

Neural field model for the recognition of biological motion patterns

M.A. Giese

Center for Biological and Computational Learning
Massachusetts Institute of Technology
45, Carletonstreet E25-206
Cambridge, MA 02139, USA
Tel.: 617 253 0549
FAX: 617 253 2964
E-mail: giese@mit.edu

Keywords: biological motion, recognition, neural field, prototype, learning

Paper contribution for the NC 2000,
May 23-26, Berlin, Germany

Neural field model for the recognition of biological motion patterns

Martin A. Giese

Center for Biological and Computational Learning, M.I.T., Cambridge, 45, Carletonstreet, MA 02139

Email: giese@mit.edu

Abstract

Neurophysiological research has revealed evidence that the recognition of *stationary* three-dimensional objects in the cortex seems to be based on neurons that encode prototypical two-dimensional views of the object. Much less is known about the neural mechanisms for the recognition of complex motion patterns, like biological motion and actions. This paper investigates if complex motion patterns can be recognized based on a similar neural principle, using dynamic neural networks to represent learned prototypical motion patterns. Based on this idea, a biologically plausible model for the recognition of biological motion is derived that is compatible with the known neurophysiological facts. The model combines neural mechanisms that have provided a valid account for neurophysiological data on the recognition of stationary objects with a recurrent neural network structure that can be most adequately analyzed in the mathematical framework of dynamic neural fields. Several simulation results are presented that show the computational feasibility of different plausible neural mechanisms, and that lead to a number of predictions that can be tested experimentally.

1 Introduction

Neurophysiological research in the last decade has revealed important insights in the physiological basis of the visual recognition of stationary objects [12]. A central principle that is supported by neurophysiological and psychophysical evidence is the representation of three-dimensional objects in terms of learned prototypical views. Poggio and Edelman [17] have postulated that cortical neurons exist which represent learned two-dimensional views of objects with tuning curves that decay gradually with increasing dissimilarity between the stimulus and the learned prototypical view. This gradual decay ensures good generalization of the representation. This is important to represent classes of similar views

of an object by interpolation between a small number of learned prototypical views.

Psychophysical evidence for this hypothesis was provided by demonstrating that the recognition of artificial three-dimensional objects (paper clips) is view-dependent and shows a gradual decrease of the recognition performance with the dissimilarity between test and training views [4]. Neurophysiological experiments by Logothetis *et al.* showed later that area IT of the macaque contains cells that can be trained to respond to individual views of a paper clip, and that have tuning curves that decay gradually with the orientation difference between test and training view of the paper clip [12]. To keep the number of prototypical views that must be stored as small as possible, it is important to base the recognition on features that are invariant, for instance, with respect to translation and scaling of the object in the visual field. Neurons in area IT show in fact substantial invariance against scaling and translation of the stimulus [12]. A simple neural mechanism to achieve such invariance was proposed by Riesenhuber and Poggio and verified by reproducing neurophysiological data from area IT in simulations [19]. The model assumes a hierarchy of feature detectors with increasing complexity of the detected features, and in parallel, an increasing invariance against translation and scaling from hierarchy level to hierarchy level. The invariance is achieved by pooling the responses of translation- and scaling-variant detectors on lower levels of the hierarchy using a nonlinear pooling operation.

Compared with this relatively detailed knowledge about neural processes that are relevant for the recognition of stationary objects, relatively few is known about the recognition of complex motion patterns, such as biological motion. The recognition of biological motion has been a popular topic in psychology several decades ago (see [9] for a review), but the interest has intermediately strongly decayed.

Very few neurophysiological results exist on the recognition of biological motion patterns and actions. Among those are the experiments by Perrett et al. who have found neurons in in the superior temporal sulcus that are sensitive to body shape and motion direction [15], or to gestures [16]. Such results seem to be corroborated by functional imaging studies showing activity during the perception of actions in similar areas in humans [3].

It seems, therefore, an interesting theoretical question whether the recognition of motion patterns can be based on neural principles that are similar to the ones that have successfully accounted for the recognition of stationary objects in area IT. The main purpose of this paper is to explore this idea.

