massachusetts institute of technology — artificial intelligence laboratory

Biologically Plausible Neural Model for the Recognition of Biological Motion and Actions

Martin Alexander Giese and Tomaso Poggio

AI Memo 2002-012 August 2002
CBCL Memo 219

Abstract

The visual recognition of complex movements and actions is crucial for communication and survival in many species. Remarkable sensitivity and robustness of biological motion perception have been demonstrated in psychophysical experiments. In recent years, neurons and cortical areas involved in action recognition have been identified in neurophysiological and imaging studies. However, the detailed neural mechanisms that underlie the recognition of such complex movement patterns remain largely unknown. This paper reviews the experimental results and summarizes them in terms of a biologically plausible neural model. The model is based on the key assumption that action recognition is based on learned prototypical patterns and exploits information from the ventral and the dorsal pathway. The model makes specific predictions that motivate new experiments.

We thank I. Bülthoff, E. Curio, Z. Kourtzi, M. Riesenhuber, T. Sejnowski, P. Sinha, I. Thornton, and L. Vaina for very useful comments. We are grateful to A. Benali, Z. Kourtzi, and C. Curio for help with the data acquisition, and to the Max-Planck Institute for Biological Cybernetics, Tübingen, for providing support.We thank M. Fitzgerald for help with the final layout.

This report describes research done within the McGovern Institute and the Center for Biological & Computational Learning in the Department of Brain & Cognitive Sciences and in the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology. M. Giese was supported by the Deutsche Forschungsgemeinschaft, Honda R&D Americas Inc., and the Deusche Volkswagen Stiftung. T. Poggio is supported in part by the Whitaker chair.

Research at CBCL was sponsored by grants from: Office of Naval Research (DARPA) under contract No. N00014-00-1-0907, National Science Foundation (ITR) under contract No. IIS-0085836, National Science Foundation (KDI) under contract No. DMS-9872936, and National Science Foundation under contract No. IIS-9800032. Additional support was provided by: Central Research Institute of Electric Power Industry, Eastman Kodak Company, DaimlerChrysler AG, Compaq, Komatsu, Ltd., NEC Fund, Nippon Telegraph & Telephone, Siemens Corporate Research, Inc.

Complex motion patterns, such as biological movements or actions, are biologically important visual stimuli. They are useful for recognizing predators from large distance, and are important for the courtship behavior in multiple species. Many predators exploit complex movements for detecting their prey, and for minimizing their own risk during attack by selecting weak animals (*1*). Gestures and facial expressions play a central role in the communication behavior of primates and humans *(2, 3)*. In spite of the high biological relevance of such stimuli, a detailed, biologically plausible theory for the neural mechanisms that underlie their recognition is still lacking. We provide in this article a review of the most important experimental results and present a model that consistently accounts for them on the basis of well-established cortical neural mechanisms. From this theoretical analysis a number of questions arises that motivate new psychophysical, neurophysiological and functional imaging experiments.

Neural Basis of Motion Recognition: Review of Some Basic Results:

Classical psychophysical experiments by Johansson have demonstrated that complex movement patterns, such as walking or dancing, can be recognized from highly impoverished stimuli that consist only of a small number of moving dots. Such "point light displays" can be generated by filming illuminated markers that are fixed on the joints of actors who perform complex movements (*4*). Subsequent studies have shown that the same stimuli are also sufficient for the recognition of other complex actions, such as American Sign Language and facial expressions (*5*). The recognition of biological motion from point light displays is highly selective, and subjects are able to identify the actor and the gender of the walking person on the basis of such stimuli (*6*). Since action patterns can be recognized from point light stimuli and strongly blurred movies (*7*) well-defined shape information seems not to be necessary for the recognition process. Other psychophysical studies suggest that gait patterns in some cases can be identified on the basis of form features in individual stationary images, or from stimuli with strongly degraded motion information. This implies that pure form information sometimes can be sufficient (*8*).

