Neurodynamical Model of the Visual Recognition of Dynamic Bodily Actions from Silhouettes

Prerana Kumar ✉ [1,2], Nick Taubert[1], Rajani Raman[3], Anna Bognár[3], Ghazaleh Ghamkhari Nejad[3], Rufin Vogels[3], and Martin A. Giese[1]

[1] Section for Computational Sensomotorics, Centre for Integrative Neuroscience and Hertie Institute for Clinical Brain Research, University Clinic Tübingen, Germany
[2] International Max Planck Research School for Intelligent Systems, Tübingen, Germany
[3] Laboratory of Neuro- and Psychophysiology, Department of Neurosciences, KU Leuven, Belgium
{prerana.kumar,martin.giese}@uni-tuebingen.de

Abstract. For social species, including primates, the recognition of dynamic body actions is crucial for survival. However, the detailed neural circuitry underlying this process is currently not well understood. In monkeys, body-selective patches in the visual temporal cortex may contribute to this processing. We propose a physiologically-inspired neural model of the visual recognition of body movements, which combines an existing image-computable model ('ShapeComp') that produces high-dimensional shape vectors of object silhouettes, with a neurodynamical model that encodes dynamic image sequences exploiting sequence-selective neural fields. The model successfully classifies videos of body silhouettes performing different actions. At the population level, the model reproduces characteristics of macaque single-unit responses from the rostral dorsal bank of the Superior Temporal Sulcus (Anterior Medial Upper Body (AMUB) patch). In the presence of time gaps in the stimulus videos, the predictions made by the model match the data from real neurons. The underlying neurodynamics can be analyzed by exploiting the framework of neural field dynamics.

Keywords: Action recognition · Silhouettes · Neurodynamical model · Neural field · Visual cortex.

1 Introduction

Electrophysiological and neuroimaging studies have uncovered the presence of body-selective neurons and regions in the visual cortex. Body-selective regions in the occipitotemporal cortex (OTC), the extrastriate body area (EBA) [3] and the fusiform body area (FBA) [8, 17] have been discovered in human functional magnetic resonance imaging (fMRI) studies. fMRI studies in monkeys have demonstrated the presence of numerous body-selective patches [2, 18, 21, 23] in

the visual temporal cortex, including the Superior Temporal Sulcus (STS). In these regions, work on fMRI and single-unit responses has demonstrated stronger responses to bodies than to faces and other categories of objects.

However, the focus of most of these studies has been on static bodies. The detailed neural computations underlying the visual recognition of dynamic body actions are not yet well understood. While there have been some single-cell studies investigating the responses of neurons, especially in the STS, to biological motion and body motion [14,15,22], detailed physiologically-inspired neural models of the processing of dynamic bodies are required to clarify the underlying neural computations.

Biologically-inspired models have previously been proposed for the recognition of dynamic bodies [6,9,12]. Older models largely used hierarchies of primitive detectors, such as Gabor filters, for modeling the initial layers of the visual pathway. More recent studies predominantly model the visual pathway using feedforward convolutional neural networks (CNNs) [25] and other studies use different hierarchical neural network architectures [16], but these models do not make use of physiologically-plausible dynamical neural circuits. Our model combines approaches from deep learning in the form of a front-end CNN architecture (ShapeComp network [13]) which has been trained to produce perceptually relevant shape features of objects, with a neurodynamical model based on neural fields that reproduces the dynamic properties of action-selective neurons in the STS and premotor cortex [4,6].

In this paper, we aim to present, as a proof-of-concept, a physiologically-inspired model of the neural circuitry involved in the visual recognition of static body poses and dynamic body movements. The purpose of our model is to reproduce the invariance properties of cortical neurons, and not primarily to achieve maximum classification performance on large data sets. In its present form, the model can learn to classify actions from body silhouettes and reproduces activation dynamics of a population of body-responsive neurons in the macaque STS. Previous studies [10,19] have shown strong and selective responses in STS body patches to silhouettes comparable to those to shaded images, which is in agreement with the well-known shape-bias of human vision [11]. We present some initial comparisons with macaque electrophysiological data recorded from the AMUB body patch [2] of the rostral dorsal bank of the STS using dynamic silhouettes extracted from videos of real macaques. This work provides a starting point for the development of a detailed model of the shape selectivity and dynamic properties of body-selective neurons in the macaque STS. The model also makes predictions about the output dynamics for stimulus videos, including the responses to stimuli with time gaps of different durations, which qualitatively match the data from real neurons.