In computer vision, the recognition of actions, gestures and facial expressions has been a popular theme during the last years. To our knowledge, there are only very few biologically motivated neural accounts for motion pattern recognition. The most similar work is a connectionist model by Goddard [8] which is not based on learning, and which is not directly linked to details of the neurophysiology of the visual system.

Different neurophysiological results might hint to mechanisms that might be relevant for the recognition of complex movement patterns. One result is the existence of detectors for complexly structured optic flow fields in area MST (e.g. [11]). Even though such neurons are usually interpreted in the context of ego-motion detection, such detectors might be also computationally useful for the analysis of complex motion patterns [6]. The model presented in this paper shows that this is the case. Neurophysiological results show that IT neurons are able to associate image patterns over time (e.g. [13]). Such phenomena are usually interpreted in the context of visual short- and long-term memory and have been associated with reverberatory activity in recurrent neural networks (e.g. [2]). The question arises if such dynamic neural mechanisms can also support the recognition of complex motion patterns. It is shown in this paper that this is the case, and that recurrent neural networks permit a discrimination between different complex motion patterns, and between patterns with different temporal order.

2 Model

Figure 1 shows an overview of the model. The model has two separate pathways for the analysis of shape and motion information. Beyond general neuroanatomical arguments, a strong functional argument for this separation is that biological motion can be recognized from motion information alone (as demonstrated by Johansson's point light walker) [9], and from shape information alone (e.g. from characteristic key frames from stick figure stimuli) [20].

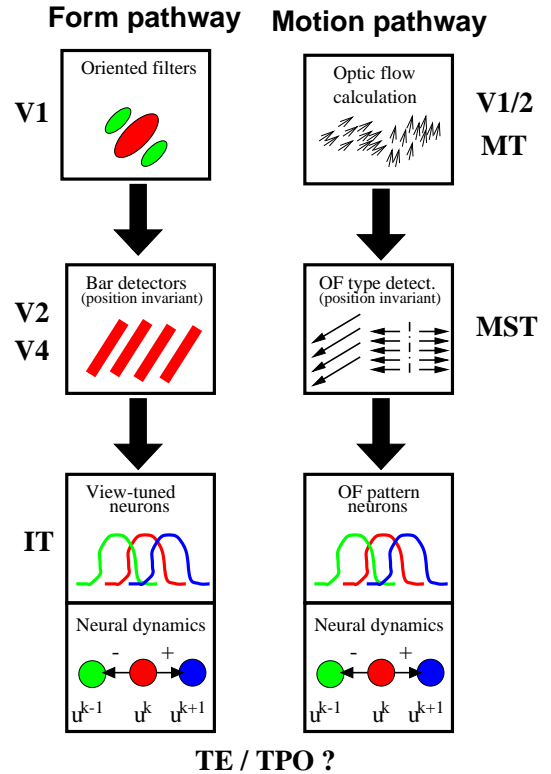


Figure 1: Overview of the Model

The postulated "form pathway" includes computational functions that are usually associated with cortical areas V1/2, V4 and IT. The "motion pathway" encompasses computational functions that are typically assigned to areas MT and MST, and potentially higher motion sensitive visual areas, like for instance area TPO in the superior temporal sulcus. In particular, the anatomical localization of the final neural recognition circuits is still relatively unclear, and parts of area IT and different higher areas in the superior temporal sulcus seem to be the most probable candidates.

2.1 Form pathway

The initial stages of the form pathway were modeled using a simplified version of the hierarchical model for stationary object recognition in [19]. The first stage of the form pathway consists of local oriented Gabor filters that model the properties of V1 simple cells [10]. The model contains 2056 symmetric Gabor filters with strongly overlapping receptive fields each of which covers about 4% of the simulated retinal area¹ and eight different preferred orientations. The receptive field centers are ordered within an equidistant rectangular grid.

¹The motion stimulus covers on average about half of this area.

