its k nearest neighbors and a classwise voting is updated by the distance of a descriptor to the respective neighbor weighted by the relative number of classes per descriptor. In Baysal et al. (2010), noise reduced edges of an image are chained into a contour segmented network (CSN) by using orientation and closeness properties and transformed into a 2-adjacent segment descriptor (k-AS; Ferrari et al., 2008). The most characteristic descriptors are determined by identifying k candidate key poses per class using the k-medoids clustering algorithm and selecting the most distinctive ones among the set of all classes using a similarity measure on the 2-AS descriptors. Classification is performed by assigning each frame to the class of the key pose with the highest similarity and sequence-wide majority voting. Cheema et al. (2011) follow the same key pose extraction scheme, but instead of selecting only the most distinctive ones, key pose candidates are weighted by the number of false and correct assignments to an action class. A weighted voting scheme is then used to classify a given test sequence. Thus, although key poses with large weights have an increased influence on the final class assignment, all key poses take part in the classification process. Zhao and Elgammal (2008) use an information theoretic approach to select key frames within action

sequences. They propose to describe the local neighborhood of spatiotemporal interest points using an intensity gradient based descriptor (Dollar et al., 2005). The extracted descriptors are then clustered, resulting in a codebook of prototypical descriptors (visual words). The pose prototypes are used to estimate the discriminatory power of a frame by calculating a measure based on the conditional entropy given the visual words detected in a frame. The frames with the highest discriminatory power are marked as key frames. Chi-square distances of histogram based spatiotemporal representations are used to compare key frames from the test and training datasets and majority voting is used to assign an action class to a test sequence.

For a given pose descriptor and/or classification architecture, *inter-class* based key pose selection methods in principle minimize the recognition error, either for the recognition of the key poses (e.g., Baysal et al., 2010; Liu et al., 2013) or for the action classification (e.g., Weinland and Boyer, 2008). But, on the other hand, key poses obtained by *inter-class* analysis inherently do not cover the most characteristic poses of an action, but the ones which are the most distinctive within a specific set of actions. Applying this class of algorithms on two different sets of actions sharing one common action might result in a different selection of key poses for the same action. Thus, once extracted, key pose representations do not necessarily generalize over different datasets/domains and, in addition, sharing of key poses between different classes is not intended.

Feature-Driven Approaches

Feature-driven key pose selection methods do not rely on the distribution of features or descriptors at all and define a key pose as a pose which co-occurs with a specific characteristic of an image or feature. Commonly employed features, such as extrema in a motion energy based signal, are often correlated with pose properties such as the degree of articulation or the extendedness. Compared to statistical methods, this is a more pose centered perspective, since parameters of the pose itself are used to select a key pose instead of parameters describing the relationship or differences between poses.

Lv and Nevatia (2007) select key poses in sequences of 3D-joint positions by automatically locating extrema of the motion energy within temporal windows. Motion energy in their approach is determined by calculating the sum over the L^2 norm of the motion vectors of the joints between two temporally adjacent timesteps. 3D motion capturing data is used to render 2D projections of the key poses from different view angles. Single frames of an action sequence are matched to the silhouettes of the resulting 2D key pose representations using an extension of the Pyramid Match Kernel algorithm (PMK; Grauman and Darrell, 2005). Transitions between key poses are modeled using action graph models. Given an action sequence, the most likely action model is determined using the Viterbi Algorithm. In Gong et al. (2010), a key pose selection mechanism for 3D human action representations is proposed. Per action sequence, feature vectors (three angles for twelve joints) are projected onto the subspace spanned by the first three eigenvectors obtained by PCA. Several instances of an action are synchronized to derive the mean performance (in terms of execution) of an action. Motion energy is then defined by calculating the Euclidean distance between two adjacent poses in the mean performance. The local extrema of the motion energy are used to select the key poses, which after their reconstruction in the original space are used as the vocabulary in a bag of words approach. During recognition, each pose within a sequence is assigned to the key pose with the minimum Euclidean distance resulting in a histogram of key pose occurrences per sequence. These histograms serve as input to a support vector machine (SVM) classifier. In Ogale et al. (2007), candidate key poses are extracted by localizing the extrema of the mean motion magnitude in the estimated optical flow. Redundant poses are sorted out pairwise by considering the ratio between the intersection and the union of two registered silhouettes. The final set of unique key poses is used to construct a probabilistic context-free grammar (PCFG). This method uses an *inter-class* metric to reject preselected key pose candidates and thus is not purely feature-driven.

Feature-driven key pose selection methods are independent of the number of different actions within a dataset. Thus, retraining is not necessary if, e.g., a new action is added to a dataset and the sharing of key poses among different actions is in principle possible. Naturally, there is no guarantee, that the selected poses maximize the separability of pose or action classes.

3. MODEL/METHODS

To realize an energy efficient implementation for key pose based action recognition, the proposed model uses a neuromorphic deep convolutional neural network (DCNN) to selectively learn representations of key poses which are assigned to different action classes. In the preprocessing phase, optical flow is calculated on the input sequences and key pose frames are selected in an unsupervised manner. Form and motion information is calculated for each key pose frame. The concatenated form and motion information is then used as the input to the DCNN. In the following, detailed information about the image preprocessing, the key pose selection automatism and the structure and functionality of the DCNN are presented. All simulations were carried out using a neuromorphic computing paradigm and mapped to the IBM TrueNorth hardware platform (Merolla et al., 2014).

3.1. Key Pose Selection and Image Preprocessing

During preprocessing, two elementary processing steps are performed. First, the key pose selection is performed by automatically analyzing simple motion and form parameters. Second, the final input to the network is calculated by combining the form and motion representations I^{form} and I^{motion} obtained by simple image-based operations.

Key Pose Selection

The key pose selection process operates upon two parameters, namely (I) local temporal extrema in the motion energy and (II) the extendedness of a subject at a given timestep. Optical flow is calculated using a differential method, as suggested in the *Lucas-Kanade* optical flow estimation algorithm (Lucas and

Kanade, 1981). Given an image sequence $I(\mathbf{x}, t)$, the optical flow $\mathbf{u}(\mathbf{x}, t) = (u(\mathbf{x}, t), v(\mathbf{x}, t))$ at timestep t and position $\mathbf{x} = (x, y)$ is estimated in a local neighborhood $N(\mathbf{x})$ by minimizing

$$\sum_{\mathbf{y}\in N(\mathbf{x})} W(\mathbf{x}-\mathbf{y})^2[I_x(\mathbf{y}, t)u(\mathbf{x}, t) + I_y(\mathbf{x}, t)v(\mathbf{x}, t) + I_t(\mathbf{y}, t)]^2, \quad (1)$$

where $W(\mathbf{x} - \mathbf{y})$ increases the influence of the optical flow constraints within the center of the local neighborhood (for details see Barron et al., 1994). The spatiotemporal derivatives I_x, I_y and I_t are estimated by convolution of the image sequences with the forth-order central difference $[-1, 8, 0, -8, 1]/12$ and it's transpose in the spatial and the first-order backward difference $[-1, 1]$ in the temporal domain. A separable 2D kernel with 1D coefficients of $[1, 4, 6, 4, 1]/16$ is used to realize the weighted integration of the derivatives within a 5×5 spatial neighborhood ($N(\mathbf{x})$)[1]. The use of the *Lucas-Kanade* algorithm is not a hard prerequisite for the proposed approach. Other types of optical flow estimators might be applied as well (e.g., (Brosch and Neumann, 2016), which is capable to be executed on neuromorphic hardware). The overall motion energy E^{flo} is then calculated by integrating the speed of all estimated flow vectors within the vector field.

$$E^{\text{flo}}(t) = \sum_{\mathbf{x}\in I(\mathbf{x},t)} \|\mathbf{u}(\mathbf{x}, t)\|_2 = \sum_{\mathbf{x}\in I(\mathbf{x},t)} \sqrt{u(\mathbf{x}, t)^2 + v(\mathbf{x}, t)^2}, \quad (2)$$

Motion energy is smoothed by convolving the estimated motion energy with a Gaussian kernel, $\tilde{E}^{\text{flo}}(t) = (E^{\text{flo}} * G_\sigma)(t)$. In the performed simulations, $\sigma = 2$ and $\sigma = 4$ were used dependent on the dataset[2]. Potential key pose frames are then marked by identifying the local extrema of the motion energy signal.

$$\mathcal{K}^{\text{flo}} = \{I(t), t \in [1, ..., T] | \text{t is a local extremum of } \tilde{E}^{\text{flo}}(t)\}, \quad (3)$$

The relative horizontal and vertical extent of a given pose at time t is then used to reject local extrema with an extent smaller than a predefined percentual threshold λ, as defined by:

$$\mathcal{K} = \mathcal{K}^{\text{flo}} \cap \mathcal{K}^{\text{ext}}. \quad (4)$$

with

$$\begin{aligned}\mathcal{K}^{\text{ext}} = \{I(t), t \in [1, ..., T] \,|\, &(\text{Ext}^{\text{ver}}(t) > (1+\lambda)\overline{\text{Ext}}^{\text{ver}}) \\ &\vee (\text{Ext}^{\text{ver}}(t) < (1-\lambda)\overline{\text{Ext}}^{\text{ver}}) \\ &\vee (\text{Ext}^{\text{hor}}(t) > (1+\lambda)\overline{\text{Ext}}^{\text{hor}}) \\ &\vee (\text{Ext}^{\text{hor}}(t) < (1-\lambda)\overline{\text{Ext}}^{\text{hor}})\}\end{aligned} \quad (5)$$

In the performed simulations, values of $\lambda = 0.1$ and $\lambda = 0.05$ were used for the two different datasets. The percentual thresholds were determined manually with the aim to compensate for differences in the temporal resolution of the datasets. The horizontal and vertical extent Ext^{hor} and Ext^{ver} are derived framewise by estimating the width and the height of the bounding box enclosing the body shape. The extent of a neutral pose is used as the reference extent $\overline{\text{Ext}}^{\text{hor}}$ and $\overline{\text{Ext}}^{\text{ver}}$, which are derived from the width and height of the bounding box in the first frame of a sequence. Silhouette representations, and thus the bounding boxes of the bodies, are available for both datasets used in the simulations. In constrained recording scenarios, silhouettes can be extracted by background subtraction or using the optical flow fields calculated for the selection of the key pose frames. **Figure 1A** shows the motion energy signal $\tilde{E}^{\text{flo}}$ together with the extent Ext^{hor} and Ext^{ver} and their reference values. A strong correlation between the motion energy and the extent of the pose can be seen. In **Figure 1B**, examples for the horizontal and the vertical extent are displayed for a neutral and a extended posture. While the motion energy allows an identification of temporal anchor points in a motion sequence, the extent helps in selecting the most characteristic ones.

Form and Motion Representations

For each selected key pose frame $I^{\text{key}} \in \mathcal{K}$, a form representation is derived by estimating the spatial derivatives I_x^{key} and I_y^{key} and combining them into one contour representation I^{con} by concatenating the orientation selective maps (see **Figure 2**, second row). The final form representation is then obtained by applying a logarithmic transformation emphasizing low range values and normalizing the response amplitudes, using the transformation:

$$I_{\log}^{\text{con}} = \log(1 + 5|I^{\text{con}}|) \quad (6)$$

$$I^{\text{form}} = \frac{I_{\log}^{\text{con}}}{\max(I_{\log}^{\text{con}})} \quad (7)$$

Likewise, for each key pose frame I^{key}, optical flow is separated into vertical and horizontal components and concatenated (see **Figure 2**, first row). The resulting motion representation I^{flo} is log-transformed and normalized. As for the contrast mapping, the transformation is given through:

$$I_{\log}^{\text{flo}} = \log(1 + 5|I^{\text{flo}}|)) \quad (8)$$

$$I^{\text{motion}} = \frac{I_{\log}^{\text{flo}}}{\max(I_{\log}^{\text{flo}})} \quad (9)$$

The form representations I^{form} and the motion representations I^{motion} are combined to an overall input representation I^{input} (**Figure 2**, last column). I^{input} is then used as an input for the training of the DCNN described in the following section.

3.2. Learning of Class Specific Key Pose Representations

A neuromorphic deep convolutional neural network was used to establish classwise representations of the preselected and suppress wrapping key pose frames using a supervised learning scheme. The network was implemented using the *energy-efficient deep neuromorphic networks* (*Eedn*) framework (Esser

[1]In the presented simulations, the MATLAB® implementation of the *Lucas-Kanade* flow estimation algorithm was used.

[2]The values of σ were chosen manually to take different temporal resolutions into account.

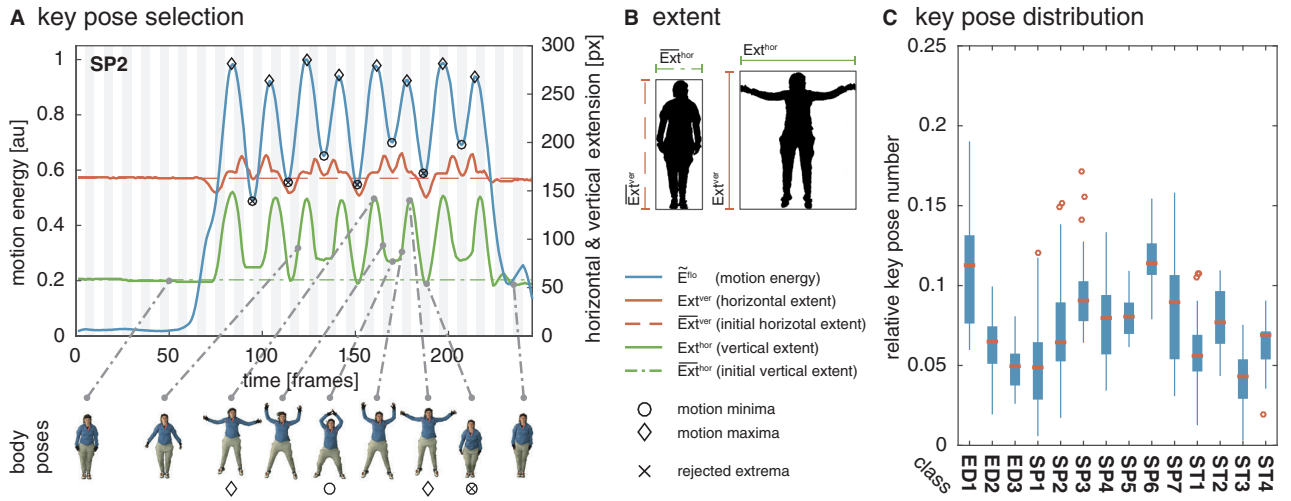


FIGURE 1 | Key pose selection. For each action sequence, key pose frames are selected by identifying minima and maxima in the motion energy which co-occur with a sufficiently increased extent of the body pose. In **(A)**, the smoothed motion energy $\tilde{E}^{\text{flo}}$ (blue) together with the horizontal (red) and vertical (green) extent of the body pose Ext^{hor} and Ext^{ver} and their reference values $\overline{\text{Ext}}^{\text{ver}}$ and $\overline{\text{Ext}}^{\text{hor}}$ (dashed, dash-dotted) are displayed for one of the actions used as part of the simulations (*SP2*/jumping jack). At the bottom, body poses for several frames are shown. Local minima in the motion energy are marked by a circle, local maxima by a diamond. Extrema which are rejected as a key pose frame because of an insufficient extent are additionally marked with ×. **(B)** Shows an example for the horizontal and vertical extent of a neutral and a highly articulated body pose. The first frame of each action sequence is defined to be the neutral pose. **(C)** Shows the relative number of identified key poses per action sequence for the uulmMAD dataset used for the simulations (see Section 4.1). Written informed consent for the publication of the exemplary images was obtained from the displayed subject.

et al., 2016), which adapts and extends the training and network functions of the *MatConvNet* toolbox (Vedaldi and Lenc, 2015). In the following for readers' convenience, we will briefly recapitulate and summarize key aspects of the framework and its extensions presented in Esser et al. (2016). In the framework, the weights established through learning match the representation scheme and processing principles used in neuromorphic computing paradigms. The structure of the DCNN follows one of the network parameter sets presented by Esser et al. (2016), which show a close to state-of-the-art classification performance on a variety of image datasets and allow the trained network to be run on a single IBM TrueNorth chip (Merolla et al., 2014).

A deep convolutional neural network is typically organized in a feedforward cascade of layers composed of artificial neurons (LeCun et al., 2010), which process the output of the proceeding layer (afferent synaptic connections) and propagate the result to the subsequent one (efferent synaptic connections). Following the definition in Esser et al. (2016), an artificial cell j in a DCNN calculates a weighted sum over the input to that cell, as defined by:

$$s_j = \sum_{xy} \sum_{f} in_{xyf} w_{xyfj}, \tag{10}$$

where in_{xyf} are the signals in the input field of cell j at locations (xy) in the spatial and (f) in the feature domain and, w_{xyfj} the respective synaptic weights. In the following, we will use the linear index i to denote locations in the (xyf) space-feature cube. Normalizing the weighted sum over a set of input samples (batch normalization) allows to accelerate the training of the network by standardizing s_j as defined through:

$$\tilde{s}_j = \frac{s_j - \mu_j}{\sigma_j + \epsilon} + b_j, \tag{11}$$

with $\tilde{s}_j$ the standardized weighted sum, μ_j the mean and σ_j the standard deviation of s calculated over the number of training examples within a batch (Ioffe and Szegedy, 2015). b_j is a bias term, allowing to shift the activation function $\phi(\bullet)$, and ϵ guarantees numerical stability. The output activation of the artificial neuron is calculated by applying an activation function on the standardized filter response:

$$r_j = \phi(\tilde{s}_j). \tag{12}$$

Weight adaptation is performed through gradient descent by applying error backpropagation with momentum (Rumelhart et al., 1986). In the forward phase, an input pattern is propagated through the network until the activations of the cells in the output layer are obtained. In the backward phase, the target values of an input pattern are used to calculate the cross entropy C given the current and the desired response of the output layer cell activations, as defined by:

$$C = -\sum_{j=1}^{M} v_j \ln(r_j) = -\sum_{j=1}^{M} v_j \ln(\phi(\tilde{s}_j)), \tag{13}$$

with M denoting the number of cells in the output layer. Here, v_j is the one-hot encoded target value (or teaching signal) of a cell

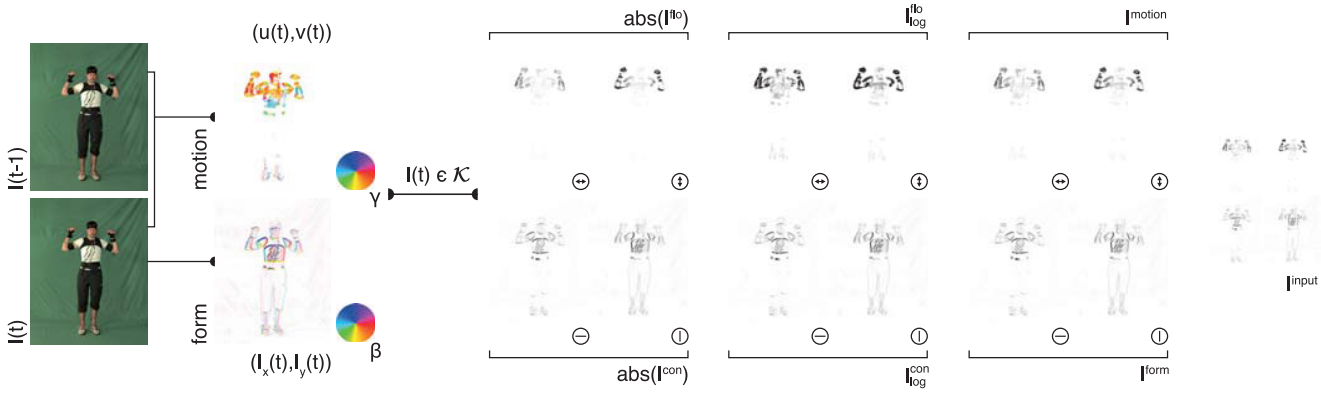


FIGURE 2 | Input generation. Each frame of an action sequence is transformed into a combined motion and form representation. In the top row, the estimated optical flow is displayed in the second column (direction γ color-encoded) for two consecutive frames (first column). The optical flow field is then separated into horizontal and vertical components (third column) and their absolute value is log transformed (forth column) and normalized (fifth column). Form representations (bottom row) are derived framewise by estimating the horizontal and vertical derivatives I_x and I_y (second column, gradient orientation with polarity β color-encoded). The resulting contrast images are then log-transformed and normalized. The form and motion representations are combined into a single feature map I^{input} which is then fed into the convolutional neural network. Image sizes are increased for a better visibility. Written informed consent for the publication of exemplary images was obtained from the shown subject.

j with activation r_j. A *softmax* function is employed as activation function in the output layer, as defined through:

$$\phi(\tilde{s}_j) = \frac{e^{\tilde{s}_j}}{\sum_{k=1}^{M} e^{\tilde{s}_k}}. \tag{14}$$

The cross entropy error $E(t) = C$ is then propagated backwards through the network and the synaptic weight adaptation is calculated for all cells in the output and hidden layers by applying the chain rule. The strength of weight adaptation Δw_{ij} is given through:

$$\Delta w_{ij}(t) = -\eta \frac{\partial E(t)}{\partial w_{ij}} + \alpha \Delta w_{ij}(t-1) = -\eta \delta_j in_i + \alpha \Delta w_{ij}(t-1), \tag{15}$$

$$\text{with } \delta_j = \begin{cases} (r_j - v_j) & \text{if } j \text{ is a neuron in the output layer} \\ \phi'(\tilde{s}_j) \sum_k \delta_k w_{jk} & \text{if } j \text{ is a neuron in a hidden layer,} \end{cases} \tag{16}$$

which includes a momentum term for smoothing instantaneous weight changes. Here, k is the index of cells in the layer succeeding cell j, t describes the current training step, or iteration, and η denotes the learning rate. The momentum factor $0 \leq \alpha \leq 1$ helps the network to handle local minima and flat plateaus on the error surface. After the backward pass, weights are finally adapted by:

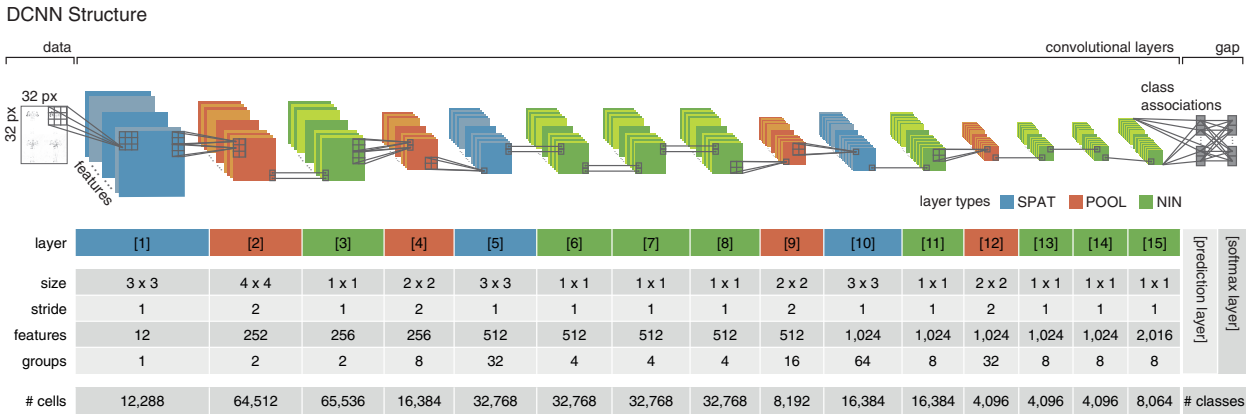
$$w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij}(t). \tag{17}$$

To ensure the compatibility to neuromorphic processing principles, a binary activation function $\phi(\tilde{s}_j)$ is applied in the hidden layers (for details see Section 3.3).

Within a convolutional layer, weights w_{ij} of a cell j are shared over multiple input fields, which are arranged as a regular grid in the source layer. The calculation of the weighted sum during the forward, as well as the integration of the error derivative during the backward pass can be formulated as a convolution with the input from the source, or the error signal from the succeeding layer. The weights w_{ij} act as the filter (or convolution) kernel, $\tilde{s}_j$ as the filter response and r_j as the output of an artificial cell. The *size* and *stride* of a filter allow to adjust the size and the overlap of the input fields to a filter in the source layer. A small stride results in an increased overlap and thus a large number of output values. The number of *features* defines how many different filters are employed in a layer. The weight matrices of the cells within a layer can be separated into *groups* of filters, which define the set of input features from the source layer covered by a filter[3].