In the following sections of the paper, we will first present the architecture of the model. We will then describe the results of the recognition of dynamic human silhouettes performing actions, after training the model on only a few exemplars. Following this, we will compare our simulations with macaque electrophysiological data. After briefly explaining the behavior of the model for stimuli with

time gaps by analyzing the underlying neurodynamics, we will finally discuss the implications of this work.

2 Architecture of the Model

The model combines an image-computable model ('ShapeComp' [13]) that produces high-dimensional vectors describing the shapes of objects, with an existing neurodynamical model [6] which has previously replicated the neural dynamics in higher areas of the visual and premotor cortex. The model takes videos (image sequences) of silhouettes performing various actions as input, and the output layer consists of neurons that classify the various learned body actions. The shape features extracted from the ShapeComp network are used to train radial basis function networks whose outputs feed into sequence selective neural fields (recurrent neural networks) that encode temporal sequences of keyframes (dynamic stimuli). The outputs of individual neural fields representing the different body actions or movements are temporally summated by motion pattern neurons that comprise the highest (readout) level of the model.

An overview of the model architecture is shown in Fig.1. Sections 2.1 and 2.2 will describe the components of the model in more detail.

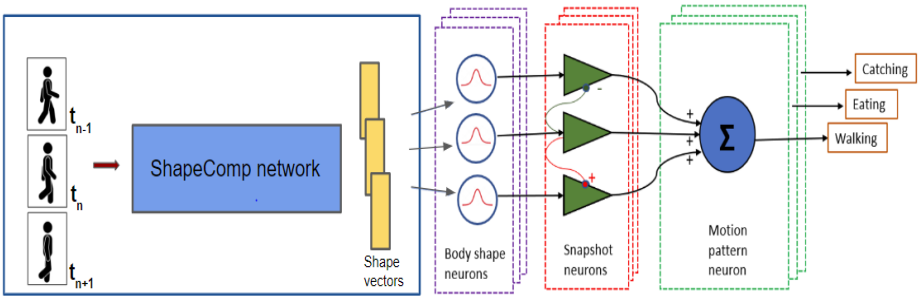

Fig. 1: Overview of model architecture: A CNN architecture (ShapeComp) is combined with a recurrent network of snapshot neurons that integrate information over time.

2.1 Extraction of Mid-Level Shape Features

The initial layers of the visual pathway that detect mid-level features are modeled by a CNN. We initially tested some standard CNNs from the computer vision literature as alternative front-ends of our model. For the detection of body postures across different individuals, we found that these networks did not facilitate robust recognition of key poses with invariance across different individuals when the model was only trained on moderately-sized data sets. A more robust recognition of body pose could be realized using the ShapeComp model [13]. This psychophysically-validated model uses the shape boundaries of objects to

produce high-dimensional vectors that represent the shapes of objects. As the shape vectors produced by the ShapeComp model predict human shape similarity judgments better than features output by standard CNN architectures, we used this architecture as the front-end of our model [13].

The version of the ShapeComp architecture used in the model is a multilayer feedforward CNN called KerNet1, pre-trained on 800,000 shapes produced by a Generative Adversarial Network (GAN), spanning the high-dimensional shape space. The network takes silhouettes of objects as input and produces 22-dimensional feature vectors as output, that describe the objects' shapes in a compact manner. These 22 dimensions are weighted linear combinations of the original 109 image-computable shape features from the ShapeComp model [13]. The architecture generates shape vectors for every keyframe, which form the input to the dynamic layers of the network that are described in the following section.