The next stage of the form pathway contains position-invariant bar detectors, modeling the properties of V4 neurons (e.g. [5]). Other features, like edges or crosses, could be easily added, but sufficient selectivity for the discrimination of our test stimuli was already achieved with the bar detectors. The position-invariant bar detectors respond to bars with a certain orientation within their receptive field, independent of the spatial position of the bar within the receptive field. The receptive fields of these invariant detectors are much larger than the receptive fields of the Gabor filters, and cover about one quarter of the simulated retinal area. Position-invariance is achieved by a nonlinear pooling of the responses of of bar detectors without position invariance. (Scaling invariance can be achieved in the same way, but was not simulated for this paper.) The pooling is achieved by taking the maximum of the responses of the non-invariant detectors. In this way, it is possible to pool multiple responses of local non-invariant detectors without destroying the specificity of the pooled response for the features that are extracted by the local detectors. (For a detailed discussion see [19]). The model contains 144 invariant bar detectors with eight different optimal orientations.

The next processing stage is given by a radial basis function network that is trained with prototypical motion patterns which are specified by image sequences. The neural units represent individual frames from these image sequences. They encode thus the identity of the motion pattern, as well as time (because each motion pattern consists of multiple image frames). An individual cell is active when a particular image frame from the right motion pattern is present. The basis functions are gaussians, leading to a gradual decay of the response of the neural units when the dissimilarity between the stimulus and the learned prototypical image frame increases. The inputs of the basis function network are given by the responses of a subset of the invariant bar detectors. The subset is defined by the detectors that show significant variation of their response when different biological motion patterns are presented. A detector was treated as significant if the variance of its response over a set of training patterns, and over time, exceeded a certain threshold value. Let in the following the variables y_l specify the relevant detector outputs. The response of a basis function unit encoding frame number k of the prototypical motion pattern p is then given by the gaussian function:

$$s^{kp} = \prod_l \exp\left(-\frac{(y_l - y_l^{kp})^2}{2\alpha\sigma_l^2}\right)$$

The variables y_l^{kp} signify the bar detector responses when the k -th frame of the p -th prototypical motion pattern is presented as stimulus. σ_l is the output variance of bar detector l , and α is a positive constant. The output signals s^{kp} of the basis function units are time-dependent and

provide the inputs for a dynamic recognition network that is described in section 2.3.

2.2 Motion pathway

The functional architecture of the motion pathway is analogous to the form pathway. However, the underlying feature detectors are sensitive to complex optic flow features instead of shape features. The first stage of the motion pathway calculates the optic flow from the stimulus image sequence. This function is typically associated with cortical areas V1/2 and MT. The underlying processing was not modeled in detail. The model was tested using stick figure stimuli which permitted to calculate the associated optic flow directly on the basis of simple geometrical considerations.

The next stage of the motion pathway consists of detectors for different types of local optic flow fields² (expansion, contraction, and translation). These cells model the tuning properties of neurons in higher motion-sensitive cortical areas, like area MST (e.g. [11]). These detectors have large receptive fields extending over about one fifth of the simulated retinal area.

The model contains 72 translation detectors for four different directions and two different optimal speeds (fast and slow). The outputs of these detectors are given by the "motion energy" that is consistent with the detector specificity. To determine this energy, we determine the number of optic flow field vectors within a certain interval of angles and speeds. The second class of optic flow detectors are sensitive for contraction and expansion flow. The center of expansion or contraction is given by a line (cf. dashed-dotted line in the inset in figure 1). Such detectors exist for two different orientations of this center line (horizontal and vertical). MST cells that are sensitive to contraction or expansion flow show gradual invariance with respect to the position of the expansion center (e.g. [11]). To model this behavior, invariant detectors were again constructed by pooling the responses of non-invariant detectors for contraction and expansion using a maximum operation. The non-invariant detectors have two subfields that are adjacent to the center line. The subfield response is calculated by adding the speeds of the optic flow vectors within the subfield that point to (for contraction detectors), or away from the line (for expansion detectors). The response of the detector is then given by the square root of the product of the subfield responses³. The model contains 64 invariant contraction and expansion detectors.