It is a common practice to construct deep neural networks by employing convolutional layers for feature extraction in the lower layers and connect them with (one or more) fully connected layers (equivalent to *Multilayer Perceptrons*/MLPs) on top for classification purposes. In contrast, the proposed network follows the strategy of *global average pooling* (gap) proposed in Lin et al. (2013) and applied in Esser et al. (2016). In the final convolutional layer of the network, one feature map is generated for each category of the classification problem. Instead of a full connectivity, the average value of each class-associated feature map is propagated to the output (softmax) layer. Due to their association to classes, the feature maps can directly be interpreted as confidence maps. Following the softmax layer, the cross-entropy error is calculated using one-hot encoded target values v_j and propagated back through the network (according to Equation 16). Networks using parameter-free *global average pooling* layers in combination with softmax are less prone to overfitting (compared to MLPs) and increase the robustness to spatial translations (for details see Lin et al., 2013).

[3] In **Figure 3**, the weight matrices in the convolutional layer 5 have a dimensionality of $3 \times 3 \times 8$, since they receive input from 256 feature maps in layer 4 which are separated into 32 groups of filters, each receiving input from 8 feature maps.



layer	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[prediction layer]	[softmax layer]
size	3 x 3	4 x 4	1 x 1	2 x 2	3 x 3	1 x 1	1 x 1	1 x 1	2 x 2	3 x 3	1 x 1	2 x 2	1 x 1	1 x 1	1 x 1		
stride	1	2	1	2	1	1	1	1	2	1	1	2	1	1	1		
features	12	252	256	256	512	512	512	512	512	1,024	1,024	1,024	1,024	1,024	2,016		
groups	1	2	2	8	32	4	4	4	16	64	8	32	8	8	8		
# cells	12,288	64,512	65,536	16,384	32,768	32,768	32,768	32,768	8,192	16,384	16,384	4,096	4,096	4,096	8,064	# classes	

FIGURE 3 | Deep convolutional neural network structure. The implemented DCNN follows the structure proposed in Esser et al. (2016) and employs three different convolutional layer types (layers 1–15). *Spatial layers* (SPAT; colored in blue) perform a linear filtering operation by convolution. *Pooling layers* (POOL; colored in red) decrease the spatial dimensionality while increasing the invariance and diminishing the chance of overfitting. *Network-in-network* layers (NIN; colored in green) perform a parametric cross channel integration (Lin et al., 2013). The proposed network consists of a data (or input) layer, 15 convolutional layers and a prediction and softmax layer on top. Each of the cells in the last convolutional layer (layer 15) is associated with one class of the classification problem. In the prediction layer, the average class-associated activations are derived (*global average pooling*/gap) and fed into the softmax layer (i.e., one per class), where the cross-entropy error is calculated and propagated backwards through the network. The parameters used for the convolutional layers of the network are given in the central rows of the table. In the last row, the number of artificial cells per layer is listed. The cell count in the *prediction* and *softmax* layer depends on the number of categories of the classification task (i.e., the number of actions in the dataset).

The employed network consists of 15 convolutional layers, which implement three different types of convolutional operations. *Spatial* layers (SPAT) perform a standard convolution operation, *pooling* layers (POOL) reduce the spatial dimensions by applying a convolution with a large stride (Springenberg et al., 2014), *network-in-network* layers (NIN) are realized by convolutional layers with a size of $1x1$ and a stride of 1 and act as cross channel integration layers (Lin et al., 2013). The network structure is summarized in **Figure 3**. Each of the cells in the last convolutional layer (layer 15) is assigned to one class. During learning, activities in this layer are averaged per feature map and fed into the *softmax* layer. For recognition, the average output of the cell populations associated to the individual classes are used as prediction values and serve as the final output r_c^{class} of the network (*prediction* layer in **Figure 3**).

3.3. Neuromorphic Implementation

Processing actual spikes in hardware, the execution of a DCNN on a neuromorphic platform poses several constraints on the activity and weight representation schemes. Since the processing architecture of the TrueNorth neuromorphic platform is based on event-based representations, the gradual activations need to be mapped onto a spike-based mechanism. To be in conformity with these processing principles, Esser et al. (2016) employ a binary activation function, as defined by:

$$\phi(\tilde{s}_j) = \begin{cases} 1 \text{ if } \tilde{s}_j \geq 0 \\ 0 \text{ otherwise,} \end{cases} \tag{18}$$

and ternary synaptic weights ($w_{xyf} \in \{-1, 0, 1\}$). For the backpropagation of the error signal, the derivative of the binary activation is approximated linearly in the range of $[0, 1]$, as given through:

$$\frac{\partial \phi(\tilde{s}_j)}{\partial \tilde{s}_j} \approx \max(0, 1 - |\tilde{s}_j|). \tag{19}$$

During training, a copy of the model weights is held in a *shadow network*, which allows gradual weight adaptation. Weight updates are performed on values in the shadow network using high precision values. For the forward and backward pass, the hidden weights w_{ij}^h in the shadow network are clipped to $[-1, 1]$ and mapped to the ternary values using rounding and hysteresis, following:

$$w_{ij}(t) = \begin{cases} -1 & \text{if } w_{ij}^h(t) \leq -0.5 - h \\ 0 & \text{if } w_{ij}^h(t) \geq -0.5 + h \wedge w_{ij}^h(t) \leq 0.5 - h \\ 1 & \text{if } w_{ij}^h(t) \geq 0.5 + h \\ w_{ij}(t-1) & \text{otherwise} \end{cases} \tag{20}$$

(for details refer to Esser et al., 2016). The hidden weights w_{ij}^h allow the synaptic connection strengths to switch between the ternary values based on small changes in the error gradients obtained during backpropagation, while the hysteresis factor h prevents them from oscillating. The parameters for the training of the network were chosen according to Esser et al. (2016), using a momentum factor of $\alpha = 0.9$ and a learning rate of $\eta = 20$ (reduced by a factor of 0.1 after 2/3 and 5/6 of the total training iterations). The hysteresis factor h was set to 0.1. The mapping of the training network on the TrueNorth platform was performed by the *Eedn* framework. Training was carried out on

Nvidia GPUs, testing was performed on the IBM TrueNorth NS1e board.

The IBM TrueNorth chip consists of 4,096 interconnected neurosynaptic cores with 1 million spiking neurons and 256 million configurable synaptic connections. For the execution of the network on the TrueNorth chip, the trained network parameters are mapped to hardware using an abstraction of a TrueNorth program called *Corelet* (Amir et al., 2013). The platform independent *Corelets* translate the network parameters into a TrueNorth specific configuration, which can be used to program the parameters of the neurons and synaptic connection strengths on the chip. For details on *Corelets* and the mapping of the DCNN on neuromorphic hardware platforms refer to Amir et al. (2013); Esser et al. (2016).

3.4. Temporal Integration of Framewise Class Predictions

After the training of the DCNN, classification is either performed framewise by directly selecting the class corresponding to the cell population in layer 15 with the maximum average activation, or by integrating the individual framewise classification results using majority voting in temporal windows or over the full sequence.

For framewise classification, a key pose frame is identified in an input image sequence $I(\mathbf{x}, t)$ and preprocessed as described in Section 3.1. The resulting input map I^{input} is fed into the DCNN and the class label c associated to the cell population in layer 15 with the maximum average output r_c^{class} defines the class prediction for I^{input}. The value of r_c^{class} can directly be interpreted as the confidence in the prediction.

In sliding window based classification, the predicted class labels for key pose frames are collected within temporal windows of size n [frames], which are shifted over the input sequence $I(\mathbf{x}, t)$. The class with the most frequent occurrence of key pose frames determines the class predicted for the window (majority voting). At the moment, we do not use the confidence r_c^{class} of the predictions as weights for the voting. Note that it is not guaranteed, that key pose frames occur in all temporal windows. Windows which do not contain key poses are not used for evaluation.

Full sequence classification follows the same principle as sliding window based classification, but collects all key pose frames within a sequence. Thus, the amount of temporal information integrated in the voting process might differ substantially from sequence to sequence.

4. DATASETS

The proposed action recognition approach was evaluated using two different action datasets. Due to the higher number of subjects and actions, we focused our analysis on the *uulm multiperspective action dataset* (uulmMAD). In addition, we analyzed the performance on the widely used *Weizmann* dataset to allow a comparison to other approaches and to perform a cross-dataset evaluation of overlapping classes. In the following, we will briefly describe the main characteristics of the two datasets.

4.1. uulmMAD

The *uulm multiperspective action dataset*[4] (uumlMAD; Glodek et al., 2014) consists of data from 31 subjects performing actions from the areas of everyday life (ED), sport/fitness (SP) and stretching (ST). Eight of the actions are repeated three times, six actions are performed four times with varying speed. Altogether, each action is performed either 93 or 124 times. Actions were recorded in front of a greenscreen using three synchronized cameras and the body posture was captured in parallel by an inertial motion capturing system worn by the subjects. To decrease the likelihood of similar visual appearances, the motion capture suit was covered by additional clothes whenever possible. **Figure 4** shows the 14 actions together with a characteristic picture, an abbreviation and a short description for each action.

[4] Available via https://www.uni-ulm.de/imagedb.

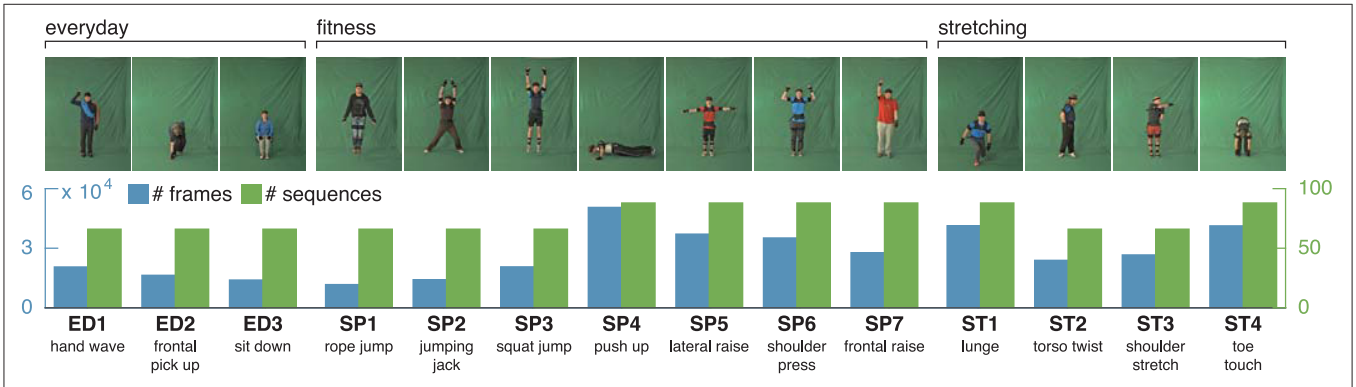


FIGURE 4 | uulmMAD – uulm multiperspective action dataset. The uulmMAD dataset contains 14 actions in the area of everyday activities, fitness/sports and stretching performed by 31 subjects. Per subject, eight of the actions are repeated three times, six actions are performed four times with varying speed. Actions were recorded by three synchronized cameras (frontal, diagonal and lateral) with a frame rate of 30 Hz and an inertial motion capturing system with a sample rate of 120 Hz. Silhouettes were extracted using chromakeying. At the time we carried out the simulations, silhouettes were available for 22 subjects. In the first row exemplary pictures are shown for all actions. The number of videos (green) and total sum of frames (blue) which were available for the evaluation are displayed in the second row. At the bottom, an abbreviation for each action is defined and a short description is given. Written informed consent for the publication of exemplary images was obtained from the displayed subjects.

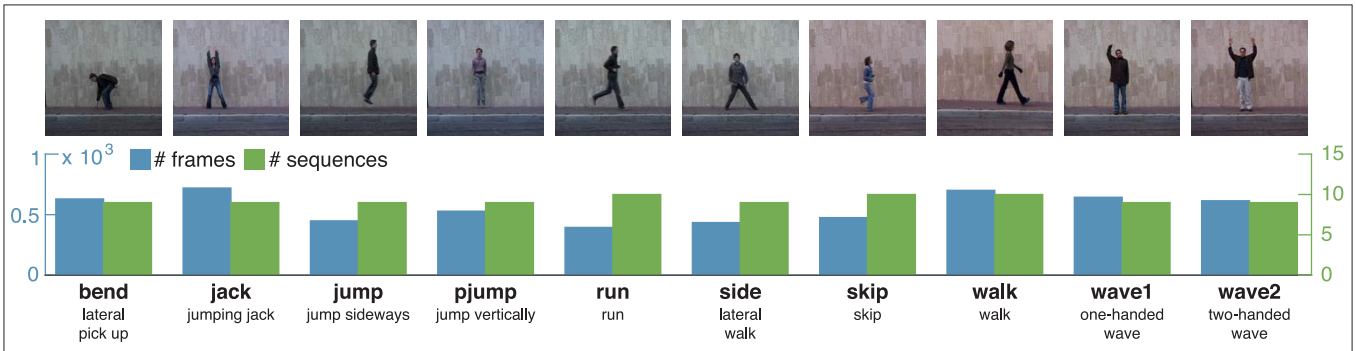


FIGURE 5 | Weizmann action dataset. The Weizmann dataset (Gorelick et al., 2007) is one of the most commonly used action recognition datasets and consist of ten actions, recorded for nine subjects. Actions are performed once (occasionally twice) per subject in front of a static background. Silhouette representations are provided for all sequences. Representative images are displayed alongside with the number of frames and sequences, a label and a short description per class.

At the time we carried out the simulations, silhouette representations were available for all sequences of 22 subjects. Since the silhouettes are used to calculate an estimate of the horizontal and vertical extent of a pose, only the frontal recordings of this subset of subjects were used within the evaluation. Some action pairs (e.g., ED2 and ST4) in the dataset are deliberately intended to appear visually similar and thus be difficult to separate. In total, the sequences used for evaluation contain 381, 194 frames, of which 28, 902 are selected by the key pose selection procedure.

4.2. Weizmann Dataset

To allow a comparison with different action recognition approaches, simulations were additionally carried out using a well established action dataset. The *Weizmann dataset*[5] (see **Figure 5**; Gorelick et al., 2007) consists of ten actions performed by nine subjects. Actions are mostly performed once per subject, although some actions are occasionally performed twice. Actions are captured in 25 *Hz* from a frontoparallel perspective in front of a uniform background.

Silhouettes are available for all subjects and sequences. In total, the sequences contain 5, 594 frames, 1, 873 of which are identified as key pose frames by using the procedure described in Section 3.1.

5. RESULTS

Several simulations were carried out to evaluate the performance of the proposed key pose based action recognition approach. The simulations were intended to address questions related to (I) the overall performance of the approach on different datasets using a framewise, as well as windowed and full sequence majority voting recognition schemes, (II) a comparison to other action recognition methods, (III) a juxtaposition of key pose based and full sequence learning, and (IV) cross-dataset evaluation. Since action recognition datasets—in particular, in case of framewise recognition—are often highly imbalanced, we provide different types of performance measures, as well as classwise performance values for the most essential results. Since the nomenclature and definition of performance measures vary largely in the pattern recognition and machine learning community we will briefly define and describe the reported measures to allow a better comparability. For a comprehensive discussion on performance measures, we refer to Sokolova and Lapalme (2009) and the contributions of D. Powers, e.g. (Powers, 2013).

[5]Available via http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html.

TABLE 1 | Performance measures.

Measure	Abbreviation	Definition
Recall	$\mathrm{Rec_M}$	$\frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{tp}_i}{\mathrm{tp}_i+\mathrm{fn}_i}$
Informedness	$\mathrm{Inf_M}$	$\sum_{i=1}^{N}\frac{\mathrm{tp}_i+\mathrm{fp}_i}{\mathrm{tp}_i+\mathrm{fn}_i+\mathrm{tn}_i+\mathrm{fp}_i}\times(\frac{\mathrm{tp}_i}{\mathrm{tp}_i+\mathrm{fn}_i}+\frac{\mathrm{tn}_i}{\mathrm{fp}_i+\mathrm{tn}_i}-1)$
Markedness	$\mathrm{Mark_M}$	$\sum_{i=1}^{N}\frac{\mathrm{tp}_i+\mathrm{fn}_i}{\mathrm{tp}_i+\mathrm{fn}_i+\mathrm{tn}_i+\mathrm{fp}_i}\times(\frac{\mathrm{tp}_i}{\mathrm{tp}_i+\mathrm{fp}_i}+\frac{\mathrm{tn}_i}{\mathrm{fn}_i+\mathrm{tn}_i}-1)$
Matthews Correlation	$\mathrm{MCC_M}$	$\pm\sqrt{\mathrm{Mark_M}\times\mathrm{Inf_M}}$

In a multiclass classification problem with N classes tp_i (true positives) are commonly defined as the number of correct acceptances (hits) for a class C_i ($i \in [1, ..., N]$), fn_i as the number of false rejections (misses), tn_i as the number of correct rejections of samples of different classes $C_{j\neq i}$ and fn_i (false negatives) as the number of false acceptances (false alarms). Together, these four counts constitute the confusion matrix and allow to derive a variety of measures describing the performance of a trained classification system. The ones used for the evaluation of the presented results are listed alongside with an abbreviation and their definition in **Table 1**.

All multiclass performance measures are calculated using macro averaging (M), since using micro averaging, classes with a large number of examples would dominate the averaging. $\mathrm{Rec_M}$, often referred to as (average) recognition rate or somewhat misleading as (classification) accuracy, might be the performance measurement most frequently used in the action recognition literature and describes the average percentage of correctly identified positive examples per class. $\mathrm{Inf_M}$ reflects how informed the decision of a classifier is in comparison to chance, whereas $\mathrm{Mark_M}$ follows the inverse concept by describing how likely the prediction variable is marked by the true variable

(Powers, 2013). Note, that when calculating the average per class values of Inf_M and Mark_M are weighted by the $\text{Bias}_i = \frac{\text{tp}_i+\text{fp}_i}{\text{tp}_i+\text{fn}_i+\text{tn}_i+\text{fp}_i}$ and the $\text{Prevalence}_i = \frac{\text{tp}_i+\text{fn}_i}{\text{tp}_i+\text{fn}_i+\text{tn}_i+\text{fp}_i}$, respectively. The Matthews Correlation Coefficient MCC_M can be derived by calculating the geometric mean of Inf_M and Mark_M and expresses the correlation between predicted classes and true values.

Leave-one-subject-out cross-validation (LOSO) was performed in all test scenarios and the resulting average performance measures are reported together with the corresponding standard deviations. In the following, rates are either reported in a range of [0, 100] or [0, 1] (due to limited space).

5.1. Classification Performance

The equivalent network structure (see Section 3.2) was used to train the network on the two datasets described in Section 4. In case of the uulmMAD dataset, 28,902 key pose frames (per class average 2,064.43, std 1,097.16) were selected and used as the training input. 576 cells in the last convolutional layer (layer 15) of the CNN were assigned to each of the 14 classes in the dataset. The network was trained in 150,000 iterations. Testing was performed using the preselected key pose frames of the test subject as input. The average population activation of the cells assigned to each class was used to infer the final classification decision (for an exemplary activation pattern see **Figure 8**). **Figure 6** summarizes classification results obtained for different temporal integration schemes of single frame classification results. A framewise classification scheme allows to recognize an action in an instant when the key pose frame is presented to the network. This kind of immediate decision might be crucial for systems which rely on decisions in real time. Not only the processing speed, but also the time necessary to sample and construct the action descriptors is relevant in this context. **Figure 6A** summarizes the framewise

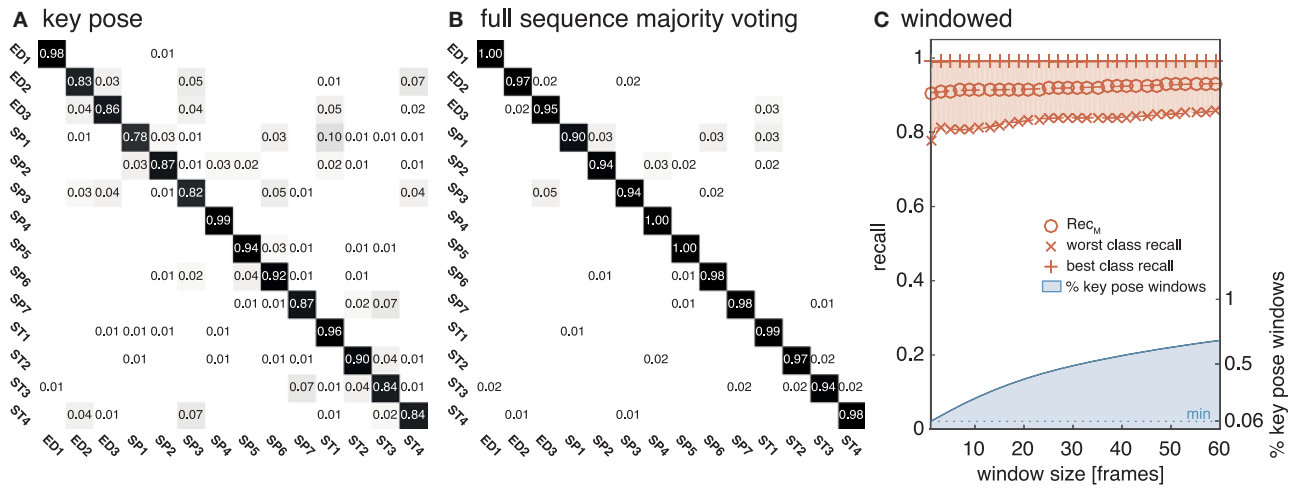


FIGURE 6 | uulmMAD classification performance. The proposed network was trained on the key pose frames extracted from the uulmMAD action recognition dataset. **(A)** Shows the per class classification rates obtained by single key pose frame classification. This allows the recognition of an action in the instant a key pose frame emerges in the input sequence. Average classwise recall (on the diagonal) ranges from 0.78 to 0.98. Some of the notable confusions between classes can be explained by a large visual similarity (e.g., between ED2 and ST4). In **(B)** sequence level majority voting was applied. The final decision is made after evaluating all key pose frames within an action sequence and determining the class with the most frequent occurrence of key poses. The resulting per class values of Rec_M range from 0.94 to 1.00. A sliding window based classification scheme was evaluated in **(C)**. The best and worst per class average recall values together with the average value of Rec_M are displayed for temporal window sizes from 1 to 60 frames. In addition, the percentage of windows containing one or more key pose frames (and thus allow a classification of the action) is shown (blue line).

TABLE 2 | uulmMAD classification performance.

			Rec_M	Inf_M	Mark_M	MCC_M
Single		All frames	67.56 ± 6.06	0.703 ± 0.062	0.762 ± 0.041	0.732 ± 0.051
		Key poses	88.65 ± 5.66	0.915 ± 0.045	0.915 ± 0.043	0.915 ± 0.044
Majority	Windowed*	5 [4]	89.64 ± 4.97	0.919 ± 0.041	0.920 ± 0.040	0.920 ± 0.040
		10 [9]	89.98 ± 4.69	0.922 ± 0.039	0.922 ± 0.039	0.922 ± 0.039
		20 [19]	90.47 ± 4.48	0.924 ± 0.038	0.924 ± 0.037	0.924 ± 0.037
		Full sequence	96.73 ± 2.84	0.981 ± 0.014	0.970 ± 0.033	0.975 ± 0.021

Size [overlap]

classification rates per class (average Rec_M of 0.887, std 0.057). Some of the confusions between classes might be explained by similar visual appearances of the key poses (e.g., ED2 and ST4). Accumulating the classified key poses over a sequence by majority voting increases the classification performance (average Rec_M of 0.967, std 0.028, compare **Figure 6B**), but requires to analyze all frames of a sequence and is thus not well suited for real time applications. As a compromise between classification speed and performance, a sliding window based approach was evaluated. In **Figure 6C**, the best and worst average per class recall is displayed together with the Rec_M for window sizes of $n = [1, ..., 60]$, each with an overlap of $n - 1$. In addition, the relative number of windows which contain at least one key pose (and thus allow a classification) is shown. **Table 2** summarizes the classification performance for different single frame and temporal integration schemes. Single frame performance is, in addition, reported for the evaluation of not only the key pose but the full set of frames. As can be seen, the classification performance decreases significantly but the average recall of Rec_M of 67.56 (std 6.06) indicates, that the learned key pose representations are still rich enough to classify a majority of the frames correctly. Note, that the relative number of correct classifications clearly exceeds the percentage of key pose frames in the dataset (per class average of 7.46 %, std 2.19 %, compare **Figure 1C**).