2.2 Dynamic Recognition Network

Body Shape Neurons The mid-level feature output from the ShapeComp network was used as input for body shape detectors, which were modeled by Gaussian Radial Basis functions (RBFs), which we refer to as body shape neurons in the following text. The centers $\boldsymbol{z}_n^a$ of these RBFs were defined by the 22-dimensional shape vectors $\boldsymbol{z}_n^a$ from the previous layer representing different keyframes from the training movies, indexed by n. The different actions are represented by the integer variable a. The outputs of the body shape neurons were given by the equation:

$$r_n^a = \exp\left(-|\boldsymbol{z} - \boldsymbol{z}_n^a|^2/2\sigma^2\right) \ . \tag{1}$$

The outputs of the body shape neurons $r_n^a(t)$ were smoothed along the neuron axis using a Gaussian filter (of width 2 neurons) to generate the inputs $s_n^a(t)$ for the next layer.

Snapshot Neurons The smoothed output of the body shape neurons $s_n^a(t)$ provides input to sequence-selective snapshot neurons that encode temporal sequences of keyframes. Asymmetric lateral connections between these snapshot neurons encoding the image keyframes result in recurrent neural networks that show sequence selectivity i.e. the network only responds strongly if the learned keyframes occur in the correct temporal order. The underlying network dynamics can be interpreted as a dynamic neural field [1]. Each learned action is encoded by such a network. The dynamics of the discretely approximated neural field [1] is given by the following equation ($[u]_+ = u$ for $u > 0$, and 0 otherwise):

$$\begin{aligned} \tau \dot{u}_n^a(t) =& -u_n^a(t) + \sum_m w(n-m)\,[u_m^a(t)]_+ + s_n^a(t) - h - w_c I_c^a(t) \ , \\ w(n) =& A\ \exp\left(-(n-C)^2/2\sigma_{ker}^2\right) - B \ , \end{aligned} \tag{2}$$

where $u_n^a(t)$ denotes the activity of the neuron in the neural field that encodes the keyframe n of the body action category a, and where the index m runs over all neurons. The resting level of the neurons is determined by the positive parameter h (= 1), and τ (= 28 ms) defines the time constant of the dynamics. The function w is an asymmetric interaction kernel. The neural sub-networks encoding different actions compete with each other. This is accomplished by the cross-inhibition term $I_c^a(t)$ that is given by the equation $I_c^a(t) = \sum_{m,a' \neq a} [u_m^{a'}(t)]_+$. The parameter w_c (= 1.5) determines the strength of the cross-field inhibition.

The snapshot neurons are keyframe-selective as well as action-selective, and exhibit phasic activity during the temporal progression of the presented action stimuli.

Motion Pattern Neurons The responses of the snapshot neurons encoding the same action are temporally smoothed and summated by *motion pattern neurons* that form the next layer of the model. The response of these neurons is dependent on the sum of the (thresholded) activity of the snapshot neurons encoding the corresponding actions and given by the equation:

$$\tau_v \dot{v}_a(t) = -v_a(t) + \sum_n [u_n^a(t)]_+ \ . \tag{3}$$

In the above equation, $v_a(t)$ denotes the activity of the motion pattern neurons, and τ_v (= 28 ms) denotes the time constant of their dynamics.

Each motion pattern neuron encodes a particular action and is active during the corresponding action. It is at this level of the hierarchy that the model classifies the different types of actions. The responses of these motion pattern neurons have been compared (at the population level) with single-unit responses recorded from the macaque STS in section 3.2.

3 Results

3.1 Testing the Model on Sequences of Human Silhouettes

Videos of silhouettes of 9 human subjects performing 5 types of actions that were clearly distinguishable from silhouettes were used to test the model's ability to learn actions. The selected action sequences, chosen from the publicly available Weizmann Human Action Dataset [7], were: walking, running, jumping jacks, waving with one arm, and waving with two arms. The Weizmann Human Action Dataset includes videos of the silhouettes of the subjects performing the actions, which were used for training and testing the model.

Image sequences of 25 black-and-white silhouette keyframes per action were extracted from the longer video sequences (deinterlaced 50 frames/s). The image sequences from the various subjects were coarsely time-normalized by visual inspection, such that the first and last images of the sequences of a particular action contained the corresponding poses across all subjects. The available silhouette images were already aligned/centered to a reference point, removing

any effects of translation of the subject within the image. The silhouettes contained multiple "leaks" and "intrusions", which served as a test of the model's robustness. The images, which were of different sizes, were resized to uniform dimensions of 224 x 224 pixels.