²In neurophysiological literature, also cells sensitive for rotational flow have been described. From the inspection of the optic flow fields that were generated by our stimuli, it seemed that such rotation detectors would be only rarely adequately stimulated so that it was decided not to simulate them.

³The model is presently tested with a new MST detector model that is fitted in more detail to neurophysiological data

As in the case of the form pathway, significant features were determined by calculating the variances of the detector responses over a set of training examples. The last stage of the motion pathway is given by a radial basis function network that is trained in the same way as the network for form features. The output signal of this network serves as input of a second dynamical neural network that associates the different detected complex optic flow field patterns over time.

2.3 Dynamic recognition network

The neural architecture described so far permits to recognize shape and complex optic flow patterns at a single moment in time. For the recognition of biological motion, such information must be associated over longer periods in time. A simple neurally plausible mechanism for such a temporal association of information is given by a dynamic recurrent neural network that receives input from the radial basis functions, and that has asymmetric lateral connections. There are other possible mechanisms for a temporal association of information, and we are presently evaluating different alternatives. The mechanism that is described in this paper seems intriguing, because it has been claimed already that recurrent neural networks are relevant models for sequence effects that occur during the memorization of images in area IT (e.g. [2]). It seems therefore interesting to evaluate if similar network structures can support the recognition of image sequences. A further argument for the evaluation of this neural mechanism is that recurrent network structures of this type can be learned with a biologically plausible Hebbian learning rule, and that they account successfully neurophysiological data on the direction specificity of V1 neurons [18]. A final decision about the valid neural mechanism for temporal pattern association can probably not be made without detailed neurophysiological data, which is not yet available.

The proposed dynamic mechanism can be described in a mathematically convenient way by a modification of a *neural field model* was proposed by Amari [1]. In comparison with other recurrent neural network structures, this description permits a relatively simple analysis of the occurring dynamic phenomena. Interestingly, such Amari fields with symmetric lateral connections support memory solutions [1], which can in principle account for delay activity in area IT [13]. The assumption of asymmetric interactions leads to a new class of stable solution that are useful for the recognition of complex motion patterns.

The network dynamics that was used in the model can be described as follows: Each radial basis function unit is characterized by a dynamically changing output signal $s^{kp}(t)$. These signals provide input to a field of dynamic neurons that encode the presence of individual frame of

the prototypical image sequences. Let $u^{kp}(t)$ be the activation of the neuron that encodes the presence of the k -th frame of the p -th prototypical motion pattern. It is assumed in the following that the neurons are ordered along an axis with respect to the number k of the frame that they encode resulting in a one-dimensional *neural field*. If a learned prototypical image sequence is presented as stimulus the neurons that encode this prototypical pattern receive positive input from the corresponding basis functions. A *traveling wave* of activity arises in the neural field that encodes the prototype.

The neurons within the fields are coupled through lateral interactions. These interactions within the field have a well ordered structure and their strength is described by a convolution kernel $w(k)$ that has the form of an *asymmetric Mexican hat*. This lateral connectivity implies that, once a certain neuron is active, neurons encoding subsequent image frames of the prototypical motion pattern are pre-excited, whereas neurons that encode previous frames of this pattern are inhibited. The asymmetric lateral interactions lead to a strong dependence of the amplitude of the activation in the neural field on the order of the presentation of the stimulus frames. If a prototypical image sequence is presented in the right temporal order, a stable wave of activation is arising in the neural field. In this case, the lateral interaction stabilizes the propagating activation wave. A mathematical analysis that exceeds the scope of this article shows that this wave corresponds to a *stable traveling pulse solution* of the neural field dynamics. If the stimulus frames are presented in wrong temporal order the activation in the field is strongly suppressed and the a stable traveling pulse solution can not be formed. In this case, the lateral interaction does not stabilize the spatio-temporal activation patterns that are compatible with the input signal from the basis function network.