The model was additionally trained using the Weizmann dataset (Gorelick et al., 2007, see Section 4.2). 1,873 frames (per class average 187.30, std 59.51) were selected as key pose frames utilizing the combined criterion developed in Section 3.1. Except for the number of output features encoding each class (806), the same network and learning parameters were applied. As for the uulmMAD dataset, **Figure 7** gives an overview over the classification performance, by showing confusion matrices for single key pose frame evaluation (**Figure 7A**), full sequence majority voting (**Figure 7B**), as well as best and worst class recall for different sized windows of temporal integration (**Figure 7C**). In comparison to the results reported for the uulmMAD dataset, the gap between the best and worst class recall is considerably increased. This might be explained by a different overall number of available training examples in

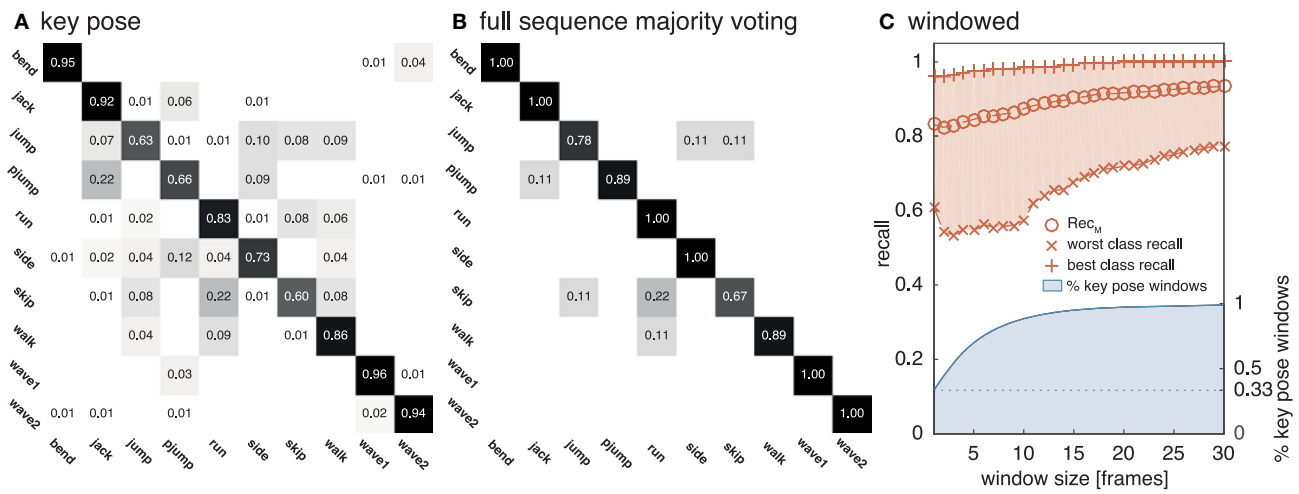


FIGURE 7 | Weizmann classification performance. The network was evaluated on the Weizmann dataset to allow a comparison to other approaches. As in **Figure 6**, **(A)** shows the classifications rates for a classification of single key pose frames per class. **(B)** Displays classwise recognition results for a full sequence evaluation using majority voting. Similar visual appearances might explain the increased rate of confusions for some of the classes (e.g., *run* and *skip*). In **(C)** the average best and worst per class recall values and Recl_M are reported for temporal window sizes between 1 and 30 frames together with the relative number of windows which contain at least one frame classified as key pose.

TABLE 3 | Weizmann classification performance.

			Rec_M	Inf_M	Mark_M	MCC_M
Single		All frames	77.15 ± 6.46	0.810 ± 0.056	0.794 ± 0.068	0.801 ± 0.061
		Key poses	82.15 ± 5.81	0.844 ± 0.061	0.827 ± 0.070	0.835 ± 0.065
Majority	Windowed*	5 [4]	83.50 ± 5.12	0.877 ± 0.043	0.860 ± 0.053	0.868 ± 0.047
		10 [9]	86.40 ± 5.58	0.920 ± 0.027	0.878 ± 0.067	0.899 ± 0.044
		20 [19]	90.35 ± 7.34	0.966 ± 0.023	0.898 ± 0.093	0.930 ± 0.057
		Full sequence	92.22 ± 8.33	0.980 ± 0.023	0.879 ± 0.128	0.927 ± 0.079

**Size [overlap]*

the datasets (the per class average of training examples in the uulmMAD dataset exceeds the Weizmann dataset by a factor of 11.02), higher visual similarities between the classes (the most prominent confusions are observed for *skip*, *jump* and *pjump*), the lack of a sufficient number of descriptive key poses, or a combination hereof. A direct relationship of the classwise performance and the per class number of key pose frames available for training cannot be observed. Even though the least number of key pose frames was extracted for the class *bend*, the second best recall value was achieved. As for the uulmMAD dataset, performance measures are reported for different single frame and temporal integration schemes in **Table 3**. Again, the trained key pose representations achieve a considerable performance even when tested per frame on all frames of the action sequences ($Rec_M = 77.15$, std 6.46). **Table 4** compares the reported classification results on the Weizmann dataset to state-of the art single frame based (second block) and sequence level approaches (third block). In particular, other key pose based action recognition approaches are listed (first block). The direct comparison of different classification architectures, even when evaluated on the same dataset, is often difficult, since different evaluation strategies may have been applied. Thus, whenever possible, the number of considered classes (sometimes the class *skip* is excluded) and the evaluation strategy is listed together with classification performance and speed. Evaluation strategies are either leave-one-subject-out (LOSO), leave-one-action-out (LOAO) or leave-one-out (LOO, not specifying what is left out) cross-validation.

On a sequence level, the classification performance of the proposed approach is on par with almost all other key pose based methods. Only Liu et al. (2013) achieved a noteworthy higher performance (recall of 100). It is important to stress that the compared methods substantially differ in their key pose selection procedures and thus in the underlying conceptual definition of key poses. For example, Weinland and Boyer (2008) and Liu et al. (2013) select key poses that maximize the classification performance in a validation subset of the dataset, whereas (Baysal et al., 2010; Cheema et al., 2011) select and weight candidate pose descriptors dependent on their distinctiveness with respect to the other classes contained in the dataset. In Chaaraoui et al. (2013), key poses are selected independently per class using clustering in combination with a compactness metric. All the above mentioned approaches, except the last one, rely on *inter-class* distributions of pose descriptors to identify key poses, implicitly stating that representativeness is equivalent to distinctiveness (among a known set of classes). If the task at hand is to separate an a priori defined set of actions, this seems to be the superior way of defining key poses for the establishment of temporally sparse representations of actions. On the other hand such poses always describe differences based on comparisons and do not necessarily capture characteristic poses of an action.

The presented approach follows a different principle. Certain properties of image or skeleton based pose features are assumed

TABLE 4 | Weizmann comparison to other approaches.

Catagory		# Actions	Evaluation	fps	Temporal range		
					Sub-sequence		Full sequence
					Recall	# Frames	Recall
Key pose	Weinland and Boyer, 2008	10	LOSO	–	–	–	93.6
	Baysal et al., 2010	9	LOO	–	–	–	92.6
	Cheema et al., 2011	9	LOO	–	–	–	91.6
	Chaaraoui et al., 2013	9	LOSO	124	–	–	92.8
	Liu et al., 2013	10	LOSO	–	–	–	**100**
Single frame	Niebles and Fei-Fei, 2007	9	LOSO	–	55	1	72.8
	Fathi and Mori, 2008	10	LOO	0.25–5	**99.9**	1	**100**
	Schindler and van Gool, 2008	9	LOSO	–	93.5	1	**100**
	Hoai et al., 2011	10	LOSO	–	87.7	1	–
Full sequence	Jhuang et al., 2007	9	–	0.83	–	–	98.8
	Klaser et al., 2008	10	LOSO	–	–	–	84.3
	Grundmann et al., 2008	9	LOAO	–	–	–	94.6
	Ikizler and Duygulu, 2009	9	LOO	–	–	–	**100**
	Bregonzio et al., 2009	10	LOSO	–	–	–	96.7
	Sun and Liu, 2012	10	LOO	–	–	–	97.8
	Beaudry et al., 2016	10	LOO	51.63	–	–	**100**
	Presented approach	10	LOSO	**1,000**	82.2	1	92.2

Bold values indicate maximum recall/fps values per column.

to co-occur with characteristic body configurations and thus are used to identify key pose frames. The feature characteristic indicating a key pose and the representations/descriptors used for the recognition of a pose do not necessarily have a close relationship. In doing so, we accept the fact that the selected poses are not guaranteed to be very distinctive and some even may occur in more than one action in exactly the same way. Key poses are assumed to be the most representative poses of a particular action, not in comparison, but in general. Nevertheless, the presented results demonstrate that a *feature-driven*, pose centered key pose selection mechanism is capable of achieving the same level of performance, without loosing generality.

Most key pose based approaches in the literature try to assign single frames of an image sequence to key pose frames with a high similarity, temporally integrate the result (e.g., by using histograms or majority voting) and perform a classification of the action on a sequence level. The result of single frame action recognition based on the extracted key poses (directly linking key poses to actions) is rarely reported. Single frame based approaches (see **Table 4**, second block), however, try to perform action classification using information solely extracted within one frame (two frames if optical flow is part of the descriptor) and achieve impressive results. In direct comparison, the single frame performance of the presented approach (Rec_M of 82.15 for key pose evaluation and 77.15 for the classification of all single frames, compare **Table 3**) cannot compete with the other methods, which, on the contrary, utilize all frames during learning to maximize classification performance in the test training dataset. The presented approach, however, achieves a single frame performance of $Rec_M = 77.15$ when evaluated over all frames, although in case of the Weizmann dataset only a per class average of 33.84 % (std 8.63 %) of all frames is used for training.

In the third block of **Table 4**, selected approaches performing action recognition on a sequence level using a variety of different representations and classification architectures are listed. Note that in an overall comparison, (I) due to the transfer on neuromorphic hardware, the presented approach achieves the highest processing speed[6] while consuming a minimal amount of energy, and (II) due to fact, that we aim at executing the model on a single TrueNorth chip we only use input maps with a resolution of 32 × 32 (using 4,064 of the 4,096 cores available on one chip). This is no limitation of the employed *Eedn* framework, which allows to realize models which run on systems with more than one chip (Esser et al., 2016; Sawada et al., 2016). An increased input resolution, as well as the use of more than two flow direction and contour orientation maps might help in separating classes with a high visual similarity (e.g., *skip*, *jump*, and *run*).

5.2. Comparison to Full Sequence Learning

To address the question whether and how the proposed classification architecture might benefit from using all frames (as opposed to only key pose frames) during training, we performed exactly the same training and testing procedure twice on the uulmMAD dataset. First, only key pose frames were presented during training, while second, all frames were provided during the training phase. Likewise, testing was performed just on the preselected key pose frames, as well as the full set of frames. **Table 5** compares the average recall under the different training (rows) and testing conditions (columns) for single frame evaluation and sequence level majority voting.

In both cases, training and testing on key pose frames achieves the highest performance. However, the observed differences between the two training conditions could not shown to be significant, neither when testing on key poses nor on the full set of frames. Nevertheless, having a closer look at the activation patterns of the network reveals some insights on the effectiveness of the two variants of trained representations. **Figure 8** shows the average activation pattern of the 14 cell populations in layer 15 assigned to the individual classes of a network trained on key pose frames and tested on all frames of the action SP2 (jumping jack). The displayed activation levels clearly show how

TABLE 5 | uulmMAD key pose versus all frame learning.

Train \ Test	Framewise		Majority voting	
	Key poses	All frames	Key poses	All frames
Key poses	**88.65 ± 5.66**	67.56 ± 6.06	**96.73 ± 2.84**	93.29 ± 7.05
All frames	85.84 ± 7.15	72.84 ± 8.25	95.27 ± 3.93	94.70 ± 7.03

Bold values indicate the maximum average recall for framewise and full sequence majority voting classification schemes.

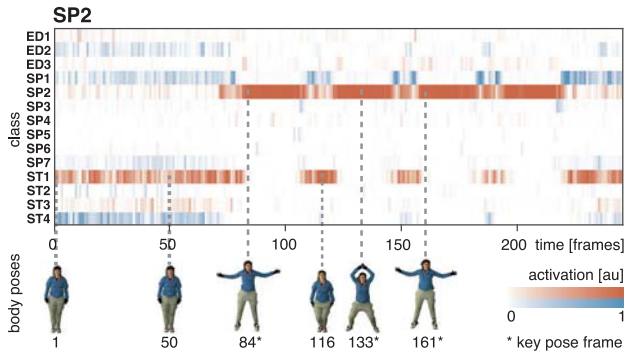


FIGURE 8 | Activation of cell populations. The activations of the cell populations in the last convolutional layer of the DCNN assigned to the 14 classes of the uulmMAD dataset are displayed for a network trained only on key pose frames and tested on all frames of the action SP2 (jumping jack). The activation level of the cell population with the maximum activation (red) and the remaining populations (blue) is encoded by color intensity. Corresponding poses are displayed for selected frames (bottom row). Key pose frames are marked by asterisks. The activation pattern shows how the cell population assigned to the class SP2 selectively responds to frames in the temporal neighborhood of the corresponding key pose frames. At the beginning and the end of the sequence, as well as in between the occurrence of key pose frames, different cell populations achieve the maximum average activation and thus result in misclassifications. Written informed consent for the publication of exemplary images was obtained from the shown subject.

[6]Image preprocessing and key pose selection is not integrated in the estimated processing time. Optical flow estimation can be performed on a second TrueNorth chip (Brosch and Neumann, 2016).

the trained representations of the corresponding class selectively respond within the temporal neighborhood of the key pose frames. Frames sampled from periods without the presence of key pose frames (at the beginning and the end of the sequence, as well as in between key pose frames) result mostly in a large activation of other cell populations and thus in misclassifications. This is in line with the results shown in **Table 5**, which indicate that classification performance increases under both training conditions when testing is only performed on key pose frames. At this point we can conclude that, compared to a training on the full set of frames, key pose based learning of class specific representations at least performs at an equal level. Whether there is any benefit of training exclusively on key pose frames next to an increased learning speed, remains, however, an open question. **Figure 9** summarizes the per class activation levels of the cell populations which resulted in a correct classification. For almost all classes (except ED3), the activation level is significantly increased when training was performed on key pose frames only. This might become a very important property in situations where it is not an option to accept any false negatives. Applying a threshold on the activation levels would allow to eliminate false negatives, while key pose based training would decrease the number of positive examples rejected by the fixed threshold. Thus, thresholding might further increase the performance for the key pose based training reported so far. Taken together, key pose based learning achieves a slightly increased classification performance with an increased selectivity of the cell populations and thus a higher confidence of the classification decisions.

5.3. Cross-Dataset Evaluation

Learning to classify input samples and the associated representations is conducted with the aim to robustly predict future outputs and, thus, generalize for new input data. Here, we evaluate such network capability by evaluating the classification of the trained network using input data across different datasets. More precisely, cross-dataset evaluation was performed to evaluate how the learned representations generalize over different datasets. The preselected key pose frames of the uulmMAD and the Weizmann dataset were used for both training and testing constellations. Performance is reported for two classes, one being *one-handed wave* (ED1 and *wave1*), which is available in both datasets. The second class was formed by combining the visually similar classes SP2/SP6 and *jack*/*wave2* during evaluation into one joint class *raising two hands*. Training was performed on the full set of classes in both cases. Thus, for *one-handed wave* a random guess classifier would achieve a recall of either 7.14 (uulmMAD) or 10.00 (Weizmann). In case of the combined class *raising two hands*, the recall chance level increases to 14.29 (uulmMAD) and 20.00 (Weizmann), respectively. **Table 6** shows the result for *one-handed wave* for the two testing (row) and training (column) setups alongside with exemplary pictures of the classes from both datasets. When training was performed on the Weizmann dataset, the recall performance for examples from the uulmMAD dataset is still considerable (loss of 24.07). Training on the uulmMAD and testing on the Weizmann dataset results in an increased performance loss, but still achieves a recall of 53.03.

In case of the combined class *raising two hands*, the performance loss is below 30 for both training and testing configurations. **Table 7** shows the achieved performance in detail for each of the four classes in isolation and their combination. Note that when trained on the uulmMAD dataset, *jumping jack* is recognized almost without any loss of performance. Vice versa, SP2 is often confused with *wave2* when training was performed on the Weizmann dataset. This may be explained by the large visual similarities between the classes.

The proposed approach shows promising generalization capabilities, which might partially be explained by the class-independent, *feature-driven* selection of the key pose frames.

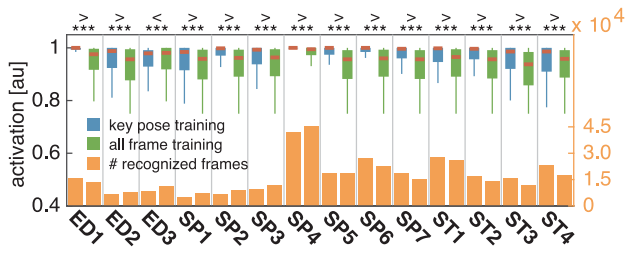


FIGURE 9 | Comparison of activation levels. The activation levels of the cell populations which resulted in correct classifications are displayed per class for key pose based (blue) and all frame (green) training alongside with the total number of correct classifications under both conditions (yellow). In case of key pose based training, activation levels are significantly increased, reflecting a higher confidence of the classification decision. Increased confidences are useful in situations where thresholding is applied on the activation level, e.g., to reduce the number of false negatives.

6. CONCLUSION AND DISCUSSION

The presented work consists of two main contributions. First, a *feature-driven* key pose selection mechanism is proposed, which builds upon evidences about human action perception. The selection mechanism does not utilize any information about the inter- or intra-class distribution of the key poses (or key pose descriptors) to optimize the classification accuracy. It is demonstrated, that the classification accuracy is on par with state-of-the-art key pose based action recognition approaches, while only motion and form related feature characteristics are used to select a key pose frame. Second, we propose a biologically inspired architecture combining form and motion information to learn hierarchical representations of key pose frames. We expect such hierarchical feature representations to make the recognition more robust against clutter and partial occlusions, in comparison to holistic shape representations of the full body configurations used in previous approaches. Form and motion pattern representations are established employing a neuromorphic deep convolutional neural network. The trained network is mapped onto the *IBM Neurosynaptic System* platform, which enables a computationally and energy efficient execution.

6.1. Relation to Other Work

The presented results demonstrate, that classifying actions using a minimal amount of temporal information is in principle

TABLE 6 | Cross-dataset evaluation one-handed wave.

Train / Test	uulmMAD	Weizmann	Loss
uulmMAD	**100**	75.93	24.07
Weizmann	53.03	**100**	46.97



Bold values indicate the maximum recall values per column. Written informed consent for the publication of exemplary images was obtained from the shown subjects (uulmMAD).

TABLE 7 | Cross-dataset evaluation raising two hands.

Train / Test		uulmMAD			Weizmann			Loss
		SP2	**SP6**	**Comb**	***jack***	***wave2***	**Comb**	**Comb**
uulmMAD	SP2	**96.97**	0.00	97.92	24.65	51.04	70.30	27.62
	SP6	1.14	**97.73**		0.00	64.90		
Weizmann	*jack*	95.96	0.00	79.80	**100**	0.00	100	20.20
	wave2	31.82	31.82		0.00	**100**		



Bold values indicate the maximum recall values per column. Written informed consent for the publication of exemplary images was obtained from the shown subjects (uulmMAD).

possible. This is in line with results from other action recognition approaches. For example, Schindler and van Gool (2008) reported that actions can be successfully recognized using snippets of three or even less frames. In their work, the length of the temporal window used for the classification of an action sequence was systematically varied. The most important result was that a reliable action recognition can be achieved by only using individual snippets, i.e. up to three consecutive frames in temporal order. The question whether there are special "key snippets" of frames, which are particularly useful for the recognition of an action and how they might be defined, however, remains open.

Inspired by evidences from perceptual studies (Thurman and Grossman, 2008; Thirkettle et al., 2009), key poses are potential candidates for representing such special events in articulated motion sequences. Unlike the majority of other approaches reported in the literature (e.g., Baysal et al., 2010; Liu et al., 2013), the proposed key pose selection mechanism identifies key pose frames without optimizing the *inter-class* distinctiveness or classification performance of the selected key poses. The *feature-driven* selection criterion proposed in this work combines form and motion information and allows the identification of key poses without any knowledge about other classes. It extends a previous proposal utilizing local temporal extrema in the motion energy as a function of time (Layher et al., 2014) by additionally taking a measure of extendedness of the silhouette shape into account. Given that these features are entirely data-driven, this has two major implications. On the one hand, the selected poses are independent of any other class and thus are more likely to generalize over different sets of actions. This property is appreciated and valuable in many applications since it does not require any prior knowledge about the distribution of classes/poses in other datasets. On the other hand, there is no guarantee, however, that a learned key pose representation is not part of more than one action and thus results in ambiguous representations. This may lead to drawbacks and deteriorations of the model performance in terms of classification rates for rather ambiguous sequences with similar pose articulations. We argue that, although the proposed key pose selection criterion might not result in the best classification performance on all action recognition datasets in isolation, it selects key pose frames which capture the nature of an action in general (independent of a specific dataset). In addition, the reported results demonstrate, that there is no substantial loss in performance when comparing the proposed *feature-driven* key pose selection mechanism to performance optimizing key pose approaches in literature. In contrast to other action recognition approaches building upon convolutional neural networks, the proposed model does not aim at establishing representations which capture the temporal relationship between successive frames. This can be accomplished by e.g., directly feeding spatiotemporal input to the network and applying 3D convolutions (e.g., Baccouche et al., 2011; Ji et al., 2013) or by applying a *multiple spatio-temporal scales neural network* (MSTNN; Jung et al., 2015). Instead, in this work, the employed DCNN exclusively aims at identifying class specific key pose frames as events in an image (and optical flow) stream.

The investigation reported in this work adds an important piece to the debate of how representations for action sequence analysis might be organized. Some previous approaches have utilized motion and form information for the classification of action categories. For example, Giese and Poggio (2003) proposed that biological motion sequences representing articulated movements of persons is subdivided into two parallel streams in primate visual cortex. In particular, the authors argue that motion patterns are represented in a hierarchy and these are paralleled by regular temporal sampling of static frames from the same input sequence. This model architecture has been extended

in Layher et al. (2014) suggesting that instead of representing sequences of static frames only key poses need to be selected. As a candidate criterion, the motion energy is calculated over time and local energy minima depict reversal points of bodily articulation. Such reversals, in turn, most likely coincide with extremal articulations and thus can be utilized to select a key pose in such articulation sequences. While these models focus on cortical architecture of visual dorsal and ventral streams, other computer vision approaches also consider combinations of motion and form information for action recognition. While the proposal of Jhuang et al. (2007) builds on a hierarchy of cascaded form and motion representations, the approach of Schindler and van Gool (2008) also utilized two parallel streams of motion and form processing. Both streams generate feature vectors of equal length which are subsequently concatenated including a weighting of the relative strength of their contribution. An evaluation of the relative weights showed that a fusion with 70 % motion against a 30 % form feature concatenation yielded the best performance on the Weizmann dataset. On the contrary, Schindler et al. (2008) demonstrated that emotion categories can be classified using static images only which are processed by a multi-scale bank of filters with subsequent pooling operation and dimension reduction. Our findings add new insights to the investigation of utilizing form/shape and motion information in biological/articulated motion analysis for action recognition. Our findings highlight that key poses defined by events of temporal extrema in motion energy and dynamic object silhouette features reliably reflect a high information content regarding the whole action sequence. In other words, key poses can be detected by an entirely *feature-driven* approach (without utilizing any a priori model of actions in the sequence) and that the associated temporal events contain a high proportion of the information about the main components of the action sequence.

We successfully trained a DCNN of 15 convolutional layers on the key pose frames used as input, which were assigned to different action classes. The network was trained using the *energy-efficient deep neuromorphic networks* (*Eedn*) framework (Esser et al., 2016) and executed on a TrueNorth NS1e board (Merolla et al., 2014). The results show that action recognition can be performed on mobile robotic platforms under real-time constraints while consuming a minimal amount of energy. The reduced energy consumption and the high performance in classification rate (compare **Table 4**) makes such a model architecture a valuable candidate for applications in mobile or remote control scenarios in which autonomy in energy supply and external control are constraints of core importance. The automatic selection of key pose information for the classification mechanism is a key step to make use of the demonstrated parameters.

