Learning Features of Intermediate Complexity for the Recognition of Biological Motion

Rodrigo Sigala[1,2], Thomas Serre[2], Tomaso Poggio[2], and Martin Giese[1]

[1] Laboratory for Action Representation and Learning (ARL), Dept. of Cognitive Neurology,
University Clinic Tübingen, Schaffhausenstr. 113, D-72072 Tübingen, Germany
Rodrigo.Sigala@tuebingen.mpg.de, Martin.Giese@uni-tuebingen.de
[2] McGovern Institute for Brain Research, Brain and Cognitive Sciences, Massachusetts
Institute of Technology, Bldg.E25-201, 45 Carleton St., Cambridge, MA 02142, USA
serre@mit.edu, poggio@ai.mit.edu

Abstract. Humans can recognize biological motion from strongly impoverished stimuli, like point-light displays. Although the neural mechanism underlying this robust perceptual process have not yet been clarified, one possible explanation is that the visual system extracts specific motion features that are suitable for the robust recognition of both normal and degraded stimuli. We present a neural model for biological motion recognition that learns robust mid-level motion features in an unsupervised way using a neurally plausible memory-trace learning rule. Optimal mid-level features were learnt from image motion sequences containing a walker with, or without background motion clutter. After learning of the motion features, the detection performance of the model substantially increases, in particular in presence of clutter. The learned mid-level motion features are characterized by horizontal opponent motion, where this feature type arises more frequently for the training stimuli without motion clutter. The learned features are consistent with recent psychophysical data that indicates that opponent motion might be critical for the detection of point light walkers.

1 Introduction

Humans can recognize biological motion (e.g. human actions like walking and running) accurately and robustly; even from stimuli consisting only of a small number of illuminated dots that move like the joints of a human actor ("point light walkers") [6]. The neural mechanism that underlies the robust generalization from normal to point-light stimuli remains largely unclear. A possible explanation is that the brain extracts specific motion features that are shared by both stimuli classes. The nature of these features is unknown, and it has been discussed whether they are based predominantly on motion or form information [7]. In a recent study, combining methods from image statistics and psychophysical experiments, it was shown that robust recognition can be accomplished based on mid-level motion features [2].

Neurophysiological studies in monkeys and imaging studies in humans suggest that the perception of biological movements and actions involves both the ventral and the dorsal visual processing stream (see [5] for a review). A recent computational model

W. Duch et al. (Eds.): ICANN 2005, LNCS 3696, pp. 241 – 246, 2005.
© Springer-Verlag Berlin Heidelberg 2005

for biological motion recognition tries to account for a variety of the existing experimental data using relatively simple physiologically-plausible mechanisms [5]. The model is based on a feed-forward architecture which has been derived by extending a "standard model" (SM) for the recognition of stationary objects in the visual cortex [8]. Like other models for object recognition in the cortex [4, 8], our model represents complex movements in terms of learned prototypical patterns that are encoded by model neurons that respond to complex body shapes.

We apply in this paper a new biologically inspired algorithm, the "Memory Trace" (MeT) learning rule, to optimize model mid-level features for motion recognition. Originally the MeT algorithm was devised for the learning of mid-level in the context of the SM [9]. It has been demonstrated that by application of this learning algorithm the detection performance of the model for real-world stimuli could be substantially improved, resulting in performance levels which exceed the ones of several state-of-the-art computer vision systems for object detection [9]. Here we use the MeT algorithm in the context of a model for the recognition of biological movements and actions in order to optimize mid-level motion features for the detection of walkers.

Our paper first describes the model and the learning algorithm. We then present the results for the detection of walkers and show that learning of optimized mid-level motion features improves the performance, in particular in presence of motion clutter.