The mathematical form of the network dynamics is given by the following differential equation system:

$$\begin{aligned} \tau \dot{u}^{kp}(t) &= -u^{kp}(t) + s^{kp}(t) - h \\ &+ \sum_{k'} w(k - k') \theta(u^{k'p}(t)) \\ &- w_I \sum_{k' \neq k} A_{k'}(t) \end{aligned}$$

h and τ are positive constants that determine the time scale and the resting activity level of the neurons. $\theta(u)$ is a step threshold function. This nonlinearity is necessary to create actively propagating solutions. The last term describes a *mutual inhibition between networks that encode different prototypes*. The positive weight w_I determines the strength of this inhibition. A_k is the sum of the thresholded activity of the network that encodes the prototype k .

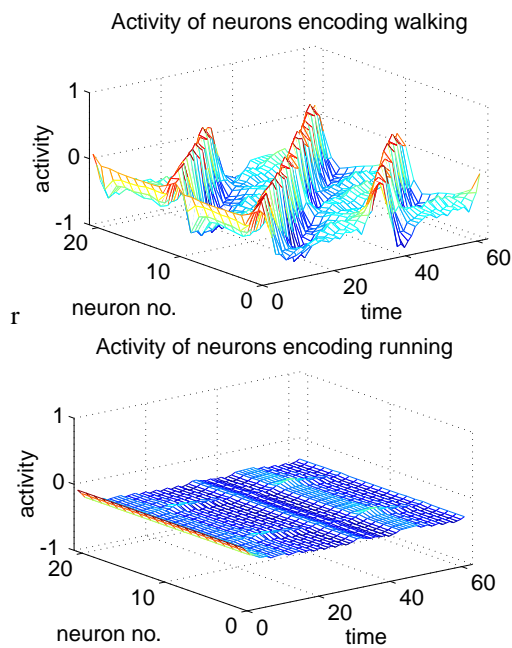


Figure 2: Activity in the "walking" and "running" field as function of time for a walking stimulus

3 Simulation results

In this section a selection of simulation results is presented. The model was tested with stimuli that showed a stick figure performing three different types of biological motion: walking, running and limping. The image sequences were generated using MATLAB. These three motion patterns could be easily distinguished by human observers, and also could be classified correctly with a computer vision algorithm [7]. Figure 2 shows the neural fields that represent the "walking" and the "running" prototype in the form pathway during the presentation of a walking stimulus. The upper part of the figure shows

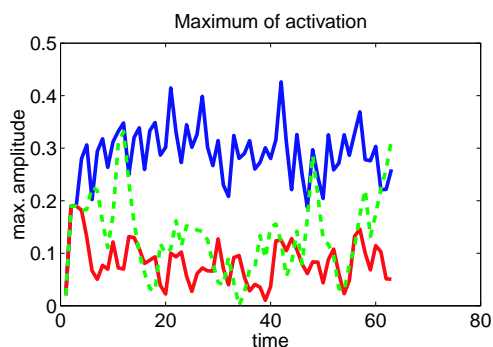


Figure 3: Maximum activation of solution for walking, walking with inverse temporal order and with scrambled temporal order

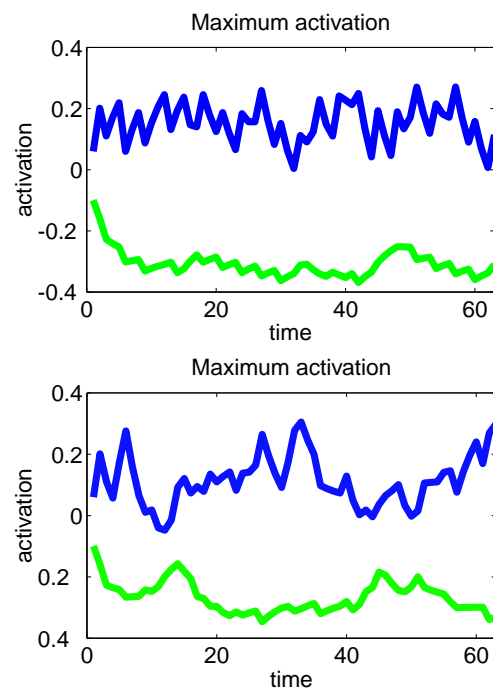


Figure 4: Activity in the "walking" and "running" field as function of time for a walking stimulus and a time-warped walking stimulus