Although some classes contained examples with highly similar visual appearances, the network shows an impressive single frame recognition performance when tested on key frames. Even when tested on the full set of frames, recognition performance is still significantly above chance level. Using a simple temporal integration scheme, we show that the results are on par with competing key pose based action recognition approaches (**Table 4**). Cross-dataset evaluation of classes with the same/a similar visual appearance in both datasets shows how the learned representations generalize over the different datasets (training was performed on the full set of classes).

6.2. Shortcomings and Possible Further Improvements

Currently, the optical flow estimation and the key pose selection are performed prior to the training and the classification of input sequences. To realize a complete neuromorphic implementation of the presented approach, optical flow can be estimated as well on neuromorphic hardware following the principles described in Brosch and Neumann (2016). A neuromorphic implementation of localizing the local extrema in the motion energy and the extendedness of a person's silhouette could be realized on top of the flow estimation process. In addition, dynamic vision sensors (e.g., iniLabs DVS128) are an option to directly feed a network similar to the proposed one with spike-based sensory streams. First attempts to realize an action recognition system using such sparse asynchronous data streams have already shown promising results (Tschechne et al., 2014).

The presented approach does not make use of any temporal relationship between the identified events (key poses) in an action sequence. Thus, the reversed, or scrambled presentation of images (and optical flow) of a sequence would result in an assignment to an action class, although, the visual appearance of the sequence is totally different. A modeling or learning of the temporal relationships between the key pose frames, e.g., their temporal order, would help in reducing ambiguities and thus increase sequence-wide or windowed classification rates. In case of the proposed approach, this could be achieved by employing, e.g., *long short-term memory* cells (LSTM; Hochreiter and Schmidhuber, 1997), which are candidates to realize the learning of temporal relationships without loosing the invariance against changes in speed. The simple majority voting based integration scheme was chosen, because of hardware limitations and to focus on an analysis of the importance of key poses in the context of action representation and recognition.

We also did not apply a weighted majority voting scheme using the confidences of the frame-wise predictions or apply thresholding on the predictions. Both strategies might further increase the classification performance but again would weaken the focus on the analysis of key pose base representations of action sequences.

The proposed architecture of a deep convolutional neural network (DCNN) as depicted in **Figure 3** builds increasingly more complex feature representations through learning from initial simple features. It would be interesting to investigate the feature selectivities of the feature representations that have been established by the learning. Such a study would potentially shed light about the structure of the feature compositions (and their hierarchical organization) which lead to the selectivity of the key poses in relation to the action sequences to be classified. Some approaches analyzing the low-, intermediate-, and higher-level feature representations have recently been proposed in the literature (Zeiler and Fergus, 2014; Güçlü and van Gerven, 2015; Mahendran and Vedaldi, 2016). Such approaches have so far investigated CNNs for static inputs only. For that reason, some principles might also be useful for the analysis of key pose representations. In addition, the consideration of short-term

spatio-temporal feature representations will help to extend the scope of the overall study of visualizing internal representations after learning. We expect necessary major efforts to carefully develop an extended set of tools which is beyond the scope of the modeling investigation presented here.

Overall, the presented results show, that the learned key pose representations allow the classification of actions using a minimal amount of temporal information. By implementing the proposed DCNN on the TrueNorth chip, we show that real-time action recognition relying on the proposed principles is possible while consuming a minimal amount of energy, as reported for the runtime environments of the *IBM Neurosynaptic System* (Esser et al., 2016).

AUTHOR CONTRIBUTIONS

Conceived and designed the approach: GL, TB, HN; Implemented the architecture: GL, TB; Performed the simulations: GL; Analyzed the data: GL; Wrote the paper: GL, HN.

FUNDING

This research has been supported by the Transregional Collaborative Research Centre SFB/TRR 62 "A Companion Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG). HN has been supported by the collaborative project "SenseEmotion" funded by the German Federal Ministry of Education and Research (BMBF).

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support via a field test agreement between Ulm University and IBM Research Almaden, and particularly the TrueNorth team at IBM for their support. In addition, the support of NVIDIA Corporation with the donation of a Tesla K40 GPU used for this research is also gratefully acknowledged. The authors like to express their gratefulness for the reviewers' efforts and suggestions that helped to improve the manuscript.

REFERENCES

Aggarwal, J., and Ryoo, M. (2011). Human activity analysis: a review. *ACM Comput. Surv.* 43, 16:1–16:43. doi: 10.1145/1922649.1922653

Akopyan, F., Sawada, J., Cassidy, A., Alvarez-Icaza, R., Arthur, J., Merolla, P., et al. (2015). Truenorth: design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip. *IEEE Trans. Comput. Aided Design Integr. Circ. Syst.* 34, 1537–1557. doi: 10.1109/TCAD.2015.2474396

Amir, A., Datta, P., Risk, W. P., Cassidy, A. S., Kusnitz, J. A., Esser, S. K., et al. (2013). "Cognitive computing programming paradigm: a corelet language for composing networks of neurosynaptic cores," in *International Joint Conference on Neural Networks (IJCNN), 2013* (IEEE), 1–10. Available online at: http://ieeexplore.ieee.org/abstract/document/6707078/?reload=true

Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., and Baskurt, A. (2011). "Sequential Deep Learning for Human Action Recognition," in *2nd International Workshop on Human Behavior Understanding (HBU),* eds A. A. Salahand and B. Lepri (Berlin; Heidelberg: Springer), 29–39.

Barron, J. L., Fleet, D. J., and Beauchemin, S. S. (1994). Performance of optical flow techniques. *Int. J. Comput. Vis.* 12, 43–77. doi: 10.1007/BF01420984

Baysal, S., Kurt, M. C., and Duygulu, P. (2010). "Recognizing human actions using key poses," in *20th International Conference on Pattern Recognition (ICPR), 2010* (IstanbulL), 1727–1730. Available online at: http://ieeexplore.ieee.org/abstract/document/5597477/

Beaudry, C., Péteri, R., and Mascarilla, L. (2016). An efficient and sparse approach for large scale human action recognition in videos. *Mach. Vis. Appl.* 27, 529–543. doi: 10.1007/s00138-016-0760-z

Bobick, A. F., and Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Trans. Patt. Anal. Mach. Intel.* 23, 257–267. doi: 10.1109/34.910878

Boussaoud, D., Ungerleider, L. G., and Desimone, R. (1990). Pathways for motion analysis: cortical connections of the medial superior temporal and fundus of the superior temporal visual areas in the macaque. *J. Comp. Neurol.* 296, 462–495. doi: 10.1002/cne.902960311

Bregonzio, M., Gong, S., and Xiang, T. (2009). "Recognising action as clouds of space-time interest points," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009* (Miami, FL), 1948–1955. Available online at: http://ieeexplore.ieee.org/abstract/document/5206779/

Brin, S., and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Comput. Netw.* 30, 107–117. doi: 10.1016/s0169-7552(98)00110-x

Brosch, T., and Neumann, H. (2016). "Event-based optical flow on neuromorphic hardware," in *Proceedings of the 9th EAI International Conference on Bio-Inspired Information and Communications Technologies (Formerly BIONETICS), 2015* (New York, NY: ICST), 551–558. doi: 10.4108/eai.3-12-2015.2262447

Cao, X., Ning, B., Yan, P., and Li, X. (2012). Selecting key poses on manifold for pairwise action recognition. *IEEE Trans. Indust. Inform.* 8, 168–177. doi: 10.1109/TII.2011.2172452

Casile, A., and Giese, M. A. (2005). Critical features for the recognition of biological motion. *J. Vis.* 5:6. doi: 10.1167/5.4.6

Chaaraoui, A. A., Climent-Pérez, P., and Flórez-Revuelta, F. (2013). Silhouette-based human action recognition using sequences of key poses. *Patt. Recogn. Lett.* 34, 1799–1807. doi: 10.1016/j.patrec.2013.01.021

Cheema, S., Eweiwi, A., Thurau, C., and Bauckhage, C. (2011). "Action recognition by learning discriminative key poses," in *IEEE Workshops of the International Conference on Computer Vision (ICCV Workshops), 2011* (Barcelona), 1302–1309. Available online at: http://ieeexplore.ieee.org/abstract/document/6130402/

Dalal, N., and Triggs, B. (2005). "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2005*, Vol. 1 (San Diego, CA: IEEE), 886–893. Available online at: http://ieeexplore.ieee.org/abstract/document/1467360/

Dedeoğlu, Y., Töreyin, B. U., Güdükbay, U., and Çetin, A. E. (2006). "Silhouette-based method for object classification and human action recognition in video," in *European Conference on Computer Vision (ECCV), Workshop on HCI Computer Vision in Human-Computer Interaction, 2006*, eds T. S. Huang, N. Sebe, M. S. Lew, V. Pavlović, M. Kölsch, A. Galata, and B. Kisačanin (Berlin; Heidelberg: Springer), 64–77.

Dollár, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). "Behavior recognition via sparse spatio-temporal features," in *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (PETS), 2005* (Beijing: IEEE), 65–72. Available online at: http://ieeexplore.ieee.org/abstract/document/1570899/

Esser, S. K., Merolla, P. A., Arthur, J. V., Cassidy, A. S., Appuswamy, R., Andreopoulos, et al. (2016). Convolutional networks for fast, energy-efficient neuromorphic computing. *Proc. Natl. Acad. Sci. U.S.A.* 113, 11441–11446. doi: 10.1073/pnas.1604850113

Fathi, A., and Mori, G. (2008). "Action recognition by learning mid-level motion features," in *IEEE Conference on Computer Vision and Pattern Recognition*

(CVPR), 2008 (IEEE), 1–8. Available online at: http://ieeexplore.ieee.org/abstract/document/4587735/

Felleman, D. J., and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1, 1–47. doi: 10.1093/cercor/1.1.1

Ferrari, V., Fevrier, L., Jurie, F., and Schmid, C. (2008). Groups of adjacent contour segments for object detection. *IEEE Trans. Patt. Anal. Mach. Intel.* 30, 36–51. doi: 10.1109/TPAMI.2007.1144

Freund, Y., and Schapire, R. E. (1995). "A desicion-theoretic generalization of on-line learning and an application to boosting," in *European Conference on Computational Learning Theory*, ed P. Vitányi (Berlin; Heidelberg: Springer), 23–37. doi: 10.1007/3-540-59119-2_166

Garcia, J. O., and Grossman, E. D. (2008). Necessary but not sufficient: motion perception is required for perceiving biological motion. *Vis. Res.* 48, 1144–1149. doi: 10.1016/j.visres.2008.01.027

Gavrila, D. M. (1999). The visual analysis of human movement: a survey. *Comput. Vis. Image Unders.* 73, 82–98. doi: 10.1006/cviu.1998.0716

Giese, M. A., and Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nat. Rev. Neurosci.* 4, 179–192. doi: 10.1038/nrn1057

Glodek, M., Layher, G., Heilemann, F., Gawrilowicz, F., Palm, G., Schwenker, F., et al. (2014). "uulmMAD–a human action recognition dataset for ground-truth evaluation and investigation of view invariances," in *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction: Third IAPR TC3 Workshop on Multimodal Pattern Recognition of Social Signals in Human-Computer Interaction (MPRSS), 2014* (Cham: Springer), 77–91. doi: 10.1007/978-3-319-14899-1_8

Gong, W., Bagdanov, A. D., Roca, F. X., and Gonzàlez, J. (2010). "Automatic key pose selection for 3D human action recognition," in *6th International Conference on Articulated Motion and Deformable Objects (AMDO), 2010*, eds F. J. Perales and R. B. Fisher (Berlin; Heidelberg: Springer), 290–299.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. *Trans. Patt. Anal. Mach. Intel.* 29, 2247–2253. doi: 10.1109/TPAMI.2007.70711

Grauman, K., and Darrell, T. (2005). "The pyramid match kernel: discriminative classification with sets of image features," in *10th IEEE International Conference on Computer Vision (ICCV), 2005*, Vol. 2 (Beijing), 1458–1465. Available online at: http://ieeexplore.ieee.org/abstract/document/1544890/

Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., et al. (2000). Brain areas involved in perception of biological motion. *J. Cogn. Neurosci.* 12, 711–720. doi: 10.1162/089892900562417

Grundmann, M., Meier, F., and Essa, I. (2008). "3D shape context and distance transform for action recognition," in *19th International Conference on Pattern Recognition (ICPR), 2008* (Tampa, FL), 1–4. Available online at: http://ieeexplore.ieee.org/abstract/document/4761435/

Güçlü, U., and van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* 35, 10005–10014. doi: 10.1523/JNEUROSCI.5023-14.2015

Hoai, M., Lan, Z.-z., and De la Torre, F. (2011). "Joint segmentation and classification of human actions in video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011* (Colorado Springs, CO), 3265–3272. Available online at: http://ieeexplore.ieee.org/abstract/document/5995470/

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Ikizler, N., and Duygulu, P. (2009). Histogram of oriented rectangles: a new pose descriptor for human action recognition. *Image Vis. Comput.* 27, 1515–1526. doi: 10.1016/j.imavis.2009.02.002

Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167.

Jellema, T., and Perrett, D. I. (2003). Cells in monkey STS responsive to articulated body motions and consequent static posture: a case of implied motion? *Neuropsychologia* 41, 1728–1737. doi: 10.1016/S0028-3932(03)00175-1

Jhuang, H., Serre, T., Wolf, L., and Poggio, T. (2007). "A biologically inspired system for action recognition," in *IEEE International Conference on Computer Vision (ICCV), 2007* (Rio de Janeiro: IEEE), 1–8. Available online at: http://ieeexplore.ieee.org/abstract/document/4408988/

Ji, S., Xu, W., Yang, M., and Yu, K. (2013). 3d convolutional neural networks for human action recognition. *IEEE Trans. Patt. Anal. Mach. Intel.* 35, 221–231. doi: 10.1109/TPAMI.2012.59

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Percept. Psychophys.* 14, 201–211. doi: 10.3758/BF03212378

Johansson, G. (1976). Spatio-temporal differentiation and integration in visual motion perception. *Psychol. Res.* 38, 379–393. doi: 10.1007/BF00309043

Jung, M., Hwang, J., and Tani, J. (2015). Self-organization of spatio-temporal hierarchy via learning of dynamic visual image patterns on action sequences. *PLoS ONE* 10:e0131214. doi: 10.1371/journal.pone.0131214

Kläser, A., Marszałek, M., and Schmid, C. (2008). "A spatio-temporal descriptor based on 3D-gradients," in *19th British Machine Vision Conference (BMVC), 2008*, eds M. Everingham and C. Needham (Leeds: British Machine Vision Association), 99.1-99.10. Available online at: http://www.bmva.org/bmvc/2008/papers/275.html

Kourtzi, Z., and Kanwisher, N. (2000). Activation in human MT/MST by static images with implied motion. *J. Cogn. Neurosci.* 12, 48–55. doi: 10.1162/08989290051137594

Lange, J., and Lappe, M. (2006). A model of biological motion perception from configural form cues. *J. Neurosci.* 26, 2894–2906. doi: 10.1523/JNEUROSCI.4915-05.2006

Layher, G., Giese, M. A., and Neumann, H. (2014). Learning representations of animated motion sequences - a neural model. *Top. Cogn. Sci.* 6, 170–182. doi: 10.1111/tops.12075

LeCun, Y., Kavukcuoglu, K., and Farabet, C. (2010). "Convolutional networks and applications in vision," in *IEEE International Symposium on Circuits and Systems (ISCAS), 2010* (Paris), 253–256. Available online at: http://ieeexplore.ieee.org/abstract/document/5537907/

Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *CoRR*, abs/1312.4400.

Liu, L., Shao, L., Zhen, X., and Li, X. (2013). Learning discriminative key poses for action recognition. *IEEE Trans. Cybernet.* 43, 1860–1870. doi: 10.1109/TSMCB.2012.2231959

Lucas, B. D., and Kanade, T. (1981). "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI), 1981* (San Francisco, CA: Morgan Kaufmann Publishers Inc.), 674–679.

Lv, F., and Nevatia, R. (2007). "Single view human action recognition using key pose matching and viterbi path searching," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007* (IEEE), 1–8. Available online at: http://ieeexplore.ieee.org/abstract/document/4270156/

Mahendran, A. and Vedaldi, A. (2016). Visualizing deep convolutional neural networks using natural pre-images. *Int. J. Comput. Vis.* 120, 233–255. doi: 10.1007/s11263-016-0911-8

Mather, G., and Murdoch, L. (1994). Gender discrimination in biological motion displays based on dynamic cues. *Proc. R. Soc. Lond. B. Biol. Sci.* 258, 273–279. doi: 10.1098/rspb.1994.0173

McCann, S., and Lowe, D. G. (2012). "Local naive bayes nearest neighbor for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012*, 3650–3656. Available online at: http://ieeexplore.ieee.org/abstract/document/6248111/

Merolla, P. A., Arthur, J. V., Alvarez-Icaza, R., Cassidy, A. S., Sawada, J., Akopyan, F., et al. (2014). A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* 345, 668–673. doi: 10.1126/science.1254642

Niebles, J. C., and Fei-Fei, L. (2007). "A hierarchical model of shape and appearance for human action classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007* (Minneapolis, MN: IEEE), 1–8. Available online at: http://ieeexplore.ieee.org/abstract/document/4270157/

Ogale, A. S., Karapurkar, A., and Aloimonos, Y. (2007). "View-invariant modeling and recognition of human actions using grammars," in *Dynamical Vision*, eds R. Vidal, A. Heyden, and Y. Ma (Berlin; Heidelberg: Springer), 115–126.

Oram, M., and Perrett, D. (1994). Responses of anterior superior temporal polysensory (STPa) neurons to "Biological Motion" stimuli. *J. Cogn. Neurosci.* 6, 99–116. doi: 10.1162/jocn.1994.6.2.99

Oram, M. W., and Perrett, D. I. (1996). Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey. *J. Neurophysiol.* 76, 109–129.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image Vis. Comput.* 28, 976–990. doi: 10.1016/j.imavis.2009.11.014

Powers, D. M. W. (2013). "A computationally and cognitively plausible model of supervised and unsupervised learning," in *6th International Conference on Advances in Brain Inspired Cognitive Systems (BICS), 2013*, eds D. Liu, C. Alippi, D. Zhao, and A. Hussain (Berlin; Heidelberg: Springer), 145–156. doi: 10.1007/978-3-642-38786-9_17

Puce, A., and Perrett, D. (2003). Electrophysiology and brain imaging of biological motion. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 358, 435–445. doi: 10.1098/rstb.2002.1221

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025. doi: 10.1038/14819

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating error. *Nature* 323, 533–536. doi: 10.1038/323533a0

Sawada, J., Akopyan, F., Cassidy, A. S., Taba, B., Debole, M. V., Datta, P., et al. (2016). "Truenorth ecosystem for brain-inspired computing: scalable systems, software, and applications," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC), 2016* (Piscataway, NJ: IEEE Press), 12:1–12:12.

Schindler, K., Gool, L. V., and de Gelder, B. (2008). Recognizing emotions expressed by body pose: a biologically inspired neural model. *Neural Netw.* 21, 1238–1246. doi: 10.1016/j.neunet.2008.05.003

Schindler, K., and van Gool, L. (2008). "Action snippets: how many frames does human action recognition require?" in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008* (Anchorage), 1–8. Available online at: http://ieeexplore.ieee.org/abstract/document/4587730/

Sokolova, M., and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Inform. Process. Manage.* 45, 427–437. doi: 10.1016/j.ipm.2009.03.002

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. A. (2014). Striving for simplicity: the all convolutional net. *CoRR*, abs/1412.6806.

Sun, Q., and Liu, H. (2012). "Action disambiguation analysis using normalized google-like distance correlogram," in *11th Asian Conference on Computer Vision (ACCV), 2012*, eds K. M. Lee, Y. Matsushita, J. M. Rehg, and Z. Hu (Berlin; Heidelberg: Springer), 425–437. doi: 10.1007/978-3-642-37431-9_33

Thirkettle, M., Benton, C. P., and Scott-Samuel, N. E. (2009). Contributions of form, motion and task to biological motion perception. *J. Vis.* 9:28. doi: 10.1167/9.3.28

Thurau, C., and Hlavác, V. (2008). "Pose primitive based human action recognition in videos or still images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008* (IEEE), 1–8. Available online at: http://ieeexplore.ieee.org/abstract/document/4587721/

Thurman, S. M., and Grossman, E. D. (2008). Temporal "Bubbles" reveal key features for point-light biological motion perception. *J. Vis.* 8:28. doi: 10.1167/8.3.28

Tschechne, S., Sailer, R., and Neumann, H. (2014). "Bio-inspired optic flow from event-based neuromorphic sensor input," in *6th International Workshop on Artificial Neural Networks in Pattern Recognition (IAPR), 2014*, eds N. El Gayar, F. Schwenker, and C. Schwenker (Cham: Springer) 171–182. doi: 10.1007/978-3-319-11656-3_16

Vedaldi, A., and Lenc, K. (2015). "Matconvnet: convolutional neural networks for matlab," in *23rd ACM International Conference on Multimedia (ACMMM), 2015* (New York, NY: ACM), 689–692. doi: 10.1145/2733373.2807412

Weinland, D., and Boyer, E. (2008). "Action recognition using exemplar-based embedding," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008* (Anchorage: IEEE), 1–7. Available online at: http://ieeexplore.ieee.org/abstract/document/4587731/

Weinland, D., Ronfard, R., and Boyer, E. (2011). A survey of vision-based methods for action representation, segmentation and recognition. *Comput. Vis. Image Underst.* 115, 224–241. doi: 10.1016/j.cviu.2010.10.002

Yamato, J., Ohya, J., and Ishii, K. (1992). "Recognizing human action in time-sequential images using hidden markov model," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1992* (IEEE), 379–385. Available online at: http://ieeexplore.ieee.org/abstract/document/223161/

Zeiler, M. D., and Fergus, R. (2014). "Visualizing and understanding convolutional networks," in *13th European Conference on Computer Vision (ECCV),* eds D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars (Cham: Springer International Publishing), 818–833. Available online at: link.springer.com/chapter/10.1007/978-3-319-10590-1_53

Zhao, Z., and Elgammal, A. (2008). "Information theoretic key frame selection for action recognition," in *Proceedings of the British Machine Vision Conference (BMVC), 2008*, eds M. Everingham and C. Everingham (BMVA Press), 109.1–109.10. Available online at: http://www.bmva.org/bmvc/2008/papers/143.html

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Layher, Brosch and Neumann. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

[Layher et al., 2014b]

G. Layher, F. Schrodt, M. V. Butz, and H. Neumann (2014b). "Adaptive Learning in a Compartmental Model of Visual Cortex - How Feedback Enables Stable Category Learning and Refinement". In: *Frontiers in Psychology* 5.1287, p. 19. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2014.01287

©2014 G. Layher, F. Schrodt, M.V. Butz and H. Neumann.
This is an open-access article distributed under the terms of the Creative Commons Attribution License CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 05 December 2014
doi: 10.3389/fpsyg.2014.01287

Adaptive learning in a compartmental model of visual cortex—how feedback enables stable category learning and refinement

Georg Layher[1], Fabian Schrodt[2], Martin V. Butz[2] and Heiko Neumann[1]*

[1] Institute of Neural Information Processing, Ulm University, Ulm, Germany
[2] Department of Computer Science, University of Tübingen, Tübingen, Germany

Edited by:
Haluk Ogmen, University of Houston, USA

Reviewed by:
Oliver Obst, Commonwealth Scientific and Industrial Research Organisation, Australia
Ennio Mingolla, Northeastern University, USA

***Correspondence:**
Heiko Neumann, Institute of Neural Information Processing, Ulm University, James-Franck-Ring, Ulm 89069, Germany
e-mail: heiko.neumann@uni-ulm.de

The categorization of real world objects is often reflected in the similarity of their visual appearances. Such categories of objects do not necessarily form disjunct sets of objects, neither semantically nor visually. The relationship between categories can often be described in terms of a hierarchical structure. For instance, tigers and leopards build two separate mammalian categories, both of which are subcategories of the category Felidae. In the last decades, the unsupervised learning of categories of visual input stimuli has been addressed by numerous approaches in machine learning as well as in computational neuroscience. However, the question of what kind of mechanisms might be involved in the process of subcategory learning, or category refinement, remains a topic of active investigation. We propose a recurrent computational network architecture for the unsupervised learning of categorial and subcategorial visual input representations. During learning, the connection strengths of bottom-up weights from input to higher-level category representations are adapted according to the input activity distribution. In a similar manner, top-down weights learn to encode the characteristics of a specific stimulus category. Feedforward and feedback learning in combination realize an associative memory mechanism, enabling the selective top-down propagation of a category's feedback weight distribution. We suggest that the difference between the expected input encoded in the projective field of a category node and the current input pattern controls the amplification of feedforward-driven representations. Large enough differences trigger the recruitment of new representational resources and the establishment of additional (sub-) category representations. We demonstrate the temporal evolution of such learning and show how the proposed combination of an associative memory with a modulatory feedback integration successfully establishes category and subcategory representations.