2 Methods

2.1 Model for Biological Motion Recognition

Our model corresponds to the motion pathway of the model in [5]. It consists of a hierarchy of neural detectors that are selective for motion features with different levels of complexity (fig. 1a). The first level of the model is formed by local motion energy-detectors whose responses are derived from the optic-flow fields of the stimuli assuming physiologically plausible tuning characteristics (see [5]). The model contains detectors for 70 x 43 different spatial positions and for 4 different directions. It turned out that for the feature learning it is critical that the outputs of the motion energy units are temporally smooth. We assume a simple linear low-pass filter with a time constant of τ = 228 ms, corresponding to the differential equation $\tau\dot{u}(t) = r(t) - u(t)$, where $r(t)$ is the motion energy signal and $u(t)$ the detector output. In the second layer, motion simple (MS2) units encode prototypical motion features of intermediate complexity. They combine the responses of the motion energy detectors with different direction preferences on the previous layer within a limited spatial region. These neurons are modeled by Gaussian radial basis function (RBF) units. The centers $\mathbf{c}_k$ of these RBFs are determined by the MeT algorithm. The responses of these neural detectors depend on the similarity of the local motion energy patterns from the present stimulus, that is given by the vectors $\mathbf{e}_k$, and these learned centers through the relationship: $x_k = \exp(|\mathbf{e}_k - \mathbf{c}_k|^2 / 2\sigma^2)$. For modeling position-invariant recognition, each mid-level motion detector is realized multiple times centered at different random spatial locations. Motion complex (MC2) units pool the responses of all mid-level motion detectors of the same type within a limited spatial receptive field using a MAXIMUM operation. The responses of these units are partially position-invariant. They define the input of a classifier that detects the presence or absence of a walker in the stimulus sequence. We tested different types of classifiers (cf. section 2.4).

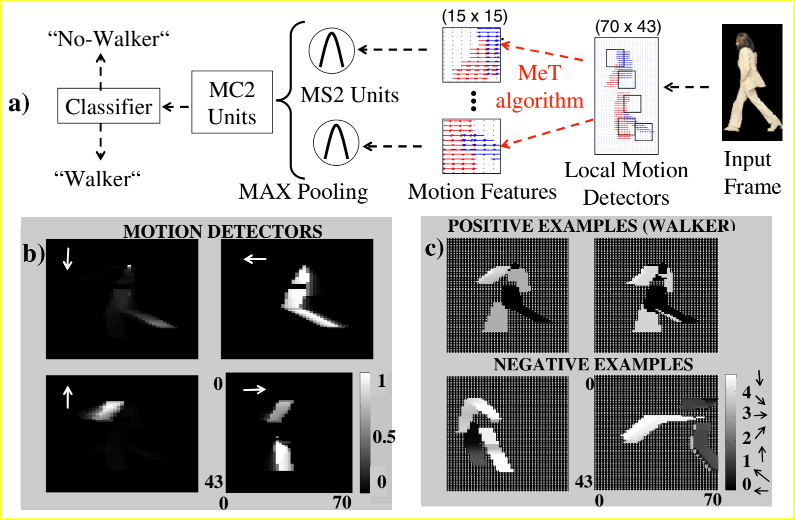

Fig. 1. a) Illustration of our model. See text for explanation. b) Activations of the local motion detectors tuned to 4 different directions (white arrows) for a walking frame shown as grey-coded maps. c) Optic-flows, with directions encoded as grey levels, for two positive (top) and two negative (bottom) examples. Zero motion energy is encoded by the dotted background (white dots on black).

2.2 "Walker-Detection" Task

The performance of our system was evaluated using a walker detection task. We used stimuli with uniform background, and with motion clutter. Stimuli were generated from five actors whose joint trajectories were tracked from videos (one gait cycle with 42 frames) [5]. The walking sequences of five different actors were used as positive examples, and other human actions (e.g. running, boxing, jumping) as negative examples. We selected randomly different sets of these sequences for training and testing the system. To introduce motion clutter for the same stimuli we added 100 moving squares (3x3) at random positions in each stimulus frame, defining random motion with uniform distribution of motion energy over the different directions.