the traveling wave of activity that arises in the "walking" field. The lower panel shows that the activity in the "running field" is suppressed. If a running stimulus is presented the situation is reversed: activity arises in the "running field" and the "walking" field is suppressed. The same behavior is observed for the motion pathway. The proposed neural system permits thus the discrimination between different biological motion patterns.

Figure 3 shows the maximum activity in the "walking field" as a function of time in the form pathway for presentation of a regular walking stimulus (solid line), and two walking stimuli with changed temporal order (inverse temporal order: gray line, and random temporal order: dashed line). For both patterns with wrong temporal order the activity in the field is strongly suppressed, showing that the proposed recurrent neural mechanism permits to achieve a selectivity for the right temporal order.

Finally, Figure 4 shows that the neural system is able to generalize to patterns with different timing. The figure shows the maximal activity as a function of time for the "walking" and the "running" field the presentation of a time-warped walking stimulus. The speed of the motion was increased during the first part of the movement and slowed in the second part by approximately 20 % such that the cycle time of the movement remained constant. The top panel of the figure shows the maximum activity of the "walking" (black) and "running" field (gray) which

are strongly different. A discrimination between the two patterns is thus easily possible. The lower panel shows the same activities for the time-warped stimulus. The activation levels still are nicely separated such that the pattern can be correctly classified as "walking". Such generalization to patterns with moderately different timing is important to avoid the necessity to store many prototypes for patterns with slightly different timing.

The motion pathway tested with the same type of stimulus shows worse generalization performance. This makes intuitively sense, since changing the speed of the motion also changes the optic flow field features leading to a reduction of the outputs of the optic flow pattern detectors. This results in a reduced input to the dynamic recognition network, and by that to worse discrimination. The model predicts thus less generalization to time-warped stimuli for patterns that are defined by motion alone (like point light walkers).

4 Discussion

The theoretical study presented in this article shows that a recognition of complex motion and action patterns is possible on the basis of a neural encoding of prototypical motion patterns. A neural mechanism was presented that is consistent with the known neurophysiological facts. Additionally, the presented computational experiments show that recurrent neural networks provide one possible mechanism for achieving selectivity for temporal order. At the same time, such networks achieve generalization to similar patterns with slightly different timing. It was also demonstrated that complex local optic flow patterns, like the ones that are extracted in area MST, can be usefully exploited for the recognition of biological motion. These results provide information about the computational feasibility of different potential neural mechanisms that might contribute to the neural basis of visual action recognition.

The proposed model leads to a number of predictions that can be experimentally tested. Here only a few examples can be given: (1) The model postulates that humans can learn to recognize artificial non-rigid motion stimuli. After learning, it should be possible to measure generalization fields that decay gradually with the dissimilarity of the test pattern from the learned pattern in space and time. (2) The generalization gradient for patterns with different timing should fall off more rapidly for patterns that are defined by motion information alone than for patterns that are mainly defined by shape information. (3) The model postulates the existence of neurons with tuning properties that reflect these generalization properties. (4) The analyzed mechanism for temporal integration postulates the existence of strong and asymmetric lateral connections between motion-pattern-sensitive neurons. (5) The proposed recognition dynamics pre-

dicts non-linear temporal integration during the recognition of motion patterns. (Evidence for this point has been provided in [14].) The main focus of our present work is to test such predictions in psychophysical experiments in order to differentiate between different possible computational mechanisms for motion pattern recognition. Our simulations also serve to select suitable experimental questions that can be tested in psychophysical and neurophysiological experiments..