A variety of computer vision methods for the recognition of complex movements and actions exist. This shows that the recognition problem is computationally solvable, even though a solution at the level of the human performance is still out of reach (*9*). Many approaches for the analysis of human movements rely on predefined or learned geometrical models of the human body, or body parts, combined with predictive filtering techniques. Other approaches rely on the analysis of the space-time patterns using flexible templates, exploiting methods from texture analysis, or by learning probabilistic models, like Hidden Markov Models. Some methods exploit form features, such as edges and colored patches. Others are based on the analysis of optical flow patterns (*10*). Most technical algorithms are not biologically plausible, since it seems difficult to imagine how they could be implemented neurally. Somewhat closer to structures that have plausible biological implementations are solutions based on neural networks or connectionist models (*10, 11*).

It is an open question whether the amazing properties of biological motion recognition can be accounted for with known neural mechanisms of cortical information processing. Only few neurophysiological results are available: Different parts of the superior temporal sulcus (STS) contain neurons that respond selectively for full-body (*12*) and hand movements (*13*). Many of these neurons show view-dependent responses: the same action elicits much smaller neural responses if it is presented from a viewing direction that differs from the preferred view of the neuron. A significant fraction of neurons in STS show strong responses for point light stimuli. Neurons selective for the visual perception of actions have also been found in area F5 of the premotor cortex of monkeys, an area that has been compared with Broca's speech area in humans (*14*). Neurons in this area respond selectively during the observation of actions, like grasping. Such responses show invariance against the effector position, e.g. the distance of the hand from the body. Some neurons even generalize over different ways of performing the same action, e.g. grasping with the hand or the mouth. The most significant property of such neurons is that they respond not only when the monkey observes an action performed by another actor, but also when the

monkey itself performs the action. It has been postulated that such „mirror neurons“ are fundamental for linking perception and motor planning, and for the learning of movements by imitation (*14*).

Functional imaging studies have suggested the existence of similar neural structures in humans (*15*). Activation of areas in the STS during observation of biological motion has been reported in PET and fMRI experiments using point light displays and natural stimuli for full-body motion (*16, 17*). Activation of such areas was also found for mouth and hand movements and facial expressions (*3, 18-20*). An analog of area F5 in humans has been reported as well. This area in the *inferior frontal gyrus* shows selective activity during grasping and during the observation and imagery of hand and body movements (*15, 19*). Biological movement stimuli have been reported in some studies also to activate other sites, like the *fusiform gyrus*, the supplementary motor area, the *amygdala,* and the *cerebellum* (*16, 17, 20*).

Model:

Despite a growing body of experimental results from single unit recordings and functional imaging there are almost no theoretical proposals of biologically plausible neural mechanisms for the recognition of complex biological movements. We have developed a model that consistently summarizes many existing results and simultaneously provides a plausibility proof that the recognition of complex biological movement patterns might be based on relatively simple, well-established neural mechanisms. In addition, our model shows that recognition of such patterns can be achieved with a limited number of learned prototypical motion patterns. This representational principle is analogous to the encoding of stationary three-dimensional shapes in terms of learned “prototypical views.” View-based encoding has been a fruitful approach for the study of stationary object recognition in primates (*21*).

The basic structure of the model is illustrated in Fig. 1A. Consistent with the known functional division of the visual cortex in a ventral and dorsal processing stream (*22*), the model contains two pathways that are specialized for the analysis of form and optic flow information. We wanted to test with our computational model whether the recognition of complex biological movements and actions can be based purely on form or on optic flow information. To test this hypothesis we made the simplifying assumption that, at least to first order, the two pathways are not coupled before the level of the STS. The fusion of both processing streams in the brain occurs likely at the level of the STS (*23*). This fusion can be easily integrated in the model and leads to an improvement of the robustness of the recognition (*24*).

