Noname manuscript No.
(will be inserted by the editor)

Modeling Perspective-Taking by Forecasting 3D Biological Motion Sequences

Fabian Schrodt · Martin V. Butz

Received: date / Accepted: date

Abstract The mirror neuron system (MNS) is believed to be involved in social abilities like empathy and imitation. While several brain regions have been linked to the MNS, it remains unclear how the mirror neuron property itself develops. Previously, we have introduced a recurrent neural network, which enables mirror-neuron capabilities by learning an embodied, scale- and translation-invariant model of biological motion (BM). The model allows the derivation of the orientation of observed BM by (i) segmenting BM in a common positional and angular space and (ii) generating short-term, top-down predictions of subsequent motion. While our previous model generated short-term motion predictions, here we introduce a novel forecasting algorithm, which explicitly predicts sequences of BM segments. We show that the model scales on a 3D simulation of a humanoid walking and is robust against variations in body morphology and postural control.

Keywords: Perspective Taking; Embodiment; Biological Motion; Self-Supervised Learning; Sequence Forecasting; Mirror-Neurons; Recurrent Neural Networks

1 Introduction

This paper investigates how we may be able to recognize BM sequences and mentally transform them to the egocentric frame of reference to bootstrap mirror neuron properties. Our adaptive, self-supervised, recurrent neural network model (Schrodt et al, 2014) might contribute to the understanding of the MNS and its implied capabilities. With the previous model, we were able to generate continuous mental rotations to learned canonical views of observed 2D BM – essentially taking on the perspective of an observed person. This self-supervised perspective taking was accomplished by back-propagating errors stemming from top-down, short-term predictions of the BM progression.

In this work, we introduce an alternative or complementary, time-independent forecasting mechanism of motion segment sequences to the model. In the brain, prediction and forecasting mechanisms may be realized by the cerebellum, which is involved in the processing of BM (Grossman et al, 2000). In addition, it has been suggested that the cerebellum may also support the segmentation of motion patterns via the basal ganglia, thereby influencing the learning of motor sequences in parietal and (pre-)motor cortical areas (Penhune and Steele, 2012). Along these lines, the proposed model learns to predict segments of motion patterns given embodied, sensori-motor motion signals. Due to the resulting perspective taking capabilities, the model essentially offers a mechanism to activate mirror neuron capabilities.

2 Neural Network Model

The model consists of three successive stages illustrated in the overview given in Fig. 1. The first stage processes relative positional and angular values into mentally rotated, motion-direction sensitive population codes. The second stage performs a modulatory normalization and pooling of those. Stage III is a self-supervised pattern segmentation network with sequence forecasting, which

Cognitive Modeling, Computer Science Department, University of Tübingen, Sand 14, 72076 Tübingen, Germany
E-mail: tobias-fabian.schrodt@uni-tuebingen.de
E-mail: martin.butz@uni-tuebingen.de

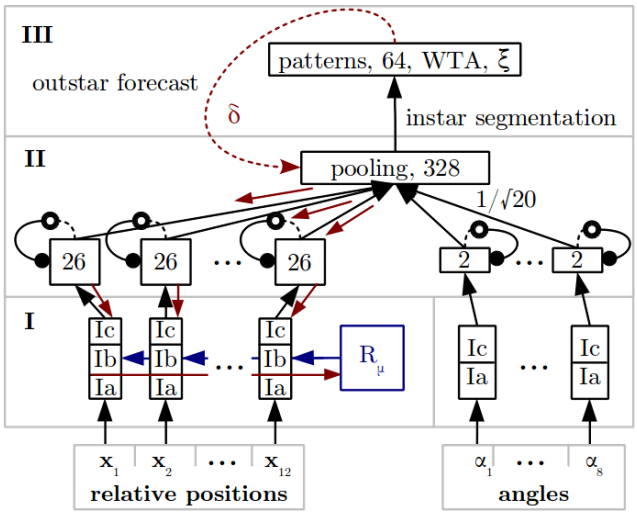

Fig. 1 Overview of the three-stage neural modeling approach in a 3D example with 12 joint positions and 8 joint angles, resulting in $n = 20$ features. Boxes numbered with m indicate layers consisting of m neurons. Black arrows describe weighted forward connections, while circled arrowheads indicate modulations. Dashed lines denote recurrent connections. Red arrows indicate the flow of the error signals.

enables the back-propagation of forecast errors. We detail the three stages and the involved techniques in the following sections.