Acknowledgments

I thank T. Poggio, G. Rainer, and M. Riesenhuber for valuable comments. This work was supported by the Deutsche Forschungsgemeinschaft Gi 305 1/1. Work at CBCL is supported by Office of Naval Research contract No. N00014-93-1-3085, and National Science Foundation under contract No. DMS-9872936. Additional support is provided by: AT&T, Central Research Institute of Electric Power Industry, Eastman Kodak Company, Daimler-Benz AG, Digital Equipment Corporation, Honda R&D Co., Ltd., NEC Fund, Nippon Telegraph & Telephone, and Siemens Corporate Research, Inc.

References

- [1] S. Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27:77–87, 1977.
- [2] D. J. Amit. The hebbian paradigm reintegrated: Local reverberations as internal representations. *Behavior and Brain Sciences*, 18:617–657, 1995.
- [3] E. Bonda, M. Petrides, D. Ostry, and A. Evans. Specific involvement of human parietal systems and the amygdala in the perception of biological motion. *Journal of Neuroscience*, 16:3737–3744, 1996.
- [4] H. H. Bülthoff and S. Edelman. Psychophysical support for a 2D-view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences (USA)*, 89:60–64, 1992.
- [5] K. Cheng, T. Hasagawa, K. S. Saalem, and K. Tanaka. Comparison of neural selectivity for stimulus speed, length and contrast in the prestriate visual cortical areas V4 and MT of the macaque monkey. *Journal of Neurophysiology*, 71:2269–2280, 1994.
- [6] B. J. Geesaman and R. A. Anderson. The analysis of complex motion patterns by form/cue invariant MSTd neurons. *Journal of Neuroscience*, 16:4716–4732, 1996.
- [7] M. A. Giese and T. Poggio. Synthesis and recognition of biological motion patterns based on linear superposition of prototypical motion sequences. In IEEE, editor, *Proceedings of the MVIEW 99 Symposium at CVPR, Fort Collins, CO*, pages 73–80. IEEE Computer Society, Los Alamitos, 1999.

- [8] N. H. Goddard. *The perception of articulated motion: recognizing moving light displays*. PhD thesis, Department of Computer Science, Rochester, NY, 1992.
- [9] G. Johansson, S. S. Bergström, and W. Epstein. *Perceiving Events and Objects*. Lawrence Erlbaum, Hillsdale NJ, 1994.
- [10] J. P. Jones and L. A. Palmer. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58:1233–1258, 1987.
- [11] L. Lagae, H. Maes, S. Raiguel, D. K. Xiao, and G. A. Orban. Responses of macaque STS neurons to optic flow components: a comparison of areas MT and MST. *Journal of Neurophysiology*, 71:1597–1626, 1994.
- [12] N. K. Logothetis, J. Pauls, and T. Poggio. Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5:552–563, 1995.
- [13] Y. Miyashita. Inferior temporal cortex: where visual perception meets memory. *Annual Review of Neuroscience*, 16:245–263, 1993.
- [14] P. Neri, M. C. Morrone, and D. C. Burr. Seeing biological motion. *Nature*, 395:894–896, 1998.
- [15] M. W. Oram and D. I. Perrett. Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey. *Journal of Neurophysiology*, 76:109–129, 1996.
- [16] D. I. Perrett, M. H. Harries, R. Bevan, S. Thomas, P. J. Benson, A. J. Mistlin, A. J. Chitty, J. K. Hietanen, and J. E. Ortega. Frameworks of analysis for the neural representation of animate objects and actions. *Journal of Experimental Biology*, 146:87–113, 1989.
- [17] T. Poggio and S. Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266, 1990.
- [18] R. P. N. Rao, M. S. Livingstone, and T. J. Sejnowski. Direction selectivity from predictive sequence learning in recurrent neocortical circuits. *Society for Neuroscience Abstracts*, 25:1316, 1999.
- [19] M. K. Riesenhuber and T. Poggio. A hierarchical model for visual object recognition. *Nature Neuroscience*, 11:1019–1025, 1999.
- [20] J. T. Todd. Perception of gait. *Journal of Experimental Psychology: Human Perception and Performance*, 9:31–42, 1983.