Both pathways consist of hierarchies of neural feature detectors. Hierarchical models for the ventral pathway have been proposed repeatedly to account for stationary object recognition (*25, 26*). Along the hierarchy the size of the receptive fields of the detectors, as well as their invariance against scaling and translation of the stimuli increases gradually. This assumption is consistent with the physiological properties of neurons in both the ventral and dorsal pathway. In addition, recent theoretical studies have demonstrated that hierarchical feed-forward networks with such gradual increase of feature complexity and invariance can account for highly selective pattern recognition with substantial degrees of invariance. Such models account in particular for the neurophysiologically measured invariance properties of neurons in the object recognition area IT of monkeys (*26*). Similar to other models for the recognition of stationary objects (*25*), invariance for translation and scaling is achieved in our model by pooling the responses from non-invariant neural detectors over multiple spatial positions and scales using a maximum-like operation (*26*). Pooling by a maximum operation, as opposed to pooling by linear summation, assures that the pooled response does not loose its selectivity for the original feature. In addition, pooling by a maximum operation makes the responses of feature detectors robust against background clutter (*26*). The maximum operation can be realized with simple, biologically plausible neural circuits (*27*). Preliminary physiological results suggest that this operation may be carried out by a subpopulation of complex cells in areas V1 and V4 (*28*).

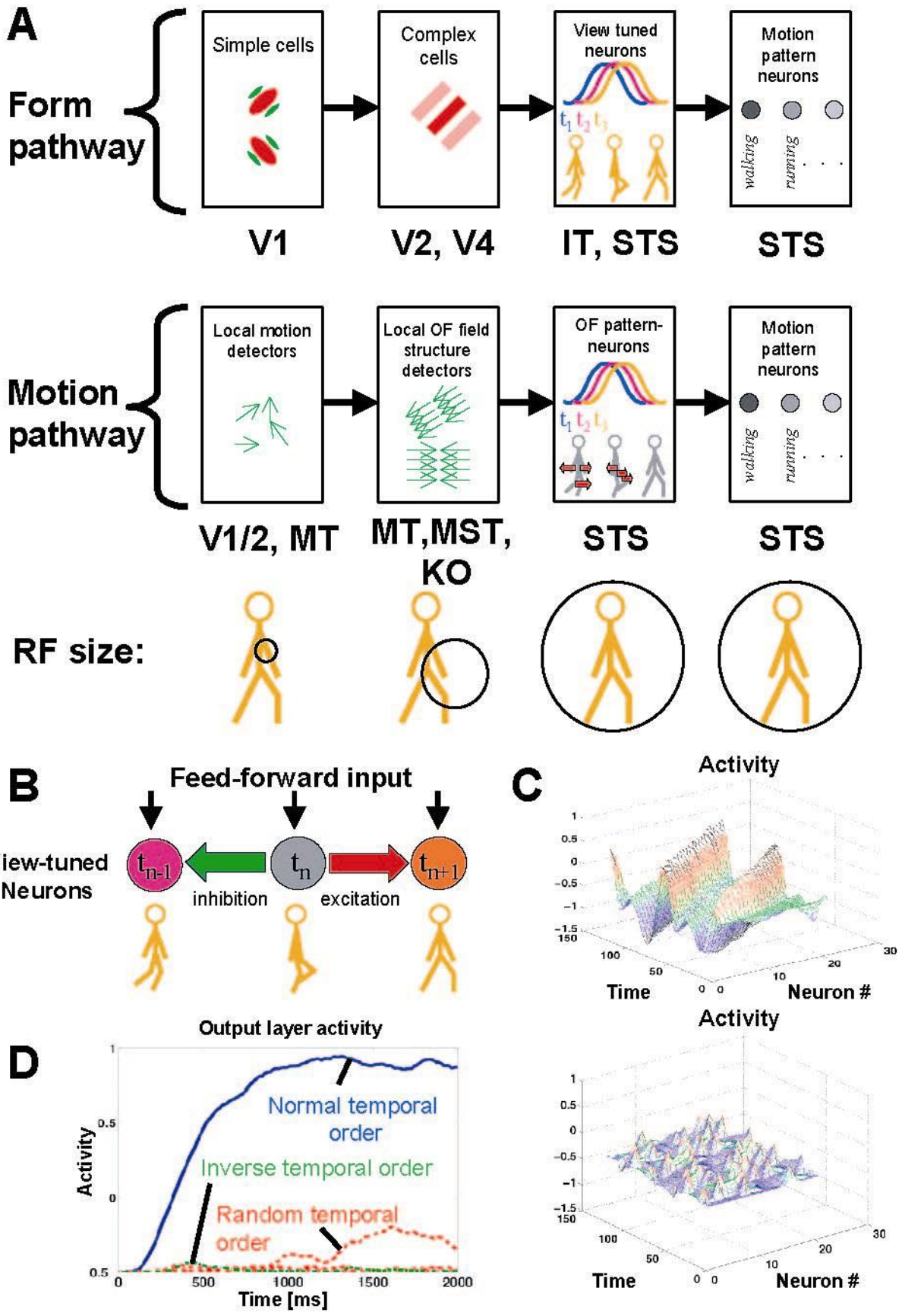
A
Form pathway
Simple cells
Complex cells
View tuned neurons
Motion pattern neurons
walking
running
V1
V2, V4
IT, STS
STS
Motion pathway
Local motion detectors
Local OF field structure detectors
OF pattern-neurons
Motion pattern neurons
V1/2, MT
MT,MST, KO
STS
STS
RF size:
B
Feed-forward input
View-tuned Neurons
inhibition
excitation
C
Activity
Time
Neuron #
D
Output layer activity
Activity
Normal temporal order
Inverse temporal order
Random temporal order
Time [ms]
Activity
Time
Neuron #