Keywords: neural model, category learning, subcategory learning, unsupervised learning, feedforward and feedback processing

1. INTRODUCTION

Stimuli presented in isolation cause cortical responses by feeding a representation defined by the feature arrangement that is contained in the current scene. The strength of the response depends on its contrast but is influenced by the local context in which it is embedded. Such (local) context information is integrated and thus made available at a neural site via lateral intra-cortical interactions, preferentially through long-range associative interactions in the superficial layers of cortex (Self et al., 2012). Larger context is integrated through the hierarchical processing of inputs over several stages of the cortical hierarchy where feature specificity of the neurons becomes more and more specific, integrating over an increasingly more widespread space-feature domain (Markov and Kennedy, 2013). At earlier stages, the result of such feature integration is made available via top-down feedback to merge feature representations of higher levels with spatially more localized responses from initial filtering. Such convergence of feedforward and feedback streams of activation has recently been demonstrated to occur at the level of individual cortical columns (Mountcastle, 1997; Larkum, 2013).

Feedback signals tend to modulate the responses of activations at the earlier representations of raw feature presence (Larkum et al., 2004; Self et al., 2013). Modulating interactions are a common principle of neuronal interaction, which have been observed at different levels of cortical processing, subserving different cognitive computational functions, such as attention, figure-ground segregation, or grouping (Roelfsema et al., 2007; Poort et al., 2012). However, the precise functional role of feedback signals along downstream pathways is largely unclear and a topic of intense research investigation. Specific theoretical frameworks have been proposed that receive support by recent experimental investigations (Markov and Kennedy, 2013). One

such theoretical framework proposes that feedforward sensory activations are amplified by matching feedback such that those cells yield enhanced activations in a competition of cells, that have received a competitive advantage via modulating feedback (biased competition; Girard and Bullier, 1989; Desimone, 1998). Another framework considers the role of feedback as a predictive signal in which a template is activated that predicts the expected input given the evidence derived from current bottom-up input signals. The interaction of feedforward and feedback signals reduces the residual discrepancy between the different signal streams (Ullman, 1995; Rao and Ballard, 1999; Bastos et al., 2012). Overall, the literal difference between these model frameworks lies in the different roles feedback exerts on the bottom-up driven representations, although under certain conditions the two frameworks yield two variants of the same generic principles (Spratling, 2008, 2014).

In this work, we investigate learning and adaptation mechanisms in hierarchical cortical systems to develop a functional account for the role of feedback mechanisms. More specifically, we address the role hierarchical feedback may play in the online learning of visual representations. The study builds upon our previous modeling of a generic cortical architecture at the level of cortical columns. Model areas are defined by regular grids of interconnected columns, which are combined to define cortical subsystems, each composed of distributed networks of interconnected areas. Each model column is described at a mesoscopic level considering a compartmental structure that subdivides a cortical site into an input stage of specific signal filters, as well as superficial and deep layers as columnar compartments. Within this framework, feeding input signals drive the activity of columns and their lateral interactions. Feedback signals are thought to act in a modulating fashion so that responses at higher level cortical stages alone cannot generate activations in earlier representations (thus implementing a no-strong-loops principle; Crick and Koch, 1998). However, we demonstrate that interaction between different groups of cells allows to segregate the feedback signal strength that modulates the feedforward input activation such that the strength of feedback could be traced to serve as a signature how the expectations or predictions converge to the activation distribution of the driving input. The feature specificity of neurons in a cortical column is established through a learning mechanism that evaluates correlative activation in a scheme of modified Hebbian weight adaptation (Grossberg, 1988). During learning the connection strengths of bottom-up weights (to propagate converging driving input signals) are adapted. The applied learning scheme imposes a constraint such that the weights conserve their total energies so that variable input that is distributed over a population of neurons in columns does not lead to any bias in the incremental input segmentation. Thus, segmentations are allowed to build different and partly overlapping categorical patterns in which the total energy of the bottom-up input weights is normalized. The recurrent feedback from higher level representations generates a prediction, which consists of a pattern of the expected input activation, that drives the receiving representation of a column best. For that reason, the modulatory top-down feedback connections are here learned by using a slightly different weight adaptation mechanism. The feedback weights define a top-down projective field, which represents the expected average input activity distribution of the cell. Taken together, feedforward learning enables the generation of prototypical form pattern representations, whereas feedback weights encode the characteristics of the category a stimulus is currently assigned to by the visual system. Thus, feedback and feedforward learning in combination realize an online associative memory mechanism, allowing the separation of an input stimulus and an according prototypical representation (see Carpenter, 1989). Using a modulation mechanism, the differences between an input pattern and an internal category representation are amplified in the input signal, yielding category building, consolidation, or refinement. The framework thus defines an important building block for the automatic incremental learning of visual categories (at different stages in the visual hierarchy). The compartmental structure and the neuronal interactions allow to stabilize the learning to prevent oscillatory learning as well as effects of overshadowing existing representations, connoted as the plasticity-stability dilemma (Grossberg, 1988). Using simple form patterns as input stimuli, we demonstrate that the model allows to automatically distinguish and refine the encoding of overlapping patterns and to trigger the learning of new categories when the input patterns differ significantly.

2. GENERIC MODEL ARCHITECTURE

2.1. OVERVIEW OF THE MODEL COMPONENTS AND FUNCTIONAL ARCHITECTURE

The function of the proposed network architecture has been discussed in the previous section in order to motivate key aspects of automatic acquisition of shape and object representations and how underlying cortical structural principles and mechanisms might contribute to its realization. In this section we present formal model mechanisms as as a sketch of how the processing might be implemented dynamically. The basic structure of the generic model architecture is defined by three layers, each of which consisting of sheets of mutually interconnected computational elements (see **Figure 1**). These layers in the model roughly correspond to areas in cortex. Henceforth, we will address these stages by calling them layers or areas, given the particular context in the text. In the three layer architecture, the input layer is sketched like a simple replica of the input field fed by the current stimulus. The inclusion of such an explicit layer implicitly states that it may represent the result of some complex preprocessing that transforms the raw input into activity distributions referring to certain feature dimensions represented in a distributed fashion in (visual) cortex. As the same structure and composition of abstract columns can be replicated and more fine-tuned at different levels of cortex-like processing, we suggest that the outlined model architecture is generic in its structure and function. The computational elements in layers two and three both consist of an abstract model representation of cortical columns. Each of such columnar units itself is organized in a cascade of three processing stages: (I) input filtering, (II) activity modulation, and (III) pool normalization (details of the functional properties are discussed in, e.g., Neumann and Sepp, 1999; Bouecke et al., 2011; Brosch and Neumann, 2014a,b). These cascade stages roughly

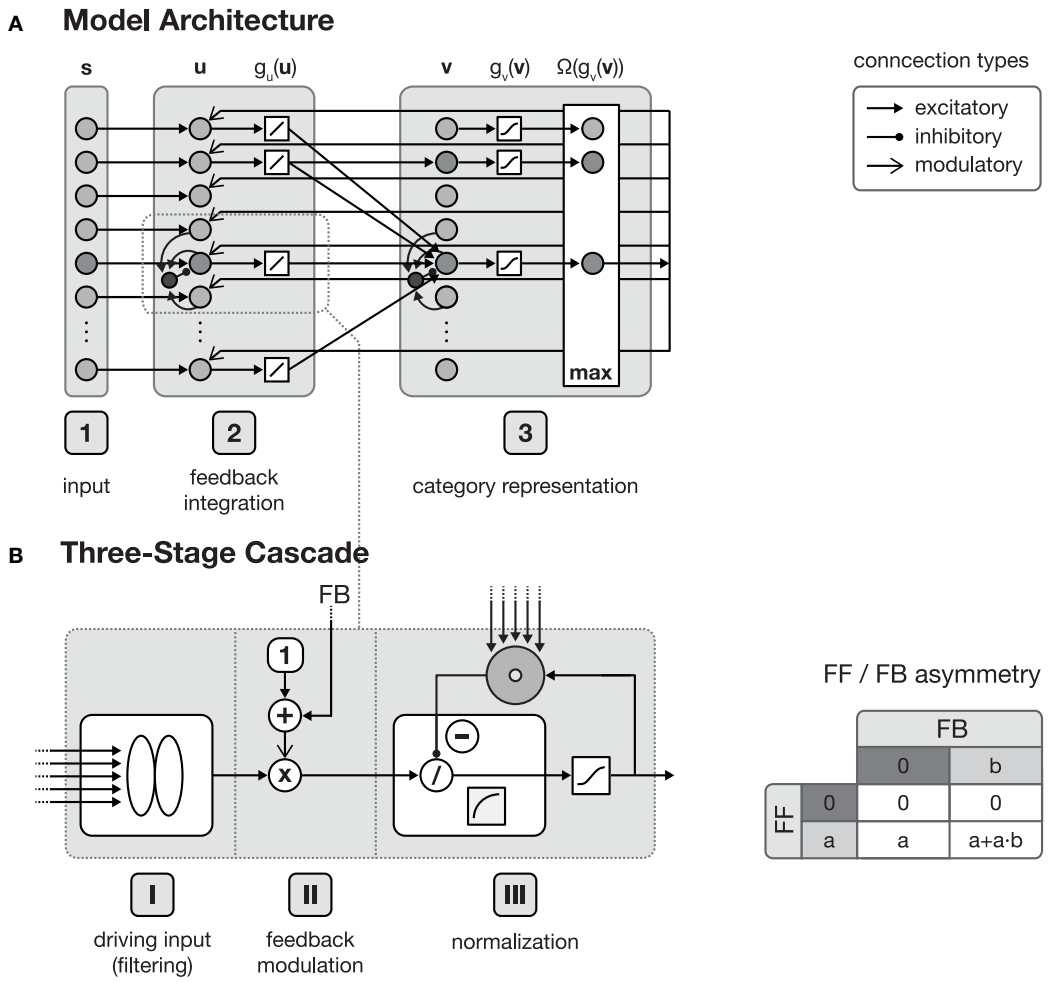


FIGURE 1 | Model architecture. (A) shows the overall structure of the proposed model, which is composed of interconnected cortical columns, subdivided into a compartmental structure of three processing layers. The first layer propagates input activities **s** to the second layer, where they are combined with a residual signal derived from feedback activities emitted from layer 3 and the current activity $g_U(\mathbf{u})$. After normalization, layer 3 category cells perform a correlation of the current layer 2 activities and their respective synaptic input weights. The cell with the strongest activation $g_V(\mathbf{v})$ is then selected by a winner-take-all (WTA) mechanism for weight adaptation and activity propagation. In principle all of the model layers consist of a three-stage processing cascade as illustrated in **(B)**. The cascade comprises an initial input filtering (stage **I**), the modulation of the activity (stage **II**) and a final pool normalization (stage **III**). Re-entrant feedback from higher level areas is incorporated in stage **II** where the current activity is modulated by $(1 + net_{FB})$. This kind of feedback integration is essential, since it results in an asymmetry of the roles the feedforward and the feedback signal play in the signal processing. As illustrated in the table on the right-hand side of **(B)**, without the presence of a feedforward signal, a feedback signal cannot evoke any activity.

correspond to the division of cortical areas, with their six layers (Lui et al., 2011), considering the layer of terminating bottom-up input, as well as the superficial and the deep layers of cortex (Self et al., 2013). Each of these stages is represented by a model neuron that itself is a single-compartment dynamic element with gradual activation dynamics representing the average potential of a group of mutually coupled neurons. A firing-rate function $g(\cdot)$ converts the potentials into an output activation. Feedforward and feedback signal streams are combined at the level of individual columns (Larkum, 2013; see Brosch and Neumann, 2014a for a model implementation). In the proposed architecture, the second layer combines the input multiplicatively with a residual signal that is derived from the current input pattern and a feedback signal emitted from the successive layer 3 which is biased by a tonic activity level (Eckhorn, 1999; Neumann and Sepp, 1999). Thus, the feedforward signal gates the re-entrant top-down signal so that the gain of existing activity can be increased by matching feedback signals. Feedback signals alone, however, cannot generate any activation for void bottom-up signal input. The feedback signal is generated here by a residual template, which contains the difference between the expected input (of the winning category node) and the current bottom-up input signal. As long as the difference does not vanish, the feedback mechanism leads to an increase in the activity gain of the current input. This mechanism deviates from the scheme described in e.g., (Bouecke et al., 2011), where the top-down signal is used instead of the residual signal. However, the dynamic properties of the non-linear circuit are retained.

Apart from the rather detailed network structure for generating an activation dynamic, the bidirectionally coupled network architecture is capable to adapt its connection weights, and is thus able to learn new category and subcategory representations as well as the expected average input distributions that have been established to drive a specific target category representation. In layer 3 of the generic architecture, category and subcategory representations are established using Hebbian learning mechanisms. Here, two complementary synaptic weight distributions are learned, each serving a different purpose within the proposed network. The feedforward synaptic weights are intended to build the category and subcategory representations during training, whereas the feedback weights are used to propagate an internal representation of the currently best matching category back to layer 2. This allows the estimation of the difference between the current input and the category assigned to the input after the feedforward sweep. Thus, layer 2 cells are able to combine the input with the derived difference signal and potentially evoke the activation of a different category/subcategory cell at the level of layer 3.

We split our presentation of the detailed model components into two major parts. First, we describe the activation dynamics, i.e., the formal definition of the generation of activities in each model computational element along the structure outlined in the previous paragraph. The activations are dependent on the input, the weightings of the spatial couplings for the input, and the current state, or activation of a model neuron. We emphasize how the incorporation of top-down feedback signal pathways can achieve rich and stable computations in such a network architecture. Second, in order to automatically acquire behaviorally relevant feature and category representations, the system can learn by adapting the weightings of the connection patterns between the model areas. We describe the weight, or learning, dynamics separately by focusing on the formal description of the weight adaptation and their key functionality. We finally link activation and learning dynamics to emphasize the capability of such building blocks for autonomous learning in cortical architectures.

In essence, category and subcategory learning is enabled using two complementary core mechanisms. First, an associative memory is realized through the combination of an instar with an outstar learning scheme (compare Carpenter, 1989; see **Figure 2**). This allows the assignment of a given input to the currently best matching internal representation, as well as the propagation of the corresponding feedback pattern to re-enter at an earlier processing layer. Second, the differences between an input signal and the pattern associated with the best matching internal representation of the input define the modulatory signal to enhance the gain of the bottom-up feedforward signal.

In the following, we first describe the overall properties of the three-stage processing cascade, which forms the generic building block for all of the model layers.

2.2. ACTIVATION DYNAMICS

2.2.1. Three-stage processing cascade

The first stage of the model cascade performs a linear filtering of the input. To model the response r of a cell, we calculate the weighted sum on the input to a cell, as defined by

$$r = \sum_{j=1}^{N} K_j \cdot s_j, \tag{1}$$

with N the number of input cells with activities $\mathbf{s}$, which are modulated by the weight distribution $\mathbf{K}$. Within the proposed model, the filtering step either results in the propagation of the impulse response to a given input (for layer 2 cells) or $\mathbf{K}$ corresponds to a weight distribution derived from the input statistics (for layer 3 cells, see Section 2.3.1).

At the second stage of the cascade, responses from the previous filtering are modulated by re-entrant input from higher-level model areas. Modulation is thereby performed in a way, such that

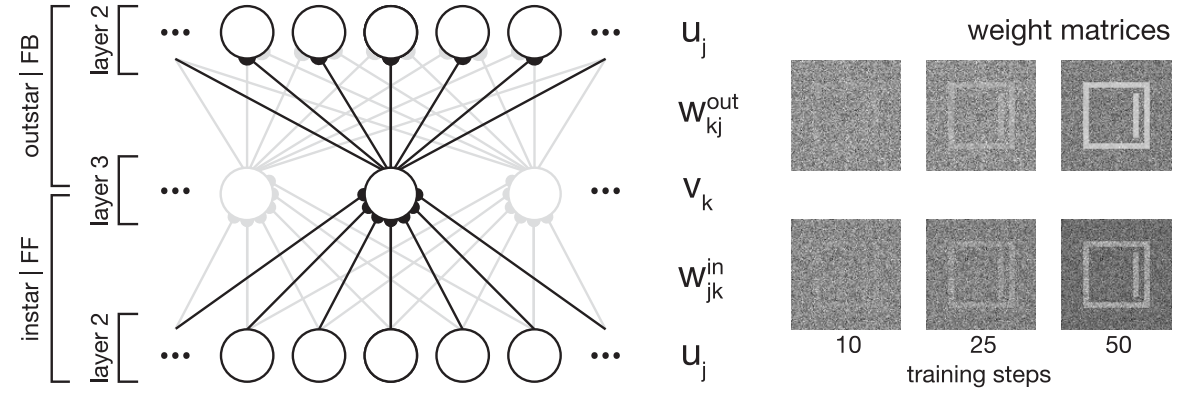


FIGURE 2 | Network plasticity. Model layer 3 cells establish categorial and subcategorial representations using Hebbian learning in combination with a modulatory feedback mechanism. As shown on the left-hand side, they realize an associative memory by combining instar and outstar learning schemes. The afferent connections weights w^{in} are used to select the best matching representation to a given input. The weights projecting away from the cell w^{out} are incorporated in the top-down feedback to layer 2 cells. For a given stimulus, only the cell with the highest activation is selected for weight adaptation. On the right-hand side, exemplary weight matrices are shown after several training steps. The matrices were obtained during the simulation of Experiment 1 (see Section 3.1).

only existing activities in an input signal can be amplified (and thus activities cannot emerge solely provoked by a feedback signal). With r being the unmodulated driving signal and net_{FB} being the strength of the feedback signal, the modulated response of a cell is given by

$$r_{FB} \propto r \cdot (1 + net_{FB}). \tag{2}$$

This kind of feedback incorporation assures that if $r = 0$ no signal is generated as output, independent of the strength of the feedback net_{FB}. On the other hand, the input signal r is left unchanged in the absence of any feedback signal (i.e., $net_{FB} = 0$, see **Figure 1B**).

Prior to the final stage of the processing cascade, we apply a transfer function to convert the responses into a cell activation level. For simplicity we employ a linear transfer function at layer 2 of the proposed model, whereas at layer 3, a non-linear sigmoidal transfer function is used.

At the final stage of the processing cascade, activity normalization through divisive mutual inhibition within a pool of neurons (shunting inhibition) is applied. In its dynamic formulation, the rate change of the a signal r_j^{norm} depends on the current activation level r_j and the amount of inhibitory input activation in the pool q_j

$$\dot{r}_j^{norm} = -\alpha_r \cdot r_j^{norm} + \beta_r \cdot r_j - r_j^{norm} \cdot q_j \tag{3}$$

$$\dot{q}_j = -q_j + \cdot \textstyle\sum_{k=1}^{M} r_k \cdot \Lambda_{jk}^{pool}, \tag{4}$$

with M denoting the size of the incorporated population in the neighborhood of location j and the weighting function Λ_{jk}^{pool}. The constant β_r controls the scale of the normalized signal, α_r denotes the passive decay rate.

In the following, we first describe the forward sweep throughout the proposed model layers. After the functional differences between the different model layers have been described in detail, we will emphasize the feedback connections and their role for the task of category and in particular subcategory learning.

2.2.2. Model layer 1/2

Layer 1 and layer 2 follow a pairwise connection scheme, such that each input cell in layer 1 is only connected to exactly one cell in layer 2 (see **Figure 1**). At the level of layer 2, the linear filtering step described in Equation (1) is equal to an identity function. Thus, the response of a layer 2 cell is defined by the following equation:

$$\dot{u}_j = -\alpha_u \cdot u_j + \beta_u \cdot s_j - u_j \cdot q_j, \tag{5}$$

where s_j denotes the output of a layer 1 cell, u_j describes the layer 2 cell response which relates to the membrane potential of real cells (j denoting the cell position). The constant α_u denotes the passive decay rate, whereas β_u describes the input scaling factor. The potentials are converted into an activation level, or firing rate, by the transfer function $g_u(u_j)$ (see Brosch and Neumann, 2014a for a formal specification and analysis). Here, we employ a linear transfer function with rectification such that no negative responses occur,

$$g_u(u_j) = \left[u_j\right]^+, \tag{6}$$

with $[u]^+ = max(u, 0)$. The competitive interaction against a pool of cells to accomplish activity normalization is defined as

$$\dot{q}_j = -q_j + \sum_{k=1}^{N} g_u(u_k) \cdot \Lambda_{jk}^{pool}, \tag{7}$$

with N denoting the size of the incorporated population in the neighborhood of location j, weighted by Λ_{jk}^{pool}. Without the incorporation of any feedback signals, layer 2 cells solely perform an activity normalization on the output activities $\mathbf{s}$ of layer 1 and propagate the result to layer 3.

2.2.3. Model layer 3

Layer 2 and layer 3 cells form a complete bipartite connection graph with connections in both directions (see **Figure 1**), with corresponding synaptic coupling strengths w^{in} for feedforward and w^{out} for feedback connections. The output of layer 2 $g_u(\mathbf{u})$ is filtered by the feedforward weights w_{ji}^{in} to generate the strength of the response v_i of a layer 3 cell, which finally enters a competition with the surrounding pool activation ($\mathbf{u}$ denoting the field of input activities represented as a vector), as defined by:

$$\dot{v}_i = -\alpha_v \cdot v_i + \beta_v \cdot \sum_{j=1}^{N} g_u(u_j) \cdot w_{ji}^{in} - v_i \cdot q_i, \tag{8}$$

with the passive decay rate α_v and the input scaling factor β_v. The response is then converted into an activity level using the non-linear sigmoidal transfer function g_v with the parameters κ_{log} (steepness) and μ_{log} (mean response level),

$$g_v(v_i) = \frac{1}{1 + e^{\kappa_{\log} \cdot (\mu_{\log} - v_i)}}. \tag{9}$$

As in layer 2, the final competition for activity normalization is defined by a non-linear competition of target activity and the integrated activation over a pool of neurons, which is determined by

$$\dot{q}_i = -q_i + \sum_{k=1}^{M} g_v(v_k) \cdot \Lambda_{ik}^{pool}, \tag{10}$$

with M denoting the number of cells in layer 3 and the weighting function Λ_{ik}^{pool}.

2.3. NETWORK PLASTICITY

In the previous part we have briefly introduced the formal description that covers the activation dynamics of the model mechanisms in the suggested generic architecture. As already mentioned, the architecture consists of three layers that roughly correspond to model areas of visual cortex. As outlined in **Figure 1**, the first area represents the input, that can be the raw responses of preprocessing the input directly (like in the early stages of the visual hierarchy, e.g., V1 and V2) or the output responses from a cascade of already more sophisticated processing to build intermediate level representations (like in the

higher stages of the visual hierarchy, e.g., V3 and V4). The second and third model areas in the model layout are connected bidirectionally representing feedforward and feedback sweeps of signal propagation in cortex (Lamme and Roelfsema, 2000). We have already explained how the two counterstream signal flows converge to build representations of integrated bottom-up evidences (from signal processing) and top-down predictions or expectations (generated by higher level stages of category representations). In this part we equip the network architecture with mechanisms of adapting the connections to learn representations in specific input weights. We suggest here that learning occurs along the feedforward as well as the feedback pathways (an outline of the learning architecture is shown in **Figure 2**). The functionality behind such a, again generic, principle is that feedforward connections learn weighting profiles that increase the probability for an input activation pattern to generate amplified responses in the recipient unit. Likewise, learning of feedback connections is intended to build up a representation in which source node activations (at the higher-level stage of the architecture) will generate a distribution of (pre-) activations as the expected average activity at the input stage that drives the node. The expectation is thus represented in the top-down connection weights (see Layher et al., 2014 for a model learning architecture that follows the same generic principles). Here, we develop a mechanism with a slightly different emphasis. The network aims to develop categories and also (later) to advance the automatic establishment of subcategories driven by significant local deviations of the already existing category representation. Therefore, the signal that is carried by the top-down feedback connections needs to be transformed into a residual signal such that the difference from the expected activation pattern is registered. We suggest that such residual patterns are generated at the neuronal *activation* pattern, instead of the weighting pattern.