2.1 Stage I - Feature Preprocessing

The input of the network is driven by a number of (not necessarily all) relative joint positions and joint angles of a person. Initially, the network can be driven by self-perception to establish an egocentric perspective on self-motion. In this case, the relative joint positions may be perceived visually, while the perception of the joint angles may be supported by proprioception in addition to vision. When actions of others are observed, joint angles may be solely identified visually.

In each single interstage Ia in the relative position pathway, a single, positional body landmark relation is transformed into a directional velocity by time-delayed inhibition, in which way the model becomes translation-invariant. Interstage Ib implements a mental rotation of the resulting directional velocity signals using a neural rotation module R_μ. It is driven by auto-adaptive mental rotation angles (euler angles in a 3D space), which are implemented by bias neurons. The rotational module and its influence on the directional velocity signals are realized by gain field-like modulations of neural populations (Andersen et al, 1985). All positional processing stages apply the same mental rotation R_μ, by which multiple error signals can be merged at the module. This enables orientation-invariance on adequate adaptation of the module's biases. In interstage Ic, each (rotated) D-dimensional directional motion feature is convolved into a population of $3^D - 1$ direction-responsive neurons.

The processing of each one-dimensional angular information is done analogously, resulting in 2-dimensional population codes. A rotation mechanism (interstage Ib) is not necessary for angles and thus not applied. In summary, stage I provides a population of neurons for each feature of sensory processing, which is either sensitive to directional changes in a body-relative limb position (26 neurons for each 3D position) or sensitive to directional changes in angles between limbs (2 neurons for each angle).

2.2 Stage II - Normalization and Pooling

Stage II first implements individual activity normalizations in the direction-sensitive populations. Consequently, the magnitude of activity is generalized over, by which the model becomes scale- and velocity-invariant. Normalization of a layer's activity-vector can be achieved by axo-axonic modulations, using a single, layer-specific normalizing neuron (shown as circles in Fig. 1). Next, all normalized direction-sensitive fields are merged by one-to-one connections to a pooling layer, which serves as the input to stage III. To also normalize the activity of the pooling layer, the connections are weighted by $1/\sqrt{n}$, where n denotes the number of features being processed.

2.3 Stage III - Correlation Learning

Stage III realizes a clustering of the normalized and pooled information from stage II (indexed by i) over time by instar weights fully connected to a number of pattern-responsive neurons (indexed j). Thus, each pattern neuron represents a unique constellation of positional and angular directional movements. For pattern learning, we use the Hebb'ian inspired instar learning rule (Grossberg, 1976). To avoid a "catastrophic forgetting" of patterns, we use winner-takes-all competitive learning in the sense that only the weights to the most active pattern neuron are adapted. We bootstrap the weights from scratch by adding neural noise to the input of each pattern neuron, which consequently activates Hebb'ian learning of novel input patterns. The relative influence of neural noise decreases while a pattern-sensitive neuron is learned (cf. Schrodt et al, 2014).

In contrast to our previous, short-term prediction approach, here we apply a time-independent forecast-

ing algorithm (replacing the attentional gain control mechanism). This is realized by feedback connections w_{ji} from the pattern layer to the pooling layer, which are trained to approximate the input net_i of the pooling layer neurons:

$$1/\eta \cdot \partial w_{ji}(t)/\partial t = \Delta w_{ji}(t) = \text{net}_i(t) - w_{ji}(t) \ , \qquad (1)$$

where neuron j is the last winner neuron that differed from the current winner in the pattern layer. In consequence, the outgoing weight vector of a pattern neuron forecasts the input to the pooling layer while the next pattern neuron is active. The forecasting error can be backpropagated through the network to adapt the mental transformation for error minimization (cf. red arrows in Fig. 1). Thus, perspective adaptation is driven by the difference between the forecasted and actually perceived motion. The difference δ_i is directly fed into the pooling layer by the outstar weights:

$$\delta_i(t) = -\Delta w_{ji}(t) \ , \qquad (2)$$

where j again refers to the preceding winner.