***Figure 1A**: Overview of the model with two pathways for the processing of form and optic flow information. Each pathway consists of a hierarchy of neural feature detectors with receptive field sizes illustrated by the circles in the bottom row. The hierarchy levels of the form pathway are formed by "simple cells", modeled by Gabor filters, complex cells that respond maximally for oriented bars independent of their exact spatial position, view-tuned neurons selective for body poses, and motion pattern-selective neurons that are selective for the whole movement sequence. The hierarchy levels of the motion pathway are motion (energy) detectors, detectors for local optic flow field patterns (translation, expansion, and contracting flow), neurons selective for complex instantaneous optic flow patterns, and motion pattern-selective neurons. The labels indicate also the areas in the macaque that contain neurons with similar properties. The highest levels of both pathways might not be separated in the brain. The Appendix provides more details of the circuitry and the properties of the cells.*

***Figure 1B**: Selectivity for temporal order is achieved by asymmetric lateral connections between neurons coding for different body configurations (or complex optic flow patterns). The activated neuron pre-excites other neurons that code for temporally subsequent configurations and inhibits neurons coding for configurations earlier or much later in the image sequence. (See Appendix for further details.)*

***Figure 1C**: A traveling pulse of activity in the representation is only stabilized by the recurrent neural network if the stimulus frames are presented in the right temporal order (upper panel). If the stimulus frames are scrambled in time (lower panel) the neural activity is strongly reduced.*

***Figure 1D**: Test of the recurrent network with movies generated from a video sequence of a walking person: The stimulus shown in the original temporal order (solid blue line) leads to a fast build-up of activity in the form pathway (latency < 200 ms) after stimulus onset. Destroying the temporal order by either presenting the frames in random order (broken red lines), or by playing the movie in reverse order (broken green line) leads to a substantially reduced activity in the neurons representing the walking movement.*

The form pathway achieves recognition of actions based on form features. In the model, neurons on the first level of the form pathway mimic V1 simple cells responding selectively for local oriented contours. Neurons on the next hierarchy level, corresponding to complex cells in areas V2 and V4, are selective for bars independently of their exact spatial position and phase information. Neurons selective for more complex shape features, like corners, as observed in area V4 (29), could be easily added at this level, even though they are not necessary for replicating the experimental results discussed in this paper. The neurons at the next level of the form pathway are functionally equivalent to the "view-tuned neurons" that have been found in area IT of macaques. Such neurons can be trained to respond selectively to views of complex three-dimensional objects (*30*). We assume that these, or functionally similar neurons, can also learn to respond selectively for the particular configurations of the human body that are characteristic for actions and biological movements. Such learning might occur during the observation of movement sequences, but we did not model the underlying learning process. After learning, such neurons encode "snapshots" from image sequences showing complex body movements. The highest level of the form pathway consists of neurons that respond selectively for complete complex motion patterns, such as walking, running, boxing, or dancing. These motion pattern neurons summate and temporally smooth the activity of all snapshot neurons on the earlier level that code for snapshots of the same motion pattern. To keep the model simple we assumed that a single motion pattern neuron codes for each training pattern. This is a simplification, because in the brain multiple motion pattern neurons are likely to contribute to the representation to each motion pattern category (*31*).