Averaged across all action types, we achieved a performance of 97.77% correct classifications on the test set, determined by cross-validation (leave-one-out analysis on 9 videos per action type). The classification accuracy of the neurodynamical model using ShapeComp as its front-end was compared with that using a CNN architecture, ResNet-101, coupled with different unsupervised dimensionality reduction algorithms. ResNet-101 has been shown to predict human shape similarity better than other standard CNNs [13] and was used to produce mid-level features. The network (pre-trained on ImageNet) was read out just before the fully connected layer (at layer "Pool5"). Only output features showing high variance over time were retained (feature selection), and 3 types of unsupervised dimensionality reduction methods were applied to construct a lower-dimensional mid-level feature space (of 15 dimensions) - Principal Component Analysis (PCA), Non-negative Matrix Factorization (NNMF) and Independent Component Analysis (ICA). Increasing the number of dimensions of these mid-level feature vectors beyond 15 was found to decrease classification accuracy, probably due to overfitting to the training data.

As shown in Table 1 below, our model outperforms the CNN architecture combined with any of the three unsupervised dimensionality reduction algorithms. As another test of performance, we also added Gaussian random noise to all the models at 2 dynamic neural levels during the simulations: at the level of the snapshot neuron responses and at the level of the motion pattern neuron responses. We then re-computed the accuracy values, shown in Table 2, and performed this analysis for 3 levels ($\sigma = 2$, $\sigma = 6$, $\sigma = 10$) of noise. The model with the ShapeComp front-end shows the highest classification accuracy even in the presence of added noise.

Table 1: Accuracy of our model compared with a CNN model and different dimension reduction methods.

Model Front-End	Accuracy
ShapeComp	**97.8%**
ResNet-101 + PCA	68.8%
ResNet-101 + NNMF	75.6%
ResNet-101 + ICA	55.6%

Table 2: Accuracy of our model compared with the other models for different levels of added noise

Model Front-End	$\sigma = 2$	$\sigma = 6$	$\sigma = 10$
ShapeComp	**86.7%**	**82.2%**	**80%**
ResNet-101 + PCA	66.7%	48.9%	48.9%
ResNet-101 + NNMF	71.1%	51.1%	51.1%
ResNet-101 + ICA	53.3%	55.6%	60%

3.2 Simulations in Comparison with Macaque Electrophysiological Data

A stimulus set of 20 videos of silhouettes of rhesus monkeys performing different dynamic body movements was created for use in the experiments and modeling. The silhouettes were centered in the videos, removing any effects of translation

of the macaque within the images. Image sequences of 60 grayscale keyframes (1s) were extracted from the videos (480 x 480 pixel images, 60 frames/s). A set of different stimulus conditions per video were used for both the experiments and modeling: Image sequences taken in the correct temporal order ("forward" condition), temporally inverted image sequences ("reverse" condition), and videos with different lengths of time gaps, during which frames of the video were replaced by blank frames for both forward and reverse conditions. The time gaps used were of 2, 4, 6, 8, 10, and 13 frames (approximately 33, 67, 100, 133, 167, and 217 ms respectively). The positions of the rest of the frames containing the macaque were unaltered in the image sequences.

For this analysis, from the responses of 32 cells recorded from the AMUB body patch, the response (averaged across 5 trials) to the video that produced the highest response over time ("best" stimulus) was chosen for each cell. There were 16 different best stimuli in total for the population. The responses of each cell were recorded for the different stimulus conditions for each video. The baseline-subtracted activity of each cell was normalized by dividing by the maximum firing rate value (bin-width = 20 ms) of the net response of that cell across all the stimulus conditions under consideration. All the cell responses were then averaged to produce the population response. Finally, the neural response curves were smoothed over time by Gaussian filtering. Likewise, the model was tested on the 16 best stimulus videos of the neurons for the same stimulus conditions. The responses of each of the motion pattern neurons to its preferred stimulus video were normalized in the same manner as in the data. The individual motion pattern neuron responses were averaged to obtain the population activity.