2.3 Feature-Learning with the "Memory Trace" (MeT) Algorithm

Motion features with intermediate complexity were learnt using the MeT algorithm [9] (cf. Fig. 1a). The MeT algorithm is a biologically inspired mechanism for the unsupervised learning of frequently occurring features. The algorithm is inspired by previous work [3] that exploits a simple trace rule for the learning of shift invariance. Our trace rule assumes that the MS2 units keep record of their recent synaptic activity by an internal memory trace signal. In addition, it is assumed that the different features compete for the activations that a given stimulus produces. Successful activation of a feature results in an increase of its memory trace signal. Otherwise, the trace signal decays. Features whose memory trace falls below a fixed threshold are eliminated, and replaced by new features. New features are generated by choosing a randomly positioned local region in the visual field and taking the outputs of the motion energy detectors within this region for the present stimulus as feature vector. (See [10] for details). Learning is online since new features can be selected for each training step.

2.4 Classification Stage

To test the validity of the learned mid-level features for the detection of biological movements, we classified the outputs of the MC2 layer using different types of classifiers: 1) The "Maximally Activated Unit" (MAU) classifier that is biologically plausible. It corresponds to a radial basis function unit whose center is trained with the output signals from the MC2 level for the learned movements. If the activation of this unit is higher than a particular threshold the stimulus is classified as the particular action. Otherwise the classification result is negative. 2) k-Nearest Neighbor (k-NN), a standard technique for classification, was also implemented using RBF units whose centers were learned in the same way as the centers of the MAU classifier. During classification, the label of a test example is set to the label of the majority of the k nearest neighbors of the training set (we tested for $k = 1$ and $k = 5$). 3) Support Vector Machine (SVM) classifiers [13], as used in many recent machine vision systems (e.g. [9]). Although SVMs are not biologically plausible, they provide a typically well-performing classification back-end, which is useful to derive a measure for the quality of the learned features.

3 Results

Performances (Area Under the Receiver Operator Characteristic (ROC) curve) for all classifiers.are shown in Table 1 using the MeT algorithm (*MeT*) without and with motion clutter in the background (*Clutter*). For comparison we also show the results for stimuli in motion clutter when the mid-level features were defined by selecting randomly positioned regions from the stimuli (*Rand*)[1].

Fig. 2 (I) and (II) show the "best" features for the walker detection task for the simulations without and with motion clutter. An important observation is that many of these best features are characterized by horizontal opponent motion. The ROC curves for the three test conditions are shown in Fig. 2 (III). Performance after training with the MeT rule without clutter is almost perfect. This is not only true for state-of-the-art classifiers but also for simpler classifiers such as NN and MAU. Even in presence of clutter the MeT rule is significantly better than for randomly selected features. (The 5-NN outperforms SVM classifier, probably due to overfitting). This robust performance is consistent with recent results from the shape pathway [9]. It suggests a key

[1] Since we were interested mainly in recognition with cluttered background, we did not compare the MeT algorithm with a random selection of features for the other case.

role for plasticity in intermediate and higher visual areas of cortex for the realization of robust recognition.

Table 1. Performances (Area under the ROC) of the system for walker detection. Bold numbers indicate the classifier that gives the best performance for each experiment

	Performances (Area Under the ROC)			
Mode	**MAU**	**SVM**	**k-NN (k=1)**	**k-NN (k=5)**
MeT	0.977	0.999	**0.981**	0.962
MeT + Clutter	0.869	0.912	0.957	**0.972**
Rand + Clutter	0.726	0.795	0.876	**0.890**

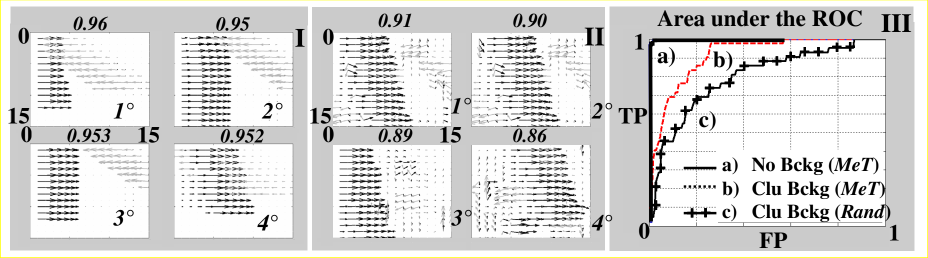

Fig. 2. I) Best four features for the stimuli without and with (II) clutter. Features were ranked according to the area under the ROC (numbers on top) computed separately for each individual feature. Features are plotted as optic-flow fields over the corresponding spatial windows. Black arrows indicate the values for local motion detectors that are selective for motion from left to right, and grey arrows indicate detectors selective for motion in opposite direction. The arrow length indicates the corresponding detector activation. III) ROC curves for the system with SVM classifier, for the MeT algorithm without (a) and with clutter (b), and for random selection of features (c).