The motion pathway analyzes optic flow information. The neurons at the first level of the motion pathway extract local motion energy, modeling direction-sensitive neurons in the primary visual cortex and area MT of the macaque (*32*). The neurons on the intermediate level of the motion pathway are selective for specific local optical flow field patterns: translation, and contraction / expansion along motion boundaries. In physiological experiments neurons with selectivity for similar local optical flow patterns have been found in areas MT and MST (*33*), and in area KO in humans (*34*). The neurons at the next higher level of the motion pathway are selective for instantaneous complex optical flow field patterns that arise during biological movements. The selectivity of these neurons is also learned from prototypical example movement patterns. These neurons are functionally equivalent to the "snapshot" neurons in the form pathway. In the brain such neurons might be located in different parts of the STS, and possibly also in the premotor cortex. The activity of these neurons is integrated by movement pattern-selective neurons at the highest level of the motion pathway that encode complex movement patterns, like walking or running. Motion pattern neurons of the form and the motion pathway may not be separated in the cortex. More details about the neural detectors on the different hierarchy levels of the model are described in the Appendix.

The recognition of complex movement patterns is selective for temporal order. Subjects who see a movie that shows a biological movement sequence in scrambled temporal order do not perceive the biological movement, even though the scrambled movie contains the same "snapshots" as the original sequence. Also, subjects can easily detect when a natural movement pattern is shown in reverse order (*35*). Multiple neural mechanisms can account for such sequence selectivity. In the model we assumed a mechanism that is based on asymmetric lateral connections between the "view-tuned" or optic-flow pattern-selective neurons *(36)*. Such asymmetric connectivity can be learned by a simple modified Hebbian rule *(37)*. Interestingly, strong effective lateral connectivity in area IT can also account for the experimentally observed memory and delay activity in this area *(38)*. Memory for stationary images and image sequences may thus be mediated by the same neural dynamics.

An experimental test of the different possible mechanisms for sequence selectivity requires detailed neurophysiological data. Fig. 1B illustrates the form of the lateral connections in the model. The network dynamics stabilizes a traveling activity pulse only if the stimulus frames are presented in the "right" temporal order. The effectiveness of this mechanism is illustrated in Figs. 1C and D. Scrambling or inversion of the temporal order of the stimulus leads to a competition between the stimulus input and the intrinsic dynamics of the network resulting in a strong reduction of neural activity. With the proposed neural mechanism a recognition of biological movement can be achieved within less than 300 ms. This is consistent with the psychophysical and neurophysiological observations indicating that biological motion stimuli can be recognized with presentation times as small as 200 ms, requiring stimulation for much less than a full walking cycle (*12, 39*). In addition, the postulated mechanism for order selectivity allows for substantial variations in the stimulus speed without abolishing recognition, consistent with psychophysical results (see Appendix for further details).