The model successfully reproduces the sequence selectivity of the population response of the real neurons (Fig. 2A and Fig. 2B). Interestingly, the model predicts that the difference in the population activity for the forward and reverse-ordered sequences should significantly decrease in the presence of large time gaps in the stimuli, which is actually found to be the case in the data from the experiments (Fig. 2C and Fig. 2D). In the model, this behavior can be explained by the recurrent network dynamics. If the input activity is not sufficiently continuous, a self-organized solution, which corresponds to a traveling pulse in the neural field [24], cannot emerge. In this case, the direction selectivity of the model disappears. This is systematically tested in the simulations shown in (Fig. 2E) showing the population responses of the motion pattern neurons, averaged over all time points, for different lengths of stimulus time gaps, for both the forward and reverse conditions. For larger durations, the strong sequence selectivity present for continuous stimuli disappears, while the output neurons are still significantly active. Fig. 2F shows the corresponding plot for the population responses from the neural data averaged during the stimulus period (accounting for the 60 ms response latency period of the neurons), which corresponds well with the model's prediction.

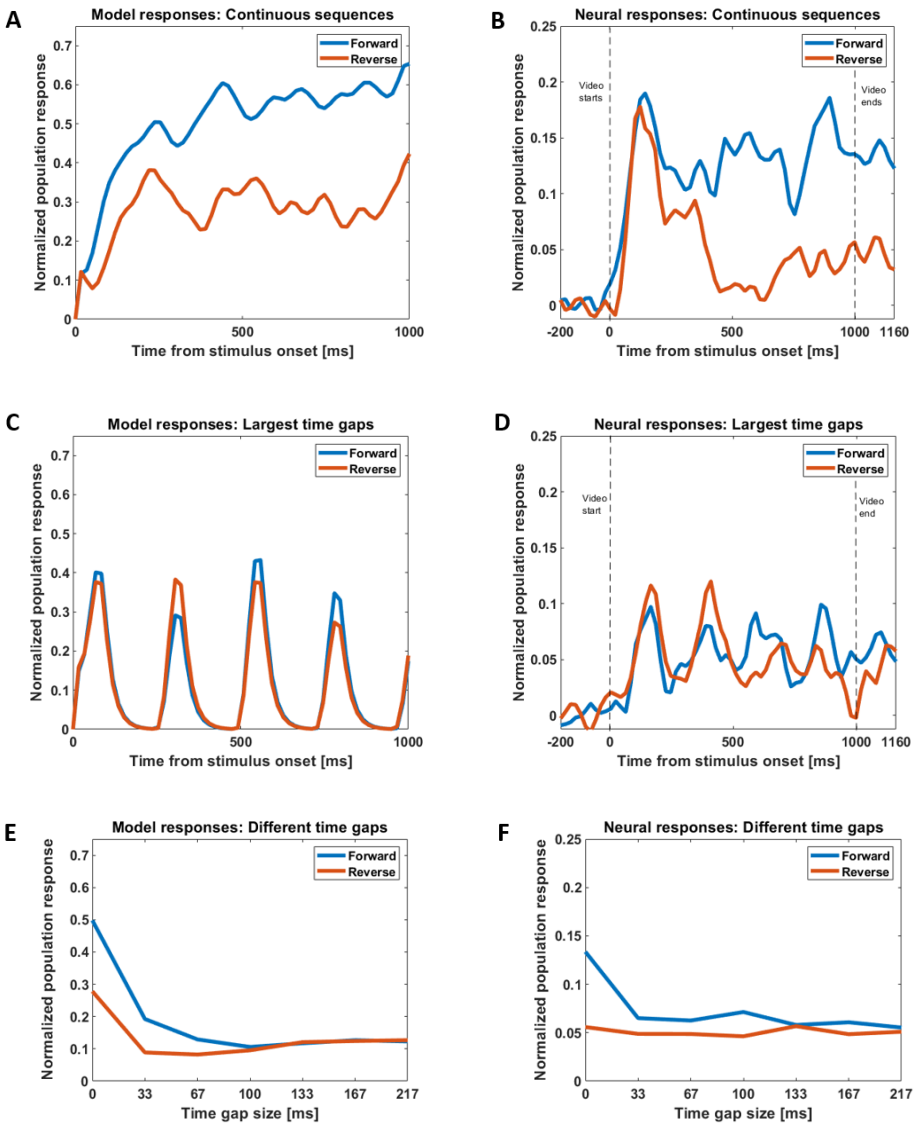

Fig. 2: Simulation results: **A** Simulated population response for forward and reverse ordered continuous sequences. **B** Population response from AMUB body patch neurons for the same continuous stimuli. **C** Simulated population response for the largest time gap (217 ms) condition of the forward and reverse-ordered stimuli. **D** Population response from AMUB body patch neurons for the same time gap length. **E** Predicted responses from the model for forward and reverse-ordered sequences containing different lengths of time gaps. **F** Population response from AMUB body patch for the same time gap stimuli.