3 Experiments

In this section, we first introduce the 3D simulation we implemented to evaluate our model. We then show that after training on the simulated movement, the learned angular and positional correlations can be exploited to take on the perspective of another person that currently executes a similar motion pattern. The reported results are averaged over 100 independent runs (training and evaluating the network starting with different random number generator seeds).

3.1 Simulation and Setup

We implemented a 3D simulation of a humanoid walking with 10 angular DOF. The movement is cyclic with a period of 200 time steps (corresponding to one left and one right walking step). The simulation provides the 3D positions of all 12 limb endpoints relative to the body's center $\mathbf{x}_1$ ... $\mathbf{x}_{12}$ as well as 8 angles α_1 ... α_8 between limbs (inner rotations of limbs are not considered). The view of the walker can be rotated arbitrarily before serving as visual input to the model.

Furthermore, the simulation allows the definition of the appearance and postural control of the walker. Each of the implied parameters (body scale, torso height, width of shoulders/hips and length of arms/legs, as well as minimum/maximum amplitude of joint angles on movement) can be varied to log-normally distributed variants of an average walker, which exhibits either female or male proportions. Randomly sampled resulting walkers are shown in Figure 2.

3.2 Perspective-Taking on Action Observation with Morphological Variance

We first trained the model on the egocentric perspective of the average male walker for 40k time steps. The rotation biases were kept fixed since no mental rotation has to be applied during self-perception. In consequence, a cyclic series of 4 to 11 winner patterns evolved from noise in the pattern layer. Each represents i) a sufficiently linear part of the walking via its instar vector and ii) the next forecasted, sequential part of the movement via its outstar vector. After training, we fed the model with an arbitrarily rotated (uniform distribution in orientation space) view of a novel walker, which was either female or male with 50% probability. Each default morphology parameter was varied by a log-normal distribution $\exp(\mathcal{N}(0, \sigma^2))$ with variance $\sigma^2 = 0.1$, postural control parameters were not varied. Instar/outstar learning was disabled from then on, but the mental rotation biases were allowed to adapt according to the backpropagated forecast error to derive the orientation of the shown walker.

Fig. 3 shows the mismatch of the model's derived walker orientation, which we term orientation difference (OD), over time. We define the OD by the minimal amount of rotation needed to rotate the derived orientation into the egocentric orientation about the optimal axis of rotation. In result, all trials converged to a negligible OD, which means that the given view of the walker was internally rotated to the previously learned, egocentric orientation. The median remaining OD converged to $\sim 0.15^\circ$ with quartiles of $\sim \pm 0.03^\circ$. The time for the median OD to fall short of 1° was 120 time steps. These results show that morphological differences between the self-perceived and observed walkers could be generalized over. This is because the model's scale-invariance applies to every positional relation perceived by the model.

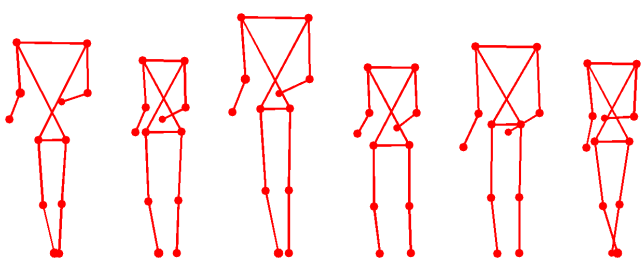

Fig. 2 Variants of the simulated walker.