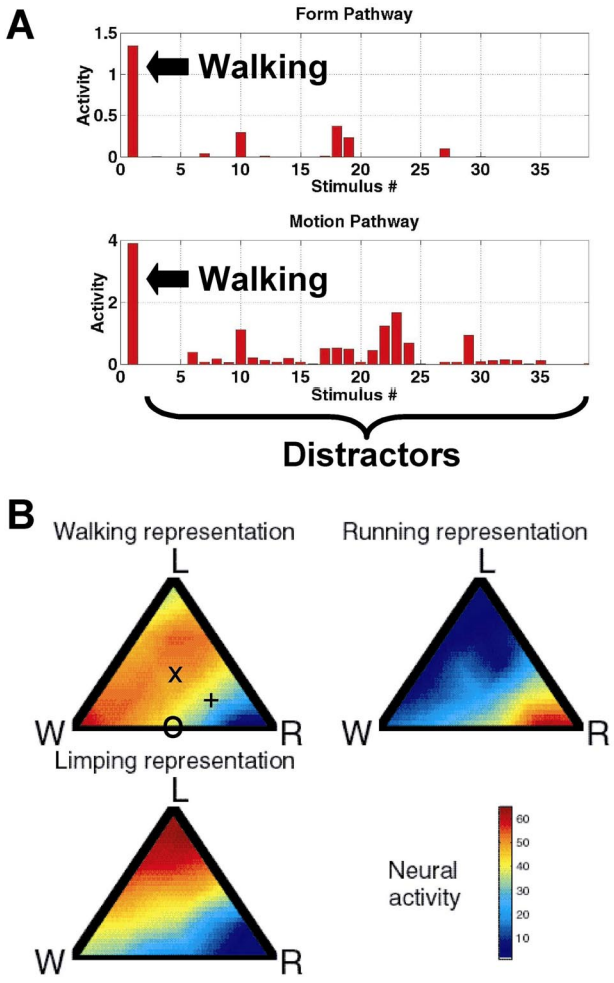

Figure 2A: *The proposed neural mechanism is characterized by high selectivity for motion patterns. The two panels show the activities of the "walking" neurons on the highest hierarchy levels of the form (top panel) and motion pathway (bottom panel). High activation arises only when "walking" is presented as stimulus. The other 38 distractor sequences, other locomotion patterns and complex movements from sports, elicit much smaller activities in the neurons that represent walking.*

Figure 2B: *Generalization fields that arise when the model, previously trained with three locomotion patterns (walking, running and limping), is tested with motion morphs between these patterns. The three panels show the color-coded activity in the motion pattern neurons in the form pathway that encode walking, running and limping. Morphs were created by linear combination of prototypical joint trajectories of walking, running and limping in space-time (see text). The positions of the pixels in the triangles code for the weights of the prototypes in the linear combination. The corners of the triangles indicate the pure prototypes (W: walking, R: running, L: limping). Points on the edges of the triangle represent two-pattern morphs between the adjacent prototypes, e.g. the point (o) is a morph between walking and running with equal weights 0.5. Points within the triangle are three-pattern morphs, where the ratios of the distances from the edge points are equal to the ratios of the corresponding weights. The center of the triangle, (x), codes for a morph with equal weights (1/3) of all patterns. Point (+) is a morph with 20 % walking, 20 % limping and 60 % running. The neural activity varies smoothly with the weights, consistent with the experimental results presented in (41). Comparing the generalization fields for the different motion prototypes depicted in the three triangles, it appears that walking is over-represented compared to the other patterns. Similar results were obtained for the model neurons in the motion pathway.*

Simulation Results and Predictions:

The model was tested using more than 100 video sequences, including different forms of human locomotion, dancing and physical exercises. We present in the following a few key results and predictions. Additional simulation results and additional details about the stimulus generation can be found in the Appendix.

A first set of simulations evaluates in how far the pattern selectivity and the generalization properties of the model match the properties of the biological system. Fig 2A illustrates that both pathways show high selectivity for different motion patterns.

The model was tested with 39 movement patterns that were presented as full-body stimuli (stick figures animated with tracking data from natural movements). The activity in the neural representation for walking is high only for walking test patterns. All other test patterns elicit only relatively small activities. Since the two pathways were not coupled this implies that the information in each pathway alone is sufficient for an accurate recognition of the presented complex motion patterns. A clear prediction follows from this property of the model. It should be possible to recognize biological movements by temporal association of stationary images, if they contain sufficient form information. This should be true even when optic flow cues are minimized, e.g. by using very long inter-stimulus intervals between the frames to suppress normal short and long-range motion perception. Psychophysical evidence consistent with this prediction was obtained with sequences of "Mooney" stimuli. Such stimuli consist of sequences of strongly degraded pictures showing body configurations of animals. Before subjects recognize the animals in the individual pictures the motion stimulus induces the percept of an incoherent optic flow. After subjects recognize the animals in the picture they could also perceive their biological motion, in spite of the seemingly incoherent optic flow information that is specified by such stimuli (*40*).