In the following, we present the formal descriptions of the mechanisms used for the weight adaptation. We also briefly sketch how these relate to achieve the target representations for the desired bottom-up and top-down processing. The adaptation of the connection weights, for both feedforward and feedback, can be considered for individual neuronal sites in layer 3: The *receptive* field, or fan-in structure, is defined for connections along the bottom-up signal transmission that converge on a target neuron, $u \rightarrow v$. The *projective* field, or fan-out structure, on the other hand, is defined for connections along the reverse direction that spread out from the target neuron back to the previous stage, $v \rightarrow u$ (compare Carpenter and Grossberg, 1987b; Lehky and Sejnowski, 1988 for discussions of the underlying function of such connection principles). The activity dependent adaptation rules of such connection weights, namely feedforward, w^{in} and feedback w^{out} weights, are governed by modified versions of Hebbian correlation learning principles (Hebb, 1949). These modifications lead to stability and proven convergence properties and it can be shown that the learning rules optimize some target functionals.

The target neurons at layer 3 (with the adaptable fan-in and fan-out connections) are considered here to represent categories in a classification or recognition mechanism. For simplicity, we consider learning by weight adaptation that is allowed only for the category node that is maximally activated, as in many other related learning paradigms (e.g., Kohonen, 1982; Carpenter and Grossberg, 2003). Such a model neuron is selected by a simple maximum selection operation, or winner-take-all (WTA) mechanism (Grossberg, 1973) and the weight adaptation is triggered subsequently,

$$\Omega(g_v(v_k)) = \begin{cases} 1 & \text{if } k = \underset{i=1\ldots M}{arg\,max}\, g_v(v_i) \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

It should be noted that the WTA selection is chosen here for simplicity. As an alternative, one could use a softmax mechanism as well (e.g., Roelfsema and van Ooyen, 2005), without changing the overall functionality of the approach. The specific learning rules for feedforward and feedback connections are presented below.

The learning of the feedforward weights w^{in}, as well as the feedback weights w^{out} is realized using Hebbian learning principles, which are described in the following.

2.3.1. Learning of feedforward connections

We utilize a variant of Hebbian correlation learning which prevents the changes of connection weights to grow without bounds. The stabilization is here achieved by a forgetting term that reduces the weight proportionally to the postsynaptic cell activation. The weight change for the receptive fields is formally defined by

$$\dot{w}^{in}_{jk} = \Omega(g_v(v_k)) \cdot \eta_{in} \cdot g_v(v_k) \cdot (g_u(u_j) - g_v(v_k) \cdot w^{in}_{jk}). \tag{12}$$

The r.h.s. of the equation is defined by the switch $\Omega(\cdot)$ to enable/disable neurons for adaptation of their weights and a learning rate η_{in}. The extended Hebbian correlation term is defined by $g_v(v_k) \cdot \left(g_u(u_j) - g_v(v_k) \cdot w^{in}_{jk}\right)$. In other words, the learning is gated by the activation of the postsynaptic neuron. Here, the Hebbian term $g_u(u_j) \cdot g_v(v_k)$ is combined with the forgetting term $g_v(v_k)^2 \cdot w^{in}_{jk}$ to balance the temporal change and bound the growth of the cell's synaptic input weights. It has been demonstrated that such a learning mechanism extracts the first Eigenvector of the input distribution (Oja, 1982, 1992). Another property of the Oja learning rule is of even more interest here: The learning of the bottom-up feedforward weights approaches a fan-in connection pattern in which the weight energy is conserved (Dayan and Abbott, 2005). The fan-in weight vector $\mathbf{w}^{in}_k$ is adapted over time to reach equilibrium, such that $\lim_{t\rightarrow\infty} \dot{\mathbf{w}}^{in}_k = v_k \cdot \mathbf{u} - \gamma v_k^2 \cdot \mathbf{w}^{in}_k = 0$ (with γ as a positive constant value that scales the balancing component). The equilibrium weight energy is then

$$\|\mathbf{w}^{in}_k\|^2 = \frac{1}{\gamma}. \tag{13}$$

Assuming $\gamma = 1$ we get a unit length for the input weights to single category nodes. This, in turn, prevents input activation distributions to bias the output activity at the category representation, given that the input activity distribution is normalized as well. The latter property is achieved by the normalization stage of the pool interaction defined in the activations dynamics of the network stages above.

2.3.2. Learning feedback connections

Again, we utilize a stabilized Hebbian weight adaptation formalism. In its dynamic formulation, the weight changes for projective fields is formally defined by

$$\dot{w}_{kj}^{out} = \Omega(g_v(v_k)) \cdot \eta_{out} \cdot g_v(v_k) \cdot (g_u(u_j) - w_{jk}^{out}). \quad (14)$$

As for the adaptation of the receptive field, or fan-in, weights (Equation 12) we utilize the switch $\Omega(\cdot)$ to enable/disable weight adaptation and a learning rate η_{out} for the projective, or fan-out, weights. The extended Hebbian term is here defined by $g_v(v_k) \cdot \left(g_u(u_j) - w_{jk}^{out}\right)$. The learning is gated by the activation of the neuron that represents the category, which is presynaptic to the projective field considering the representation generated for the top-down feedback connections. Unlike the learning rule discussed in Equation (12), the forgetting term to balance the temporal change is controlled by the weight only. Such a weight adaptation mechanism defined in Equation (14) has been suggested for gated steepest descent learning in long-term memory formation, e.g., in Adaptive Resonance, or ART networks (Grossberg, 2013b). The adaptation of the fan-out weight vector $\mathbf{w}_k^{out}$ over time reaches equilibrium, such that $\lim_{t\to\infty} \dot{\mathbf{w}}_k^{out} = v_k \cdot \mathbf{u} - \gamma v_k \cdot \mathbf{w}_k^{out} = 0$ (with γ as a positive constant value that scales the balancing component). The equilibrium weight energy is then

$$\mathbf{w}_k^{out} = \frac{1}{\gamma}\mathbf{u}. \quad (15)$$

Assuming $\gamma = 1$ we achieve a projective field, or fan-out, pattern for the connection weights corresponding to the (average) expected input activation represented in $\mathbf{u}$. Activation of a category node, thus, biases the receiving postsynaptic model neurons according to the predicted pattern the category expects to receive for its best tuning input. Feedback learning may also utilize the learning rule of Oja as for learning the feedforward connections described above. In this case the weight distribution of the projective field would converge to the first Eigenvalue of the expected input, instead of its mean. We have tested this and observed similar network performance. The latter implementation argues in favor of symmetric learning mechanisms for bottom-up and top-down connection weights. We decided to use a version in which the feedback projections approach the expected average input activation that represents the tuning of the individual categories, as in Equation (15).

2.4. FEEDBACK FOR SUBCATEGORY LEARNING

The mechanisms presented so far contributed to the feedforward as well as a generic feedback sweep of the model. The feedback sketched so far generically considered the modulatory influence a feedback signal has on any feedforward input representation. The mechanism emphasized the symmetry breaking property in which bottom-up signals gate the activity generation (at stage 2 of the processing cascade described in Section 2.2.1) which can be selectively amplified by the presence of matching feedback signals. Here, without incorporating the feedback from layer 3, the learning rules defined in Section 2.3 would successfully learn representations of input categories, but without the potential of further refining them on a subcategorial level. As stated earlier, the feedback allows the estimation of the difference between the current input and the category assigned to the input after the feedforward sweep. Thus, layer 2 cells are able to combine the input with the derived difference signal. If the difference and the modulation strength after the feedback sweep is large enough, learning is potentially triggered such that an associated new subcategory is built using a so far unused layer 3 cell. The enhancement of the layer 2 responses by modulating feedback changes (Equation 5) to

$$\dot{u}_j = -\alpha_u \cdot u_j + \beta_u \cdot s_j \cdot (1 + \lambda \cdot res_j^{templ}) - u_j \cdot q_j, \quad (16)$$

where res_j^{templ} denotes the residual signal derived from the feedback net_j^{FB} of the best matching category cell [selected by $\Omega(g_v(v_k))$] and the current activity $g_u(u_j)$. λ is controlling the influence of res_j^{templ} on u_j and thus is crucial for the extent of the difference between a modulated input and a category assigned in the feedforward sweep. The residual signal res_j^{templ} is defined by

$$\begin{aligned} res_j^{templ} &= \left[g_u(u_j) - net_j^{FB}\right]^+ \\ &= \left[g_u(u_j) - \Omega(g_v(v_k)) \cdot w_{kj}^{out}\right]^+, \end{aligned} \quad (17)$$

with $[x]^+ = \max(0, x)$ denoting a rectification operation limiting res_j^{templ} to positive values. A closer look at the presented model dynamics may help us to reveal the potential roles that feedback plays in the context of category learning. According to Equation (17), the feedback signal acts as a predictive coding scheme, since net_j^{FB} expresses what the model expects how an input of a given category looks like on average. On the other hand, the expression $s_j \cdot \left(1 + \lambda \cdot res_j^{templ}\right)$ in Equation (16) realizes a biased competition mechanism, favoring input components, which are in accordance with the residual signal res_j^{templ}. In essence, this kind of feedback incorporation results in an amplification of the differences between the currently best matching internal representation and the input. During learning, the difference between a category representation and individual instances of the category increases with the number of stimuli of the same category. If the effect of this difference on the input is large enough, a new subcategory representation is established.

3. RESULTS

In the following, we demonstrate the capabilities of the proposed model in learning category and subcategory representations using two categories of artificial input stimuli. As shown in **Figure 3**, category **A** contains four variations of a pictographical face. Category **B** is composed of four squares inclosing an either vertically or horizontally oriented bar at different positions. Without the loss of generality, we used very simplified stimuli to keep the computational complexity and in particular the necessary preprocessing steps as simple as possible. This allows us to keep the focus strictly on the role which feedback might play in the task

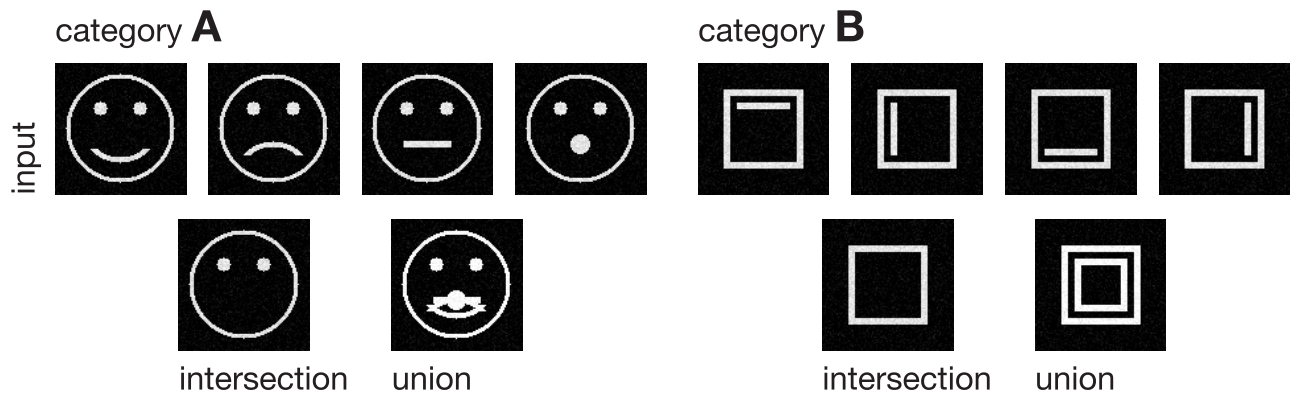


FIGURE 3 | Input stimuli. Two different categories of stimuli were used as input to the model, each with a set of subcategories. Category **A** is composed of four pictographical images of a face, only differing in the shape of their mouth. Category **B** consists of four variations of a square inclosing an either vertically or horizontally oriented bar at different positions. The bottom row shows the intra-category union and intersection for both categories, pointing out the differences and similarities in each category.

of category and subcategory learning. The stimuli were generated with the dimensions of $100 \times 100\ px$ with intensity values ranging from 0 to 1. The number of input units in layer 1 thus is always $100 \cdot 100 = 10000$ units. As mentioned in Section 2, the cells in layer 1 and those in layer 2 follow a pairwise connection scheme, so that layer 2 consists of the equivalent number of 10000 units. The number of layer 3 cells differs from experiment to experiment. Note that in all experiments, there remained at least one unused layer 3 cell after training, which was never selected for weight adaptation. Thus, the number of units in layer 3 never was a limitation to the establishment of a new category or subcategory representation. During training, Gaussian noise with mean $\mu = 0$ and a standard deviation of $\sigma = 0.05$ was added on each of the input stimuli, with values clipped to the range of [0, 1]. If not stated otherwise, we used learning rates of $\mu_{out} = \mu_{in} = 2^{-4}$ and a feedback gain factor of $\lambda = 2^5$. These values were found to be a suitable balance between the learning speed and the influence of the feedback. The parameters of the logistic function as defined in Equation (9) were set to $\mu_{log} = 700$ and $\kappa_{log} = 0.0075$, such that the transfer function results in a mean activation level of $g_v(700) = 0.5$ when roughly half of the input energy of one of the used stimuli is present in the input signal. The weights w^{in} and w^{out} of the category cells at model layer 3 were initialized with random values drawn from a normal distribution with mean $\mu = 0.75$ and a standard deviation of $\sigma = 0.1$, allowing empty category cells at layer 3 to be activated by just a small number of active input cells.

For the ease of computational complexity, we simulate the dynamics described in Section 2 using the corresponding steady-state equations. An in depth analysis of the activation dynamics can be found in Brosch and Neumann (2014a). Within the simulations, one (training) step—or iteration—corresponds to the presentation of one input stimulus, consisting of one feedforward and one feedback sweep through the model. Activities of the layer 3 cells are evaluated after the feedforward and after the feedback sweep and both trigger the adaptation of a categorial and/or subcategorial representation.

In total, we performed four experiments, each highlighting on a different aspect of the proposed model and learning mechanisms. In the first experiment, we show in principle how the model successfully learns a representation of a category of visual input stimuli and decomposes the category into subcategories. The second experiment is intended to demonstrate the invariance of the proposed learning mechanism to the order in which the stimuli are presented. Experiment 3 focuses on the importance of the feedback signal for the task of subcategory learning by contrasting Experiment 2 with a nearly identical experimental setup. The sole difference to Experiment 2 is that the incorporation of feedback is suppressed by setting the feedback gain parameter λ to $\lambda = 0$. In the last experiment we demonstrate how the model generalizes across the number of categories present in the input data and show how it successfully establishes representations for two categories of visual input and their subcategories.

All simulations were carried out using Mathworks Matlab R2014a.

3.1. EXPERIMENT 1

We trained the proposed model using the rectangular stimuli of category **B** as shown in **Figure 3**. The stimuli were presented in epochs of four blocks of sorted stimuli, each block containing 100 instances of one of the four rectangle variations. At model layer 3, six cells were used during the training. To slow down the weight adaptation process and highlight on the establishment of new subcategory representations, we used a learning rate of $\mu_{out} = \mu_{in} = 2^{-5}$, set μ_{log} to 800 and initialized w^{in} and w^{out} with random values drawn from a normal distribution with $\mu = 0.5$ and $\sigma = 0.1$. The activities of the layer 3 cells after the feedforward and the feedback sweep are shown in **Figure 4** along with the corresponding weights w^{in} and w^{out} after several training steps. Over the first training steps, the model develops a combined representation of the first and the second rectangular shape containing information about the surrounding rectangle, as well as portions of information about the interior of the two shapes. After 200 training steps, the effect of the learning mechanism starts to be twofold. After the feedforward sweep, the overall category

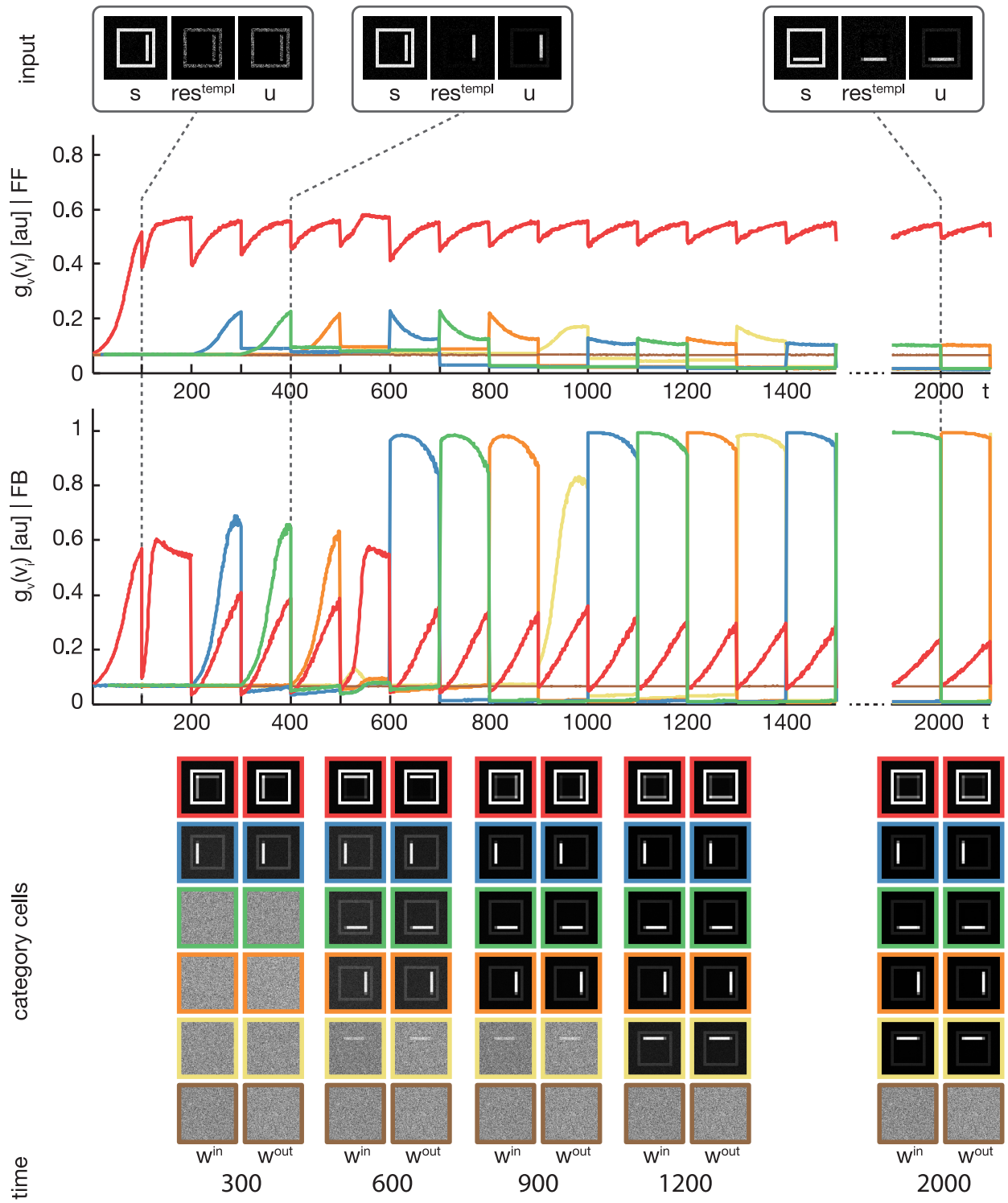


FIGURE 4 | Experiment 1. The model was trained using four rectangular shapes (see **Figure 3**, category **B**) as input stimuli. Stimuli were presented in sorted blocks. Six category cells (color-coded) were initialized with random weights. The first row shows exemplary input configurations **s**, along with the corresponding residual signal $\mathbf{res}^{\mathbf{templ}}$ and the input signal **u** after feedback modulation. In the second row, the activities of the six category cells before the feedback sweep are shown. As can be seen, before the feedback is effective on the input, only one cell (encoded in red) responds to all input configurations. This cell represents the overall category cell. The second row shows the activities after the feedback sweep. In the last row the corresponding category cell weights are displayed framed by colors according to the activity plots. It can be seen, that in the beginning all inputs are learned into one category cell. After about 200 training steps, the effect of the feedback is high enough to trigger the learning of a new subcategory representation. This process repeats several times, until each subcategory is represented by an own category cell.

representation is adapted to the current input stimulus. On the contrary, after the feedback sweep a subcategorial representation is learned by recruiting an additional layer 3 cell. The effect of the feedback signal now is large enough to suppress the outer rectangular shape and highlight on the differences between the overall category representation and the current input stimulus. This process continues until all of the four input variations are represented in an own subcategory cell. After learning, the feedforward sweep always results in a high activation level $g_v(v_i)$ of the overall category cell that represents the generic shape (refer to the second row of **Figure 4**). After the feedback sweep, however, the subcategory cell representing the specifics of the particular input stimulus is the one with the highest activation level.

3.2. EXPERIMENT 2

In the second experiment, the proposed model was trained using the pictographical faces of category **A** (see **Figure 3**) as input. Stimuli now were presented in random order. As in Experiment 1, six category cells at model layer 3 were used. All training parameters were set to their default values (see Section 3). **Figure 5** shows how category and subcategory cell representations are learned during the simulation. Again, the residual signal res^{templ} increases with the distinctiveness of the already established category representation and thus the effect of the feedback signal increases. Already after 21 training examples, the difference between the current input and the existing category cell is high enough to yield a modulation of the input effective-enough to evoke the establishment of a new subcategory. This process repeats several times, since after 127 learning iterations all of the variations of category **A** are represented in an own subcategory cell. Altogether, the model successfully learns category and subcategory representations, even though the stimuli are presented in random order.

3.3. EXPERIMENT 3

In a third experiment we conducted a simulation equivalent to the one in the second experiment, but now with disabling the feedback signal by setting $\lambda = 0$ (see **Figure 6**). As expected, without the feedback signal no subcategory representations are established and just one overall category representation is learned.

3.4. EXPERIMENT 4

For the last experiment we used both categories **A** and **B** shown in **Figure 3** as input stimuli. The parameters were equivalent to those described in Experiment 1 but now twelve category cells at layer 3 were initialized. Since the differences between the two types of stimuli (circular and rectangular) are already large enough before the feedback takes place, the model establishes two overall category representations and successively builds subcategories to these two categories. **Figure 7** shows the weights of the established two category, as well as the respective four subcategory cells after 1000 learning steps.

4. DISCUSSION

In this work we proposed a hierarchical architecture of cortical feedforward and feedback processing that builds upon previous work on the modeling of recurrent cortical dynamics (Neumann et al., 2007; Brosch and Neumann, 2014a). Here, we particularly focused on the issue how in such networks feature or category representations could be automatically acquired by unsupervised learning mechanisms, which are seamlessly integrated in the recurrent architecture. The core computational elements assumed are cortical model columns that are abstractly described by a three-stage cascade of processing steps. The same elements have been utilized as generic mechanisms in models of form and motion processing, figure-ground segregation, as well as modeling biological motion perception that fuses segregated form and motion pathways (Neumann and Sepp, 1999; Bayerl and Neumann, 2004; Raudies et al., 2011; Layher et al., 2014). As a specific model feature, we have emphasized the role of feedback that modulates feedforward driving inputs such that their gain is increased dependent on the degree of correlation between feedforward and feedback signal activation. In conjunction with subsequent pool normalization the modulatory feedback sweeps realize a way of biased competition (Girard and Bullier, 1989; Desimone, 1998; Roelfsema et al., 2002; Reynolds and Heeger, 2009). The model now incorporates learning mechanisms to automatically build feature/category representations that are generated by the connection weights through adaptation.[1] Such learning allows to build representations that adapt their specificity to the statistics of the sensory input patterns.

4.1. SUMMARY OF CONTRIBUTIONS

The main contributions of the work presented in the paper are twofold. First, the investigated learning mechanisms occur in the feedforward as well as in the feedback connections. These are driven by bottom-up sensory input and top-down feedback signals to re-enter processing at earlier stages. The latter contain context information that allows to embed local sensory input signals into a larger behavioral context and predictions generated thereof. All this is in the spirit of multi-layer learning networks as discussed in Hinton (2007). In that sense feedforward connections will learn the specific configuration of an (average) appearance of an input feature pattern that the learned category is selectively tuned to. Considering static shape and form input the underlying structural principles are based on the cortical architecture of the ventral pathway with mutual interactions between such distributed representations in different cortical areas (Markov et al., 2014). The feedback connections, on the other hand, also learn by adjusting their weights in order to improve the predicted input pattern that maximally excites the feature/category representation. Second, the top-down feedback learning mechanism combines the modulatory feedback (Girard and Bullier, 1989; De Pasquale and Sherman, 2013) with the concept of top-down predictors that tend to minimize the residual

[1] We make the distinction here between feature and category representation in order to emphasize the different locality of representations that are established at different layers in a hierarchical network architecture. With increasing integration sizes of cells at different levels more information from previous stages is integrated. The zones of lateral integration are more localized at earlier stages, thus, we refer to the learning of feature representations. At later stages the convergence zones may range over the full spatial input domain and, therefore, the representations already cover categories that could be shape or motion related.