4 Discussion

We have presented simulations using local learning rule for the optimization of mid-level motion features in a hierarchical model for the recognition of biological movements. The most important contribution of this rule compared to other approaches (e.g. [11]) relies in its neural plausibility. We found that learning of optimized mid-level features substantially improves the performance of the model, in particular in presence of motion clutter. Similar results have been obtained with a model for shape processing in the ventral pathway using the same learning rule. This suggests a key role of visual experience and plasticity throughout the whole visual cortex. Further work in this direction should implement neurally plausible mechanisms for the classification stage.

In addition, we found that for the detection of walkers, our algorithm learned optimized motion features that are characterized by horizontal opponent motion, for training with and without motion clutter. In principle, the same technique could be applied to optimize form features for the recognition of biological movements from body postures [2]. The importance of opponent motion features seems to be supported by psychophysical an imaging results that show that opponent horizontal motion might be a critical feature for the recognition of walkers, and degraded point light stimuli. Electrophysiological experiments indicate the existence of opponent motion-selective neurons, e.g. in monkey areas MT and MST [1, 12].

Acknowledgements. We are grateful to A. Casile for his support with the simulations. ARL is supported by the deutsche Volkswagenstiftung, DFG SFB 550, and HFSP. CBCL is sponsored by: Office of Naval Research (DARPA) Contract No. MDA972-04-1-0037, Office of Naval Research (DARPA) Contract No. N00014-02-1-0915, National Science Foundation (ITR/IM) Contract No. IIS-0085836, National Science Foundation (ITR/SYS) Contract No. IIS-0112991, National Science Foundation (ITR) Contract No. IIS-0209289, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218693, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218506, and National Institutes of Health (Conte) Contract No. 1 P20 MH66239-01A1.

References

1. Born, R. T. (2000). Center-surround interactions in the middle temporal visual area of the owl monkey. J. Neurophysiol. 84, 2658–2669.
2. Casile A, Giese M. (2005) Critical features for the recognition of biological motion, Journal of Vision, 5, 348-360.
3. Földiak, P. (1991). Learning invariance from transformation sequences, Neural Computation, vol. 3, pp. 194-200, 1991.
4. Fukushima, K.(1980) Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol.Cybern.36, 193–202.
5. Giese M A, Poggio T (2003): Neural mechanisms for the recognition of biological movements and action. Nature Reviews Neuroscience 4, 179-192.
6. Johansson, G. (1973) Visual perception of biological motion and a model for its analysis. Perc. Psychophys. 14, 201-211.
7. Mather, G., Radford, K., & West, S. (1992). Low-level visual processing of biological motion. Proc. R. Soc. Lon. B, 249(1325), 149-155.
8. Riesenhuber, M. & Poggio, T. (1999) Hierarchical models for object recognition in cortex. Nat. Neuroscience 2, 1019-1025.
9. Serre, T. & Poggio, T.(2005). Learning a vocabulary of shape-components in visual cortex, in prep.
10. Sigala, R. (2005) A Neural Mechanism to Learn Features of Intermediate Complexity in the Form and Motion Visual Pathway of Cortex. Thesis MSc. in Neural and Behavioural Sciences, MPI International Research School, Tuebingen, Germany.
11. Song, Y., Goncalves, L. & Perona, P. (2001) Unsupervised Learning of Human Motion Models. Advances in Neural Information Processing Systems 14, Vancouver, Cannada.
12. Tanaka, K., Fukuda, Y. & Saito, H. (1989). Analysis of motion of the visual field by direction, expansion/contraction, and rotation cells clustered in the dorsal part of the medial superior temporal area of the macaque monkey. J. Neurophysiol. 62, 626–641.
13. Vapnik, V. (1998) Statistical Learning Theory. John Wiley and Sons, New York, 1998.