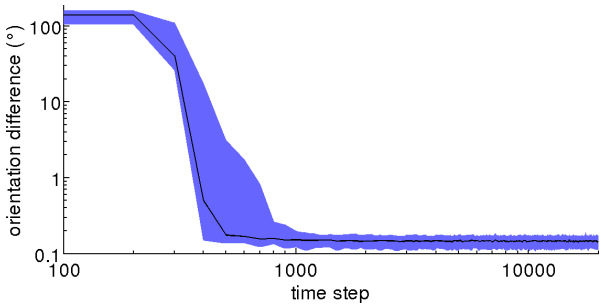

Fig. 3 The model aligns its perspective to the orientation of observed walkers with different morphological parameters (starting at t=200). Blue: quartiles, black: median.

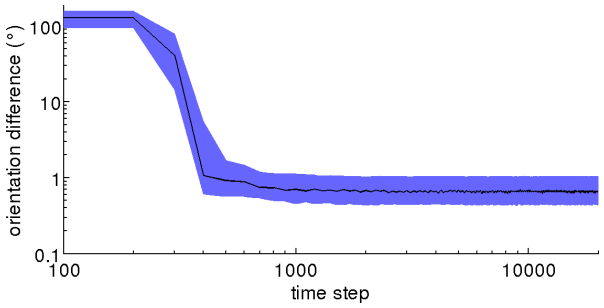

Fig. 4 The model aligns its perspective to the orientation of observed walkers with different postural control parameters.

3.3 Perspective-Taking on Action Observation with Postural Control Variance

In this experiment, we varied the postural control parameters of the simulation on action observation by a log-normal distribution with variance $\sigma^2 = 0.1$, instead of the morphological parameters. Again, female as well as male walkers were presented. The perspective of all shown walkers could be derived reliably, but with a higher remaining OD of $\sim 0.67°$and more distal quartiles of $\sim \pm 0.32°$. The median OD took longer to fall short of 1°, namely 154 time steps. This is because the directions of joint motion are influenced by angular parameters. Still, variations in postural control could largely be generalized over.

4 Conclusion & Future Work

The results have shown that the developed model is able to recognize novel perspectives on BM independent from morphological and largely independent from posture control variations. With the previous model, motion segments are also recognized if their input sequence is reordered, such that additional, implicitly learned attractors may exist for the perspective derivation. Investigations had shown that this applies especially to top-down inverted views on biological motion. The introduced, explicit learning of pattern sequences forces the model to deduce the correct perspective by predicting the patterns of the next motion segment rather than the current one. It may well be the case, however, that the combination of both predictive mechanisms may generate even more robust results. Future work needs to evaluate the current model capabilities and limitations as well as possible combinations of the prediction mechanisms further. Currently, we are investigating how missing or incomplete data could be derived by our model during action observation.

We believe that the introduced model may help to infer the current goals of an actor during action observation somewhat independent of the current perspective. Experimental psychological and further cognitive modeling studies may examine the influence of motor sequence learning on the recognition of BM and the inference of goals. Also, an additional, dynamics-based modulatory module could be incorporated, which could be used to deduce emotional properties of the derived motion – and could thus bootstrap capabilities related to empathy. These advancements could pave the way for the creation of a model on the development of a mirror neuron system that supports learning by imitation and is capable of inferring goals, intentions, and even emotions from observed BM patterns.

References

Andersen RA, Essick GK, Siegel RM (1985) Encoding of spatial location by posterior parietal neurons. Science 230(4724):456–458

Grossberg S (1976) on the development of feature detectors in the visual cortex with applications to learning and reaction-diffusion systems. Biological Cybernetics 21(3):145–159

Grossman E, Donnelly M, Price R, Pickens D, Morgan V, Neighbor G, Blake R (2000) Brain areas involved in perception of biological motion. Journal of cognitive neuroscience 12(5):711–720

Penhune VB, Steele CJ (2012) Parallel contributions of cerebellar, striatal and m1 mechanisms to motor sequence learning. Behavioural brain research 226(2):579–591

Schrodt F, Layher G, Neumann H, Butz MV (2014) Modeling perspective-taking by correlating visual and proprioceptive dynamics. In: 36th Annual Conference of the Cognitive Science Society, Conference Proceedings