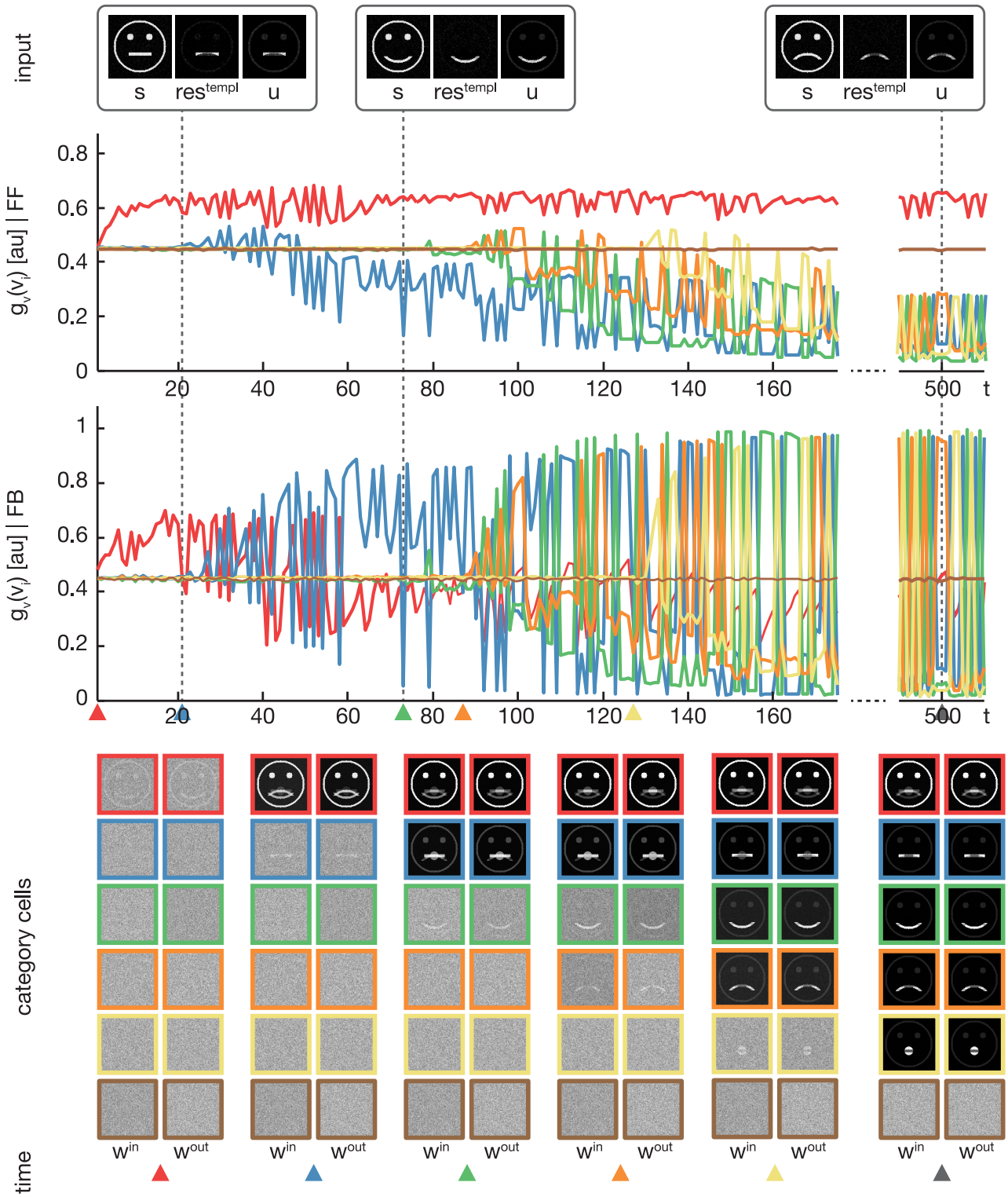


FIGURE 5 | Experiment 2. In the second experiment, we trained the model using four variations of pictographical faces (see **Figure 3**, category **A**). In contrast to Experiment 1, stimuli were presented in random order. The display of the results is organized as in **Figure 4**. In addition, colored triangles indicate the points in time, when a category or subcategory cell was selected for weight adaptation the first time. Although the stimuli were presented in random order, the model successfully separates the input stimuli into subcategories.

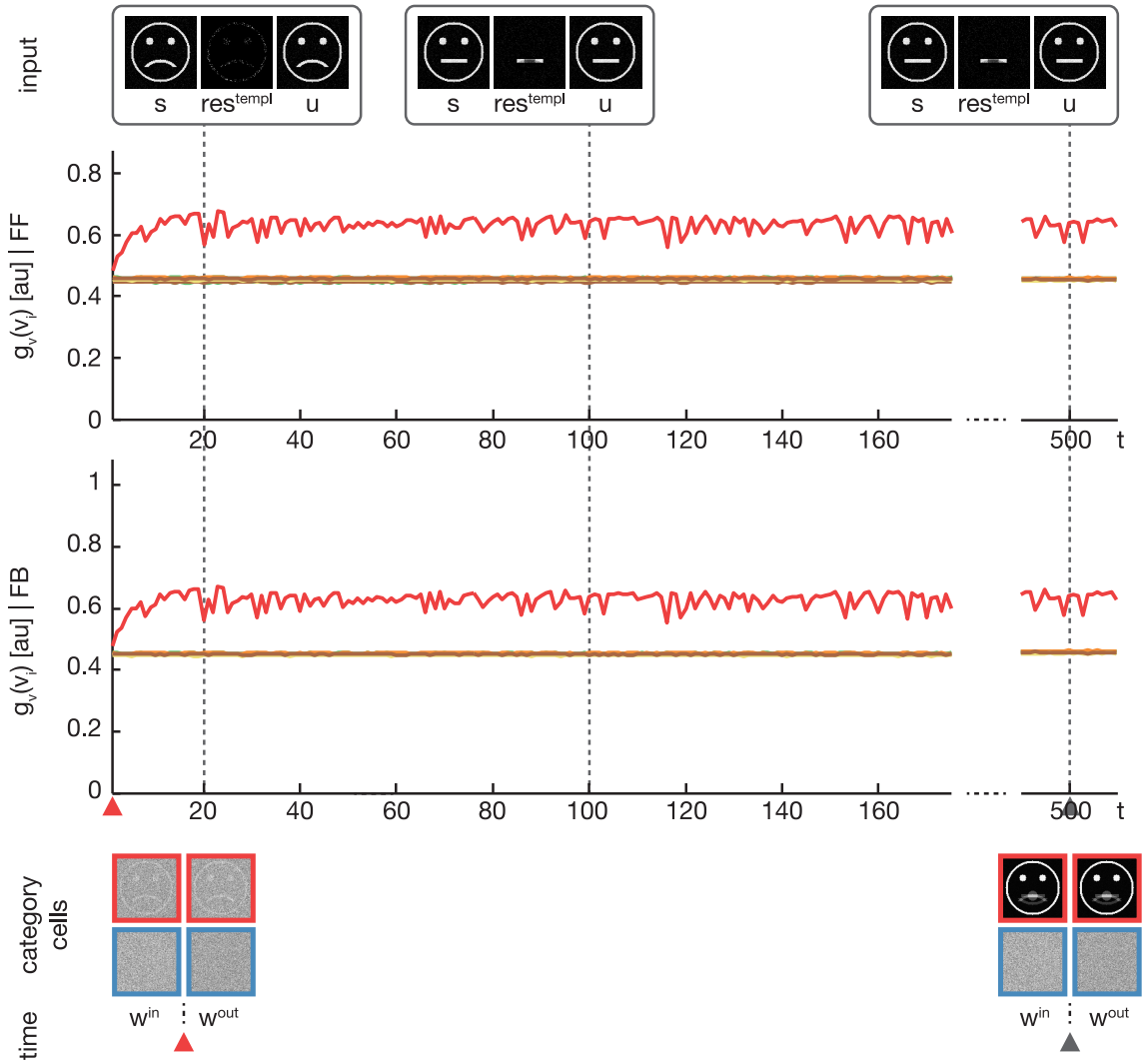


FIGURE 6 | Experiment 3. The simulation was performed with the same setup as in Experiment 2. Only the feedback gain parameter λ was adjusted to $\lambda = 0$, disabling any influence of the feedback on the learning. As expected, the activation levels $g_v(v_i)$ before (second row) and after (third row) the feedback sweep are identical. Without the feedback modulation, no subcategory representations are learned and only one overall category representation is established.

error between feedforward sensory signals and the top-down pattern (Rao and Ballard, 1999; Bastos et al., 2012). The idea behind this concept is that weights will be increased when the predicted pattern and the current input differ. The amount of this gain increase depends on the residual difference between these two patterns. The model defines the basis for more principled investigations how cortical sub-networks that are involved in different tasks might be established. In own previous work (Layher et al., 2014), distributed representations of spatio-temporal patterns in the cortical form and motion pathway were learned for articulated or biological motion perception (Johansson, 1973; Giese and Poggio, 2003). Here, sequence-selective representations were established by learning representations of convergent feedforward responses from form and motion representations. Also top-down weights are learned in which the projective field reaches the two separate pathways of form and motion. The principles proposed in this work now allow to further develop the understanding of how such complex distributed representations can be learned and how average categories are learned together with subcategories for components that deviated significantly from the average category representation.

4.2. RELATION TO PREVIOUS MODELS OF CORTICAL LEARNING OF REPRESENTATIONS

Learning of feedforward networks has been investigated intensively before. Most importantly, the connection weights in multilayer networks have been trained by using backpropagation to minimize the residual error of expected output given a specific input pattern (Lehky and Sejnowski, 1988, 1990; LeCun et al., 1989). Such approaches require a teacher signal that

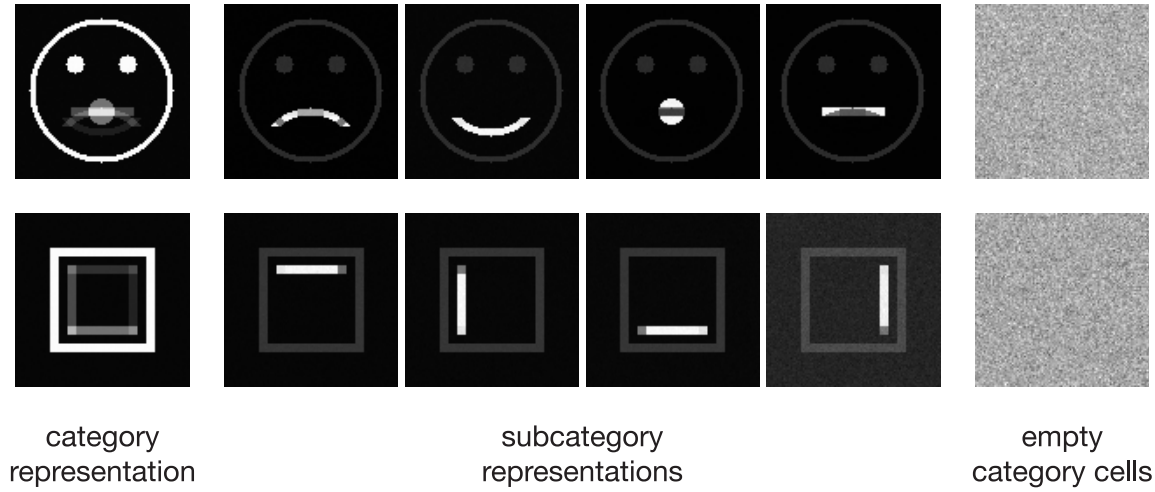


FIGURE 7 | Experiment 4. Both stimulus categories shown in **Figure 3** were used as input to the model. The distributions of the input weights w^{in} are shown after 1000 training steps for twelve category cells at model layer 3. The model successfully established representations for all input variations of both classes and two overall category representations, one for each input category.

determines the desired target output. The assumption of a supervisor involved in each teaching trial is biologically unrealistic in general. For that reason, a mechanism that is based on reinforcement learning (Sutton and Barto, 1981; Doya, 2007) has been suggested that combines an unspecific global reward-based reinforcement signal with an attentional signal that is backpropagated from the output layer to allow weight adaptation at those units that have been involved in the stimulus-response mapping in the previous processing of the input signals (Roelfsema and van Ooyen, 2005). Also, learning in hierarchical multistage architectures for object recognition has been investigated. Approaches range from random sampling of the input pattern space (Riesenhuber and Poggio, 1999; Serre et al., 2007; Mutch and Lowe, 2008; Serre and Poggio, 2010) to clustering techniques to arrive at sparse representations of the input via additional constraints on the connection weight patterns (Aharon et al., 2006) or auto-encoding that minimizes the reconstruction error of the input (LeCun et al., 1998). Recently, learning in multi-layer networks, so-called deep hierarchical networks (Bengio, 2009), has received renewed interest to build networks with high classification rate performance (LeCun et al., 1990; Hinton et al., 2006). Representations in such networks are learned in a sequential manner by learning the connection weight between pairs of layers, starting from the initial sensory-related level. Once learning converges, the next level connection weights are learned. This procedures is recurrently applied until all connections have been determined. The learning mechanisms are based on gradient descent type, for example, realizing stage-by-stage backpropagation learning. Unlike these proposals, the network mechanism here incorporates bidirectional learning of weights along the feedforward as well as the feedback path. The weight adaptation is based on variants of Hebbian correlation learning. These variants stabilize the growth properties of the input and output weight vectors to the computational elements (model columns) in the architecture. As a consequence, the representations built in the connection patterns have specific interpretations: Along the feedforward path we assume an Oja learning scheme (Oja, 1982, 1992). As a result, the fan-in (or receptive field) weight energy of the total input connections from the previous layer neurons tends to be normalized for feedforward signal filtering. This ensures that different input patterns balance their input weights such they enter any subsequent competition or selection step in an unbiased fashion. Concerning feedback learning, connection weight patterns along the recurrent projection (corresponding to the projective field of a feature or category, Lehky and Sejnowski, 1990) approach the average expected input. In other words, the driving category representation generates a prediction pattern that covers the expected input activation that tends to match the tuning of the representation (Grossberg, 1980).

The proposed architecture is influenced by the conception of adaptive resonance theory (or ART; Grossberg, 1980, 1987; Carpenter, 1989). In a nutshell, learning in ART is organized in stages of feedforward and a feedback sweep processing. During feedforward processing the input signal is weighted by the connection pattern, or filter, between nodes in the feature representation and the category layer. These weightings are initialized by some random values. One category will gain a maximal input from the feature representation activated though the input signal, similar to the feedforward sweep in other networks (Rumelhart and Zipser, 1985), and also in the model proposed in this paper. Similarly, the self-organization of feature maps has also been approached by means of connection weight adaptation in hierarchically organized networks, establishing competitive processes for automatic map formation (von der Malsburg, 1973; Kohonen, 1982). The category that is maximally activated will subsequently suppress all other category representations by recurrent lateral center-surround competition. With supra-linear firing-rate functions such a competitive stage leads to a winner-take-all strategy (Grossberg, 1973). The weightings along the feedforward path can be adjusted to approach the (average) signal features. The

feedback connections fed by the winning category node (the projective field) are then allowed to adapt their weights as well so that they approach the input activation distribution. In other words, the feedback connections learn the input that maximally drives the currently activated category node to maintain a match between the input and the expectation the category has about its input patterns it is tuned to (resonance condition). If, instead, any momentary input feature pattern maximally drives a category with a top-down expectation pattern that does *not* match the input, then a mismatch occurs and the combined bottom-up and top-down expectation patterns annihilate. In order to now select another existing category or recruit a new category item, a reset wave is triggered that instantaneously shuts off the winning category that was activated maximally but has a mismatching representation in its projection field. This allows the top-down weights of a newly selected category to adjust in order to now better match the input that is coherent with the expected pattern represented by the active category representation (for recent comprehensive summaries and overviews of the ART principle, see Grossberg, 2013a,b). Discrete implementations of ART networks for pattern recognition have been described for binary as well as continuous input pattern representations (Carpenter and Grossberg, 1987a,b). A more specific reference to possible biophysical mechanisms underlying the recurrent interaction and learning has been described in Carpenter and Grossberg (1990), while Molenaar and Raijmakers (1997) presented a continuous time network implementation.

Several other network architectures use feedback connections that can be adapted through a learning process, e.g., (Elman, 1990; Hinton et al., 2006; Hinton, 2007; Lazar et al., 2009; Rolfe and LeCun, 2013). While Elman (1990) maps temporal feature history into an explicit representation through recurrences, a more recent approach by Lazar et al. (2009) utilizes a reservoir of connected neurons in a large pool to learn representations of temporal patterns. A read-out mechanism maps the internal state trajectories onto units through reduction of state-space dimension and clustering of activities. This recurrent network architecture with spiking model neurons emphasizes different mechanisms in the learning of connections weights, namely a simplified version of spike-timing dependent plasticity (STDP; Gerstner et al., 1996; Bi and Poo, 2001; Caporale and Dan, 2008) as unsupervised weight adaptation mechanism connecting excitatory cells in the pool, a synaptic scaling mechanisms through weight normalization, and an intrinsic plasticity mechanism for firing threshold adaptation. Our approach makes use of similar mechanism in the learning procedure. Here, we are concerned with networks of gradual activation dynamics, which motivates utilizing standard Hebbian correlation learning instead of the STDP rule. Weight normalization occurs implicitly in our adaptation mechanisms by utilizing modified Hebbian learning. In particular, as discussed in Section 2.3, the bottom-up learning of receptive field weights for individual category nodes approaches a weight energy (Equations 12 and 13). The intrinsic plasticity in our scheme is accomplished through the normalization activations, or firing rates, by the pool of cells in a neighborhood defined in the space-feature domain (compare Equations 5 and 6, Brosch and Neumann, 2014a). The model of Rolfe and LeCun (2013) stresses the importance of acquisition of representations of categories and subcategories, like in our model. Their network realizes properties of deep networks establishing sparse representations of subcategories, like auto-encoder networks using binary state neuronal elements (Hinton and Salakhutdinov, 2006), and recurrently combine (hidden) representations and their predictions (Hinton et al., 2006) (see Hinton, 2007 for a review). Synaptic scaling (see a recent review in Tetzlaff et al., 2012) is addressed here from the perspective of how the receptive and projective fields learn a particular target activity distribution. In the architecture proposed by Rolfe and LeCun (2013) two types of units emerge that define parts and categories. The time course of the serial learning mechanism suggests that the network first establishes component representations mainly driven by the input. Later and with a slow learning efficacy, categories emerge that combine those units that belong to the category (while those they do not belong to are inhibited). Our proposed network architecture shares the idea of building hierarchical object representations. The acquisition of categories and subcategory, or part, representations operates oppositely: Categories are established as new representations recruiting free capacities from the long-term memory node reservoir in model layer 3 when the current input is significantly dissimilar in comparison to already existing categories. The deviations from a larger category then lead to learning subcategories and these are linked to their category representation by the temporal signature of the activation. Thus, the proposed model may start with only coarse-grained category knowledge, which is subsequently refined when more detailed information is available during the course of interacting with the environment.

While in these approaches the feedback connections serve to incorporate activations over time, feedback in ART architectures is intended to solve the stability-plasticity dilemma. The latter summarizes the necessity that an adaptive system needs to acquire or adapt to new evidence (or knowledge) and, at the same time, to keep those previously acquired representations stable (to prevent catastrophic forgetting). Our proposal differs from these previous model developments in several respects. In our architecture we build upon an abstract though biophysically plausible model of processing in cortical columns. The interaction between signal activations in bottom-up and top-down sweeps is based on modulatory feedback that enhances those sensory signal activation patterns which match the top-down template of activation that is re-entered at earlier stages of processing along the hierarchy. Thus, instead of a similarity calculation between signal patterns, a biologically plausible gain adjustment is assumed (Sherman and Guillery, 1998). The modulation signal we use for the amplification of the input signals is calculated by the difference between the current input signal and the top-down expectation pattern. This effectively combines the key mechanisms underlying the two current main theories of the role of feedback in cortex: top-down modulation and biased competition is assumed for the enhancement of the input gain. Here, the modulation strength is controlled by the difference between bottom-up and top-down signal, or the residual between these two activation patterns. Steering the amount of weight adaptation by the difference between signal and expectation template incorporates the flavor of predictive coding approaches (Rao and Ballard, 1999; Rauss

et al., 2011; Bastos et al., 2012). The logic behind this strategy is that the relative enhancement is reduced monotonically the more the top-down prediction signal approaches the bottom-up signal. As a consequence, the update of the weights will more quickly converge since both, feedforward and feedback, signal remain approximately constant and the weighting pattern approaches the prediction template. Consequently, no external reset mechanism is required that explicitly detects a mismatch discrepancies by a threshold vigilance parameter, as in ART models. In our proposal, the feedback modulatory dynamics and the learning mechanisms automatically tune the average matching activation of the responding category and also select the category or feature representation. Furthermore, and potentially of even more interest is the automatic establishment of categorical representations that capture the average of the input patterns that can drive the corresponding nodes in the columnar architecture. At the same time, subcategory representations are established that represent the significant differences in the detailed feature configurations that differ from the average case. This has been demonstrated in example cases (Section 3) in which, for example, faces are distinguished from non-faces at the categorical level. Smiling facial appearances or faces where the eyes are closed are then also automatically assigned to the average category by learning. However, to distinguish the appearance differences new subcategories are automatically established and learned. This selectivity is realized by two core mechanisms. First, the realization of an associative memory through the combination of an instar with an outstar learning scheme (see Carpenter, 1989), which allows the assignment of a given input to the currently best matching internal representation, as well as the corresponding feedback pattern. Second, the modulatory amplification of the differences between an input signal and the feedback pattern associated with the best matching internal representation of the input. If the amplification after the feedback sweep is effective enough, the correlation between the modulated input and an empty category cell will be higher than to the category representation the input was assigned to in the feedforward sweep. Thus, learning will be triggered for the so far unused category cell and a new subcategory will be built.

The computational mechanisms of activation and weight dynamics support principles that have been predicted to minimize the computational efforts of visual systems to successfully deal with the complexity problem of perception (Tsotsos, 1988, 2005). The hierarchical organization of representations in model areas, the receptive field properties of model columns, the hierarchical pooling of spatially separated input representations, and the top-down feedback together with unsupervised learning are structural principles that enable the visual system to successfully cope with complex input stimuli that are behaviorally relevant. The presented model is able to build the underlying distributed representations at low, intermediate, and higher levels in the cortical hierarchy by means of key cortical principles.

4.3. FEEDBACK—MODULATORS AND PREDICTORS

The hierarchical model architecture proposed here is composed of multiple model areas each of which is represented by a three-stage columnar cascade model. The cascade consists of input filtering, activity modulation of filter outputs by re-entrant signals, and competitive center-surround interaction of target cells against a pool of cells. The latter stage yields an activity normalization for generating net output responses. Together with the gain enhancement generated by input modulation via re-entrant signals the network interactions achieve a biased competition response characteristics (Desimone, 1998; Reynolds and Heeger, 2009; Carandini and Heeger, 2012). The proposed architecture can be interpreted as an abstracted compartment representation of the layered architecture of cortical areas (Self et al., 2012). The interplay between the normalization of activities and the selective enhancement of activities via feedback establishes the dynamics of cortical processing. Activity normalization at the output stage is computed by a mechanism of shunting inhibition, like the non-linear divisive mechanisms proposed in Carandini and Heeger (1994); Carandini et al. (1999); Kouh and Poggio (2008); Carandini and Heeger (2012) (see Brosch and Neumann, 2014a for a formal analysis of the computational properties). Feedback signals generated at higher-level cortical stages or parallel processing pathways provide context information that is re-entered at the current stage of the processing hierarchy (Grossberg, 1980; Edelman, 1993). While the presence of feedback connections is a well-established principle of cortical signal processing and integration, the exact role of how such feedback signals are re-entered at the earlier stages is a controversial topic of ongoing investigation. We adopt here two principles from the two major frameworks of the functionality of feedback, namely modulatory feedback to bias subsequent competitive mechanisms and predictive coding.