3.3 Mathematical Analysis of the Dependence of Sequence Selectivity on Gap Duration

The core module of our model that integrates information over time is the recurrent neural network (2). Sequence selectivity and its dependence on the time gaps in the stimulus are most easy to analyze by describing the recurrent neural network in a continuum limit, resulting in the following neural field (ignoring the cross-field inhibition):

$$\tau \frac{\partial u(x,t)}{\partial t} + u(x,t) = \int w(x-x')\, \theta(u(x',t))\, \mathrm{d}x' + s(x,t)\ . \tag{4}$$

Here, we integrate the resting level parameter h into the input signal $s(x,t)$ for simplicity. The function $\theta(\cdot)$ defines the output threshold characteristics of the neurons. For $\theta(u) \equiv u$, one obtains a linear neural field that is particularly easy to analyze. We treat this case here, and the analysis of nonlinear threshold functions will be treated in future publications. The input signal is assumed to be a traveling Gaussian peak of the form $s(x,t) = \exp(-(x-vt)^2/(2\eta^2)) \cdot \Xi(t)$, where the function $\Xi(t)$ is one while stimulus frames are present and zero during the time gaps. The traveling speed v of the input is determined by the frame rate of the stimulus video.

The analysis of the dynamics becomes easier in a traveling coordinate system, exploiting the identities $U(y,t) = u(x,t)$ and $S(y,t) = s(x,t) = \exp(-y^2/(2\eta^2)) \cdot \Xi(t)$, where $y = x - vt$. The resulting transformed dynamics are given by:

$$\tau \frac{\partial U(y,t)}{\partial t} - \tau v \frac{\partial U(y,t)}{\partial y} + U(y,t) = \int w(y-y')\, \theta(U(y',t))\, \mathrm{d}y' + S(y,t)\ . \tag{5}$$

For the case of $\theta(u) \equiv u$, the last equation can be solved by Fourier transformation in the space-time frequency domain. The 2D Fourier transformation of the solution is given by the following product $\tilde{U}(k,\omega) = \tilde{S}(k,\omega)\tilde{H}(k,\omega)$, where $\tilde{S}(k,\omega)$ is the Fourier transformation of the input signal. The function $\tilde{H}(k,\omega) = 1/(1 + i\omega\tau - i\tau v k - \tilde{w}(k))$ is the impulse response of the dynamics. The amplitude of this function is illustrated in Fig. 3 (panels A and D). It changes for opposite signs of the velocity v, modeling the forward and reverse temporal orders of the stimulus video frames. $\tilde{w}(k)$ is the Fourier Transform of $w(x)$.

In this analytically solvable linear neural field, we also observe a dependence of the temporal order selectivity on the presence and duration of stimulus gaps. This is illustrated in Fig. 3B and Fig. 3C that show the computed solutions $u(x,t)$ for forward vs. reverse presentation of the stimulus frames without time gaps, which show different maximum (and average) amplitudes. Contrasting with this observation, the computed solutions for the stimulus with time gaps (shown in Fig. 3E vs. Fig. 3F) show only a minimal amplitude difference between the forward and reverse presentation orders.

This dependence of the amplitude difference on the presence of gaps in the stimulus signal is caused by the fact that the stimulus with gaps activates high-frequency components along the ω axis in the (k,ω)-frequency space. For these

components, the denominator of $\tilde{H}(k,\omega)$ is effectively less sensitive to the parameter v, and thus on the presentation order of the stimulus.

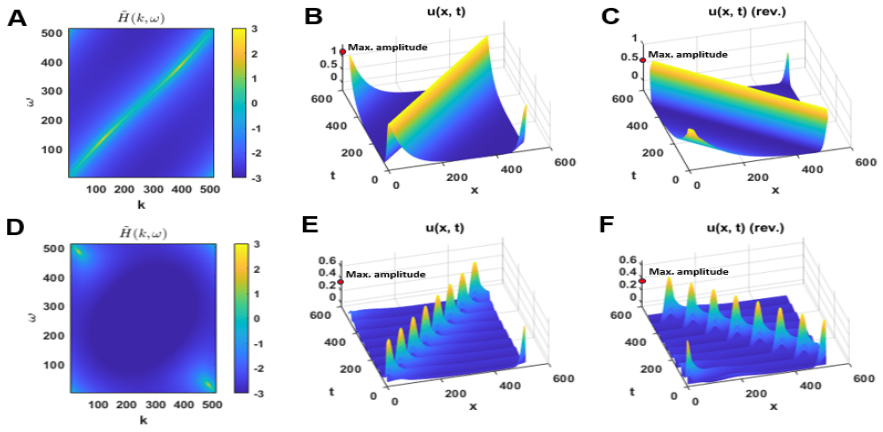

Fig. 3: Linear neural field: Fourier Spectra of the impulse response $\tilde{H}(k,\omega)$: **A** for $(v > 0)$ (forward temporal order), and **D** for $(v < 0)$ (reverse temporal order). Panels **B** and **C** show the computed solutions $u(x,t)$ for forward vs. reverse temporal orders for a stimulus without time gaps. Panels **E** and **F** show the computed solution with time gaps of a duration of 7 stimulus frames. Amplitude difference between the solutions is lower for the stimulus with gaps.

An important observation is that for the linear neural field it is critical to sum up *thresholded* output amplitudes, as assumed in equation (3) of our model. Just integrating the signal $u(x,t)$ over x turns out to be equivalent to computing the Fourier back-transformation (in time) of the function $\tilde{H}(0,\omega)$. This function does not depend on the velocity v and therefore fails to imply sequence selectivity.

4 Conclusions

In this paper, we have presented a neurodynamical model for the recognition of dynamic bodily actions performed by silhouettes, as a proof of concept. Despite the simple architecture of the model, even when trained with relatively few training examples, it accomplishes robust recognition of body poses and actions across different individuals. Using a standard CNN architecture, ResNet-101, combined with different unsupervised learning techniques, we were not able to reproduce the same robustness in body shape recognition. We think that this lack of performance could be related to the tendency of standard CNNs to overemphasize shape differences that are not relevant for keyframe pose recognition. Using modified standard CNN architectures as potential front-ends of the model with further optimization of the feature selection procedure may possibly yield higher accuracy values, but this was outside the scope of the current study.

Our model also reproduces signatures of the activity dynamics of populations of body-responsive neurons in the AMUB body patch of the STS. We could reproduce the sequence selectivity of this response, and also the fact that introducing large time gaps in the stimuli destroys this sequence selectivity. In our model, this is a consequence of the recurrent network dynamics. We mathematically analyzed this dependence of sequence selectivity on gap duration using analysis methods from linear neural field dynamics.

A major limitation of the model is that it works only on silhouettes. In future work we will try to extend the front-end of our model for more natural stimuli, where a key problem is to overcome the texture bias that is present in standard CNN architectures [5]. Furthermore, it is likely that the shape descriptors produced by the ShapeComp architecture do not exactly match those used in the brain for shape recognition. Nevertheless, we have used the ShapeComp CNN architecture to model to realize a front-end of our model that reproduces invariance properties of human shape perception better than other standard CNN networks. Another limitation of the model is the absence of neurons exhibiting both phasic and tonic responses to continuous dynamic sequences, which is an oversimplification. Finally, it is likely that neurons in the AMUB body patch are also selective for motion or optical flow features. Our model cannot account for the selectivity for local motion features. Building two pathway architectures that can reproduce these properties remains an important challenge in the modeling of the detailed properties of cortical body action-selective neurons. [6, 20]

Acknowledgements This work was supported by ERC 2019-SyG-RELEVANCE-856495; SSTeP-KiZ BMG:ZMWI1-2520DAT700.