How feedback signals interact and combine with signals delivered in the driving feedforward stream is yet unresolved. Two major conceptual ideas have been developed, each receiving support by experimental evidence (Markov and Kennedy, 2013). In a nutshell, *biased competition* suggests that signals in the feedforward pathway are enhanced by top-down templates (represented by activity distributions) such that they receive a competitive advantage in subsequent mutually competitive processes. As a result, feature responses that receive feedback have a higher gain which, in turn, leads to stronger suppression of activities that were not enhanced (Girard and Bullier, 1989; Desimone, 1998; Roelfsema et al., 2002; Reynolds and Heeger, 2009). In *predictive coding* the goal of computation is to reduce the residual error between the feedforward signal and the (top-down) templates generated at a stage that generates an expectation about the most compatible input. This idea is based upon predictor-corrector mechanisms in optimization (Ullman, 1995; Rao and Ballard, 1999; Bastos et al., 2012). As a consequence the state trajectory of such systems and their activations are different: While in biased competition the activations of the representations that match the predictions will increase, they will decrease in the predictive coding framework. Interestingly, Spratling (2008) has shown that these two approaches are functionally equivalent when the feedback in the biased competition is additive. Here, we utilize multiplicative feedback based on the linking mechanism suggested by Eckhorn et al. (1990); Eckhorn (1999) to account for activity synchronization in networks of spiking neurons and

further evidence that signal amplifications occur at the level of cortical pyramidal cells (Larkum, 2013) (see a model description in Brosch and Neumann, 2014b that accounts for these findings). An influential paper by Crick and Koch (1998) provided strong support for modulatory top-down connections based on theoretical grounds. In the model framework proposed here we adopt the framework of modulatory feedback (thus, biased competition). The feedback signals represent context-sensitive templates and are gated by feedforward driving input signals. In such a modulating feedback driven gain control mechanism spatial detail is generated by feature-driven low-level processes and representations and subsequently associated with coarse-grained context information which is provided by intermediate and higher-levels of cortical computation (Lamme and Roelfsema, 2000; Roelfsema et al., 2002; Roelfsema, 2006). In order to control the weight adaptation for learning, the strength of feedback is calculated by the difference between the feedforward signal and the predictive template that is delivered along the top-down connections. Such a difference represents the residual between the two counter stream representations (Ullman, 1995). In a nutshell, the idea is that the amount of feedback is regulated by the deviation between the two convergent streams (like in predictive coding). The re-entrant combination is, however, based on multiplicative gain enhancement. The strength of the excitatory feedback will vanish when the input is perfectly predicted by the top-down template. In that case, the feedforward signal representation will not be further enhanced. In Bastos et al. (2012) the cortical circuits are present in different compartments of a cortical area (compare Self et al., 2013 for a discussion of the possible roles of input layer and superficial and deep layer compartments in cortical area V1). Our suggested mechanism can be realized assuming subtractive interaction between driving feedforward cells and feedback signals, potentially in the superficial layer compartment. The resulting residual activations can then activate cells in columns via the apical dendrites of pyramidal cells (located either in the superficial or deep layer compartments; Larkum et al., 2004). In Brosch and Neumann (2014b) a firing-rate model of pyramidal cell interaction has been developed that explains such interactions at the level of the columnar architecture adopted here. All these feedforward and feedback interactions combine with learning mechanisms for the feedforward and the feedback connections. The equations supposed to define the weight changes lead to stable convergent weight changes. In the feedforward connection pattern the fan-in, or receptive field, weights to a unit approach a defined weight energy, or length, of the connection coefficients. This is desirable since after a representation accomplished in the weights has been settled, the activation level is not biased by the weights but is determined by the signal input and its changed gain through feedback interaction. In the feedback connection pattern the fan-out, or projective field, weights from a unit approach the (average) activity the representation is tuned to. Thus, the expected input is represented which can be activated as top-down template to instantiate the expected input signal or feature configuration. This leads to resonances in cases where the top-down expectation is retrieved from already established knowledge. In cases of mismatches new feature/category representations can be automatically recruited to establish new knowledge in the learning cortical architecture.

4.4. MODEL LIMITATIONS AND EXTENSIONS

The proposed model architecture emphasized the computational role of feedforward and feedback mechanisms in order to generate interactive states, or resonances, in a hierarchically organized model system. The re-entrant feedback is assumed to be modulatory such that bottom-up feedforward signals gate the recurrent feedback activations. The interactive processing is combined with a learning mechanism that allows to adjust connection weights along the feedforward as well as the feedback pathways. We have demonstrated the general functionality by using simple shapes that are kept under full control during the design process. Also we employed only a pair of interacting cortical model areas, each composed as a sheet of columnar units with lateral interactions. In addition, a separate input layer that represents the stimulus was incorporated. The proposed model architecture may be investigated along several lines of questions.

In its current form, the proposed model architecture separately evaluates the activities of layer 3 category and subcategory cells before and after the modulation of the residual feedback on the input signal. This results in an activity pattern in which only an overall category cell *or* a subcategory cell can be active at a time. It would be interesting to integrate an additional mechanism which prevents such fluctuations and keeps both the overall category and the subcategory cell active in parallel.

Deep hierarchies have been proposed to accomplish the build-up of rich composite feature representations at different stages of hierarchically organized networks for solving detection and recognition tasks (LeCun et al., 1998; Hinton, 2007; Bengio, 2009). A natural extension of the simplified architecture studied in this paper is to add further model cortical areas and train the feedforward and feedback connection weights at each level. We expect that such an extended architecture allows the construction of multi-level representations of pattern compositions over several stages in a hierarchy. Such an approach should provide the generic structure to automatically build representations of fragments of input stimuli in which recognition is combined with segmenting inputs using the learned top-down templates (Ullman et al., 2002; Ullman, 2007).

The proposed scheme currently utilizes simple input patterns to build categories and associated subcategories to make explicit the variations that deviate from the average category representations. It would be interesting to study the responses for more realistic shape patterns presented as gray level inputs that provide the input to the network architecture. Also in this case, it would be interesting to study the multi-level steps necessary for the proposed model cortical architecture to accomplish the category learning under even more realistic input representations. In a technical instance of processing Borenstein and Ullman (2008) proposed an image segmentation scheme based on bottom-up signal driven processing that is combined with top-down processing to utilize knowledge for improved segmentation. Although the focus there is mainly on the improvement of image processing, the approach might serve as an inspiration for modeling as well. We suggest that the potential power of the network

architecture proposed in this work lies in the automatic learning of templates for feedback expectation (at low and intermediate levels of representation; Hinton, 2007) that could be evaluated in terms of their information content for visual classification tasks (Ullman et al., 2002).

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: Georg Layher, Heiko Neumann; Performed the experiments: Georg Layher; Analyzed the data: Georg Layher, Fabian Schrodt, Martin V. Butz, Heiko Neumann; Model implementation: Georg Layher; Wrote the paper: Georg Layher, Fabian Schrodt, Martin V. Butz, Heiko Neumann.

ACKNOWLEDGMENTS

Georg Layher and Heiko Neumann have been supported by the Transregional Collaborative Research Centre SFB/TRR 62 "A Companion Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

REFERENCES

Aharon, M., Elad, M., and Bruckstein, A. (2006). K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.* 54, 4311–4322. doi: 10.1109/TSP.2006.881199

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron* 76, 695–711. doi: 10.1016/j.neuron.2012.10.038

Bayerl, P., and Neumann, H. (2004). Disambiguating visual motion through contextual feedback modulation. *Neural Comput.* 16, 2041–2066. doi: 10.1162/0899766041732404

Bengio, Y. (2009). Learning deep architectures for ai. *Found. Trends Mach. Learn.* 2, 1–127. doi: 10.1561/2200000006

Bi, G.-Q., and Poo, M.-M. (2001). Synaptic modification by correlated activity: Hebb's postulate revisited. *Annu. Rev. Neurosci.* 24, 139–166. doi: 10.1146/annurev.neuro.24.1.139

Borenstein, E., and Ullman, S. (2008). Combined top-down/bottom-up segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 30, 2109–2125. doi: 10.1109/TPAMI.2007.70840

Bouecke, J. D., Tlapale, E., Kornprobst, P., and Neumann, H. (2011). Neural mechanisms of motion detection, integration, and segregation: from biology to artificial image processing systems. *EURASIP J. Adv. Signal Process.* 2011, 6:1–6:22. doi: 10.1155/2011/781561

Brosch, T., and Neumann, H. (2014a). Computing with a canonical neural circuits model with pool normalization and modulating feedback. *Neural Comput.* 26, 2735–2789. doi: 10.1162/NECO_a_00682

Brosch, T., and Neumann, H. (2014b). Interaction of feedforward and feedback streams in visual cortex in a firing-rate model of columnar computations. *Neural Netw.* 54, 11–16. doi: 10.1016/j.neunet.2014.02.005

Caporale, N., and Dan, Y. (2008). Spike timing-dependent plasticity: a hebbian learning rule. *Annu. Rev. Neurosci.* 31, 25–46. doi: 10.1146/annurev.neuro.31.060407.125639

Carandini, M., and Heeger, D. J. (1994). Summation and division by neurons in primate visual cortex. *Science* 264, 1333–1336. doi: 10.1126/science.8191289

Carandini, M., and Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* 13, 51–62. doi: 10.1038/nrn3136

Carandini, M., Heeger, D. J., and Anthony Movshon, J. (1999). "Linearity and gain control in v1 simple cells," in *Models of Cortical Circuits*, Volume 13 of *Cerebral Cortex*, eds P. S. Ulinski, E. G. Jones, and A. Peters (New York, NY: Springer US, Kluwer Academic/ Plenum Publishers), 401–443. doi: 10.1007/978-1-4615-4903-1_7

Carpenter, G. A. (1989). Neural network models for pattern recognition and associative memory. *Neural Netw.* 2, 243–257. doi: 10.1016/0893-6080(89)90035-X

Carpenter, G. A., and Grossberg, S. (1987a). ART 2: self-organization of stable category recognition codes for analog input patterns. *Appl. Opt.* 26, 4919–4930. doi: 10.1364/AO.26.004919

Carpenter, G. A., and Grossberg, S. (1987b). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Comput. Vision Graph. Image Process.* 37, 54–115. doi: 10.1016/S0734-189X(87)80014-2

Carpenter, G. A., and Grossberg, S. (1990). {ART} 3: hierarchical search using chemical transmitters in self-organizing pattern recognition architectures. *Neural Netw.* 3, 129–152. doi: 10.1016/0893-6080(90)90085-Y

Carpenter, G. A., and Grossberg, S. (2003). "Adaptive resonance theory," in *The Handbook of Brain Theory and Neural Networks, Second Edition*, ed M. A. Arbib (Cambridge, MA: MIT Press), 87–90.

Crick, F., and Koch, C. (1998). Constraints on cortical and thalamic projections: the no-strong-loops hypothesis. *Nature* 391, 245–250. doi: 10.1038/34584

Dayan, P., and Abbott, L. F. (2005). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, MA: The MIT Press.

De Pasquale, R., and Sherman, S. M. (2013). A modulatory effect of the feedback from higher visual areas to V1 in the mouse. *J. Neurophysiol.* 109, 2618–2631. doi: 10.1152/jn.01083.2012

Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 353, 1245–1255. doi: 10.1098/rstb.1998.0280

Doya, K. (2007). Reinforcement learning: computational theory and biological mechanisms. *HFSP J.* 1, 30–40. doi: 10.2976/1.2732246/10.2976/1

Eckhorn, R. (1999). Neural mechanisms of visual feature binding investigated with microelectrodes and models. *Visual Cogn.* 6, 231–265. doi: 10.1080/135062899394975

Eckhorn, R., Reitboeck, H. J., Arndt, M., and Dicke, P. (1990). Feature linking via synchronization among distributed assemblies: simulations of results from cat visual cortex. *Neural Comput.* 2, 293–307. doi: 10.1162/neco.1990.2.3.293

Edelman, G. M. (1993). Neural darwinism: selection and reentrant signaling in higher brain function. *Neuron* 10, 115–125. doi: 10.1016/0896-6273(93)90304-A

Elman, J. L. (1990). Finding structure in time. *Cogn. Sci.* 14, 179–211. doi: 10.1207/s15516709cog1402_1

Gerstner, W., Kempter, R., van Hemmen, J. L., and Wagner, H. (1996). A neuronal learning rule for sub-millisecond temporal coding. *Nature* 383, 76–78. doi: 10.1038/383076a0

Giese, M., and Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nat. Rev. Neurosci.* 4, 179–192. doi: 10.1038/nrn1057

Girard, P., and Bullier, J. (1989). Visual activity in area V2 during reversible inactivation of area 17 in the macaque monkey. *J. Neurophysiol.* 62, 1287–1302.

Grossberg, S. (1973). Contour enhancement, short-term memory, and constancies in reverberating neural networks. *Stud. Appl. Math.* 52, 213–257.

Grossberg, S. (1980). How does a brain build a cognitive code. *Psychol. Rev.* 87, 1–51. doi: 10.1037/0033-295X.87.1.1

Grossberg, S. (1987). Competitive learning: from interactive activation to adaptive resonance. *Cogn. Sci.* 11, 23–63. doi: 10.1111/j.1551-6708.1987.tb00862.x

Grossberg, S. (1988). Nonlinear neural networks: principles, mechanisms, and architectures. *Neural Netw.* 1, 17–61. doi: 10.1016/0893-6080(88)90021-4

Grossberg, S. (2013a). Adaptive resonance theory. *Scholarpedia* 8, 1569. doi: 10.4249/scholarpedia.1569

Grossberg, S. (2013b). Adaptive resonance theory: how a brain learns to consciously attend, learn, and recognize a changing world. *Neural Netw.* 37, 1–47. doi: 10.1016/j.neunet.2012.09.017

Hebb, D. O. (1949). *The Organization of Behavior*. New York, NY: Wiley.

Hinton, G., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.* 18, 1527–1554. doi: 10.1162/neco.2006.18.7.1527

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends Cogn. Sci.* 11, 428–434. doi: 10.1016/j.tics.2007.09.004

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Percept. Psychophys.* 14, 201–211. doi: 10.3758/BF03212378

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biol. Cybernet.* 43, 59–69. doi: 10.1007/BF00337288

Kouh, M., and Poggio, T. (2008). A canonical neural circuit for cortical nonlinear operations. *Neural Comput.* 20, 1427–1451. doi: 10.1162/neco.2008.02-07-466

Lamme, V. A., and Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* 23, 571–579. doi: 10.1016/S0166-2236(00)01657-X

Larkum, M. (2013). A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. *Trends Neurosci.* 36, 141–151. doi: 10.1016/j.tins.2012.11.006

Larkum, M. E., Senn, W., and Lüscher, H. R. (2004). Top-down dendritic input increases the gain of layer 5 pyramidal neurons. *Cereb. Cortex* 14, 1059–1070. doi: 10.1093/cercor/bhh065

Layher, G., Giese, M. A., and Neumann, H. (2014). Learning representations of animated motion sequencesa neural model. *Top. Cogn. Sci.* 6, 170–182. doi: 10.1111/tops.12075

Lazar, A., Pipa, G., and Triesch, J. (2009). Sorn: a self-organizing recurrent neural network. *Front. Comput. Neurosci.* 3:23. doi: 10.3389/neuro.10.023.2009

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1, 541–551. doi: 10.1162/neco.1989.1.4.541

LeCun, Y., Boser, B., Denker, J. S., Howard, R. E., Habbard, W., Jackel, L. D., et al. (1990). "Advances in neural information processing systems 2," in *Chapter Handwritten Digit Recognition with a Back-propagation Network*, ed D. S. Touretzky (San Francisco, CA: Morgan Kaufmann Publishers Inc.), 396–404.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Lehky, S. R., and Sejnowski, T. J. (1988). Network model of shape-from-shading: neural function arises from both receptive and projective fields. *Nature* 333, 452–454. doi: 10.1038/333452a0

Lehky, S. R., and Sejnowski, T. J. (1990). Neural network model of visual cortex for determining surface curvature from images of shaded surfaces. *Proc. R. Soc. Lond. B Biol. Sci.* 240, 251–278. doi: 10.1098/rspb.1990.0037

Lui, J. H., Hansen, D. V., and Kriegstein, A. R. (2011). Development and evolution of the human neocortex. *Cell* 146, 18–36. doi: 10.1016/j.cell.2011.06.030

Markov, N. T., and Kennedy, H. (2013). The importance of being hierarchical. *Curr. Opin. Neurobiol.* 23, 187–194. doi: 10.1016/j.conb.2012.12.008

Markov, N. T., Vezoli, J., Chameau, P., Falchier, A., Quilodran, R., Huissoud, C., et al. (2014). Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. *J. Comp. Neurol.* 522, 225–259. doi: 10.1002/cne.23458

Molenaar, P. C., and Raijmakers, M. E. (1997). Exact ART: a complete implementation of an ART network. *Neural Netw.* 10, 649–669. doi: 10.1016/S0893-6080(96)00111-6

Mountcastle, V. B. (1997). The columnar organization of the neocortex. *Brain* 120(Pt 4), 701–722. doi: 10.1093/brain/120.4.701

Mutch, J., and Lowe, D. G. (2008). Object class recognition and localization using sparse features with limited receptive fields. *Int. J. Comput. Vision* 80, 45–57. doi: 10.1007/s11263-007-0118-0

Neumann, H., and Sepp, W. (1999). Recurrent V1-V2 interaction in early visual boundary processing. *Biol. Cybern.* 81, 425–444. doi: 10.1007/s004220050573

Neumann, H., Yazdanbakhsh, A., and Mingolla, E. (2007). Seeing surfaces: the brain's vision of the world. *Phys. Life Rev.* 4, 189–222. doi: 10.1016/j.plrev.2007.09.001

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *J. Math. Biol.* 15, 267–273. doi: 10.1007/BF00275687

Oja, E. (1992). Principal components, minor components, and linear neural networks. *Neural Netw.* 5, 927–935. doi: 10.1016/S0893-6080(05)80089-9

Poort, J., Raudies, F., Wannig, A., Lamme, V. A., Neumann, H., and Roelfsema, P. R. (2012). The role of attention in figure-ground segregation in areas V1 and V4 of the visual cortex. *Neuron* 75, 143–156. doi: 10.1016/j.neuron.2012.04.032

Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Raudies, F., Mingolla, E., and Neumann, H. (2011). A model of motion transparency processing with local center-surround interactions and feedback. *Neural Comput.* 23, 2868–2914. doi: 10.1162/NECO_a_00193

Rauss, K., Schwartz, S., and Pourtois, G. (2011). Top-down effects on early visual processing in humans: a predictive coding framework. *Neurosci. Biobehav. Rev.* 35, 1237–1253. doi: 10.1016/j.neubiorev.2010.12.011

Reynolds, J. H., and Heeger, D. J. (2009). The normalization model of attention. *Neuron* 61, 168–185. doi: 10.1016/j.neuron.2009.01.002

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025. doi: 10.1038/14819

Roelfsema, P. R. (2006). Cortical algorithms for perceptual grouping. *Annu. Rev. Neurosci.* 29, 203–227. doi: 10.1146/annurev.neuro.29.051605.112939

Roelfsema, P. R., Lamme, V. A. F., Spekreijse, H., and Bosch, H. (2002). Figure–ground segregation in a recurrent network architecture. *J. Cogn. Neurosci.* 14, 525–537. doi: 10.1162/08989290260045756

Roelfsema, P. R., Tolboom, M., and Khayat, P. S. (2007). Different processing phases for features, figures, and selective attention in the primary visual cortex. *Neuron* 56, 785–792. doi: 10.1016/j.neuron.2007.10.006

Roelfsema, P. R., and van Ooyen, A. (2005). Attention-gated reinforcement learning of internal representations for classification. *Neural Comput.* 17, 2176–2214. doi: 10.1162/0899766054615699

Rolfe, J. T., and LeCun, Y. (2013). Discriminative recurrent sparse auto-encoders. arXiv:1301.3775. Available online at: http://adsabs.harvard.edu/abs/2013arXiv1301.3775R

Rumelhart, D. E., and Zipser, D. (1985). Feature discovery by competitive learning*. *Cogn. Sci.* 9, 75–112. doi: 10.1207/s15516709cog0901_5

Self, M. W., Kooijmans, R. N., Supèr, H., Lamme, V. A., and Roelfsema, P. R. (2012). Different glutamate receptors convey feedforward and recurrent processing in macaque v1. *Proc. Natl. Acad. Sci. U.S.A.* 109, 11031–11036. doi: 10.1073/pnas.1119527109

Self, M. W., van Kerkoerle, T., Super, H., and Roelfsema, P. R. (2013). Distinct roles of the cortical layers of area V1 in figure-ground segregation. *Curr. Biol.* 23, 2121–2129. doi: 10.1016/j.cub.2013.09.013

Serre, T., and Poggio, T. (2010). A neuromorphic approach to computer vision. *Commun. ACM* 53, 54–61. doi: 10.1145/1831407.1831425

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 411–426. doi: 10.1109/TPAMI.2007.56

Sherman, S. M., and Guillery, R. W. (1998). On the actions that one nerve cell can have on another: distinguishing drivers from modulators. *Proc. Natl. Acad. Sci. U.S.A.* 95, 7121–7126. doi: 10.1073/pnas.95.12.7121

Spratling, M. W. (2008). Reconciling predictive coding and biased competition models of cortical function. *Front. Comput. Neurosci.* 2:4. doi: 10.3389/neuro.10.004.2008

Spratling, M. W. (2014). A single functional model of drivers and modulators in cortex. *J. Comput. Neurosci.* 36, 97–118. doi: 10.1007/s10827-013-0471-7

Sutton, R. S., and Barto, A. G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychol. Rev.* 88, 135–170. doi: 10.1037/0033-295X.88.2.135

Tetzlaff, C., Kolodziejski, C., Timme, M., and Wrgtter, F. (2012). Analysis of synaptic scaling in combination with hebbian plasticity in several simple networks. *Front. Comput. Neurosci.* 6:36. doi: 10.3389/fncom.2012.00036

Tsotsos, J. K. (1988). "How does human vision beat the computational complexity of visual perception?," in *Computational Processes in Human Vision: An Interdisciplinary Perspective*, ed Z. W. Pylyshyn (Norwood, NJ: Ablex), 286–338.

Tsotsos, J. K. (2005). "Computational foundations for attentive processes," in *Neurobiology of Attention*, eds L. Itti, G. Rees, and J. Tsotsos (Amsterdam, NL: Elsevier), 3–7.

Ullman, S. (1995). Sequence seeking and counter streams: a computational model for bidirectional information flow in the visual cortex. *Cereb. Cortex* 5, 1–11. doi: 10.1093/cercor/5.1.1

Ullman, S. (2007). Object recognition and segmentation by a fragment-based hierarchy. *Trends Cogn. Sci.* 11, 58–64. doi: 10.1016/j.tics.2006.11.009

Ullman, S., Vidal-Naquet, M., and Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nat. Neurosci.* 5, 682–687. doi: 10.1038/nn870

von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik* 14, 85–100. doi: 10.1007/BF00288907

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 July 2014; accepted: 23 October 2014; published online: 05 December 2014.

Citation: Layher G, Schrodt F, Butz MV and Neumann H (2014) Adaptive learning in a compartmental model of visual cortex—how feedback enables stable category learning and refinement. Front. Psychol. ***5****:1287. doi: 10.3389/fpsyg.2014.01287*

This article was submitted to Perception Science, a section of the journal Frontiers in Psychology.

Copyright © 2014 Layher, Schrodt, Butz and Neumann. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Declaration of Authorship

Statutory Declaration

I hereby declare that I wrote the present thesis with the title "What, Where, and When? Mechanisms of Learning Biological Motion Representations." independently and used no other aids that those cited. In each individual case, I have clearly identified the source of the passages that are taken word for word or paraphrased from other works.
I also hereby declare that I have carried out my scientific work according to the principles of good scientific practice in accordance with the current "Satzung der Universität Ulm zur Sicherung guter wissenschaftlicher Praxis" [Rules of the University of Ulm for Assuring Good Scientific Practice] (§8 Sec. 1 Nr. 5 Regulations for Doctoral Studies).
I agree that the dissertation will also be used to verify compliance with generally accepted scientific standards, in particular including electronic data processing programs (§8 Sec. 1 Nr. 8 Regulations for Doctoral Studies).

Erklärung

Ich versichere hiermit, dass ich die Arbeit mit dem "What, Where, and When? Mechanisms of Learning Biological Motion Representations." selbständig angefertigt habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht und die zur Zeit gültige Satzung der Universität Ulm zur Sicherung guter wissenschaftlicher Praxis beachtet habe (§8 Abs. 1 Nr. 5 Rahmenpromotionsordnung).
Ich bin damit einverstanden, dass die Dissertation auch zum Zweck der Überprüfung der Einhaltung allgemein geltender wissenschaftlicher Standards benutzt wird, insbesondere auch unter Verwendung elektronischer Datenverarbeitungsprogramme (§8 Abs. 1 Nr. 8 Rahmenpromotionsordnung).

.....	.....	.....
place	date	Georg Layher