References

1. Amari, S.: Dynamics of pattern formation in lateral-inhibition type neural fields. Biological Cybernetics **27**(2), 77–87 (1977)
2. Bognár, A., Raman, R., Taubert, N., Zafirova, Y., Li, B., Giese, M., De Gelder, B., Vogels, R.: The contribution of dynamics to macaque body and face patch responses. NeuroImage **269**, 119907 (2023)
3. Downing, P.E.: A cortical area selective for visual processing of the human body. Science **293**, 2470–2473 (09 2001). https://doi.org/10.1126/science.1063414
4. Fleischer, F., Caggiano, V., Thier, P., Giese, M.A.: Physiologically inspired model for the visual recognition of transitive hand actions. Journal of Neuroscience **33**(15), 6563–6580 (2013)
5. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 (2018)
6. Giese, M.A., Poggio, T.: Neural mechanisms for the recognition of biological movements. Nature Reviews Neuroscience **4**(3), 179–192 (2003)
7. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(12), 2247–2253 (2007)
8. Hadjikhani, N., de Gelder, B.: Seeing fearful body expressions activates the fusiform cortex and amygdala. Current Biology **13**, 2201–2205 (12 2003). https://doi.org/10.1016/j.cub.2003.11.049

9. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: 2007 IEEE 11th International Conference on Computer Vision. pp. 1–8. IEEE (2007)
10. Kalfas, I., Kumar, S., Vogels, R.: Shape selectivity of middle superior temporal sulcus body patch neurons. ENeuro **4**(3) (2017)
11. Landau, B., Smith, L.B., Jones, S.S.: The importance of shape in early lexical learning. Cognitive Development **3**(3), 299–321 (1988)
12. Lange, J., Lappe, M.: A model of biological motion perception from configural form cues. Journal of Neuroscience **26**(11), 2894–2906 (2006)
13. Morgenstern, Y., Hartmann, F., Schmidt, F., Tiedemann, H., Prokott, E., Maiello, G., Fleming, R.W.: An image-computable model of human visual shape similarity. PLoS Computational Biology **17**(6), e1008981 (2021)
14. Oram, M., Perrett, D.: Responses of anterior superior temporal polysensory (stpa) neurons to "biological motion" stimuli. Journal of Cognitive Neuroscience **6**(2), 99–116 (1994)
15. Oram, M., Perrett, D.: Integration of form and motion in the anterior superior temporal polysensory area (stpa) of the macaque monkey. Journal of Neurophysiology **76**(1), 109–129 (1996)
16. Parisi, G.I., Tani, J., Weber, C., Wermter, S.: Emergence of multimodal action representations from neural network self-organization. Cognitive Systems Research **43**, 208–221 (2017)
17. Peelen, M.V., Downing, P.E.: Selectivity for the human body in the fusiform gyrus. Journal of Neurophysiology **93**, 603–608 (01 2005). https://doi.org/10.1152/jn.00513.2004
18. Popivanov, I.D., Jastorff, J., Vanduffel, W., Vogels, R.: Stimulus representations in body-selective regions of the macaque cortex assessed with event-related fmri. Neuroimage **63**(2), 723–741 (2012)
19. Popivanov, I.D., Jastorff, J., Vanduffel, W., Vogels, R.: Tolerance of macaque middle sts body patch neurons to shape-preserving stimulus transformations. Journal of Cognitive Neuroscience **27**(5), 1001–1016 (2015)
20. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems **27** (2014)
21. Tsao, D.Y., Freiwald, W.A., Knutsen, T.A., Mandeville, J.B., Tootell, R.B.H.: Faces and objects in macaque cerebral cortex. Nature Neuroscience **6**, 989–995 (08 2003). https://doi.org/10.1038/nn1111
22. Vangeneugden, J., De Maziere, P.A., Van Hulle, M.M., Jaeggli, T., Van Gool, L., Vogels, R.: Distinct mechanisms for coding of visual actions in macaque temporal cortex. Journal of Neuroscience **31**(2), 385–401 (2011)
23. Vogels, R.: More than the face: Representations of bodies in the inferior temporal cortex. Annual Review of Vision Science **8**, 383–405 (2022)
24. Xie, X., Giese, M.A.: Nonlinear dynamics of direction-selective recurrent neural media. Physical Review E **65**(5), 051904 (2002)
25. Yamins, D.L., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J.: Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences **111**(23), 8619–8624 (2014)