In spite of its high selectivity, the model predicts good generalization for natural complex movement patterns. The model tested with multiple video sequences showing the same type of locomotion executed multiple times, or by different people, classifies the motion patterns correctly unless the body geometries of the actors are very different (*see* Appendix). A more accurate quantification of the generalization properties of the model can be obtained with stimuli generated by motion morphing. We used a new morphing technique for the computation of linear combinations of complex movements in space-time (*41*). Three prototypical locomotion patterns, walking, running and limping, were used to train the model. The same prototypes were subsequently linearly combined with different weights to produce the motion morphs that we used as test stimuli. Fig. 2B illustrates the neural activities that are elicited by such morphs in the neural representations for walking, running and limping.

The neural activities vary smoothly with the weights of the prototypes in the linear combination. This prediction from the model is consistent with psychophysical data that we obtained by testing human subjects with the same stimuli: We found that multiple psychophysical measures, which can be assumed to covary with the activity of the neurons that are selective for locomotion patterns, changed smoothly with the weights that characterize the morphed pattern (*42*). Fig. 2B shows the predicted "generalization fields" for the individual prototypes. These are the regions in the pattern space of the motion morphs for which a particular prototypical pattern is perceived. The model predicts an over-representation of the pattern "walking". Interestingly, we found this prediction confirmed in our psychophysical experiments. The model also reproduces the experimentally measured speed-, and position invariance of the recognition of complex movements in humans. In addition, the model is sufficiently selective to identify people by their locomotion patterns (*6*). The corresponding simulations can be found in the Appendix.

An important prediction follows naturally from the assumption of a representation in terms of learned snapshots and instantaneous optic flow patterns. Biological motion recognition should be view-dependent. Such view-dependence is psychophysically well-documented. The recognition of the

movements of point light walkers is strongly impaired when the stimulus is rotated in the image plane, or in depth (*43*). The same behavior is predicted by the model, as illustrated in Fig. 3A. Rotation of the stimulus in depth strongly reduces the response of the motion pattern sensitive neurons. A similar reduction was observed for neurons in area TPO that responded selectively for "walking" if the direction of the walker was rotated against the preferred direction of the neuron (*12*). For 2D rotations of the stimuli in the image plane against the training views similar reductions of the neural activity are observed in the model. These results are consistent with recent fMRI experiments (*44*).

A second set of simulations tests whether the model reproduces the high degree of robustness that has been observed in psychophysical experiments on biological motion recognition. Biological motion stimuli can be strongly degraded before recognition suffers. One example is the recognition of point light displays, which humans and possibly also other animals (*45*) can easily recognize, even if they have never seen such stimuli before. The model predicts the same type of robustness without assuming complex computational processes, like a reconstruction of the links between the dots or of the structure of the articulated figure. Fig. 3B shows the activities of the motion pattern neurons that have been trained with a normal full-body "walking" stimulus.

Activities are shown for the form and the motion pathway for three different stimuli: a normal walking stimulus, a point light walker, and a distractor stimulus (running). The point light stimulus does not result in substantial activity in the form pathway. In the motion pathway, however, it induces significant activation. The low response for the distractor pattern shows that this activity is selective for a specific biological motion. The generalization from normal stimuli to point light stimuli in the model is accounted for by the similarity of the optical flow fields of the normal and the degraded normal stimulus. This leads to another prediction that may be tested experimentally: Point light stimuli showing learned complex movements should elicit selective activation in the dorsal, but not in the ventral pathway. In fact, stimuli can be degraded even more without abolishing recognition of the model. Consistent with psychophysical results, the model can even recognize point light stimuli when the individual dots have a limited life time (*46*).

Point light stimuli can be degraded also by removing individual dots from the figure. Depending on the missing joints, the recognition performance is more or less reduced. This leads to another interesting prediction of the model. Removing elbows and feet has been shown to be most harmful for recognition in psychophysical experiments (*47*). The model predicts the same result (Fig. 3C). The fact that removing the elbows is very harmful for recognition, rules out trivial explanations based on the maximum speed of the stimulus dots. In the model, the critical factor is the similarity between the rudimentary optic flow field generated by the point light stimulus and the optic flow templates that are encoded in the motion pattern neurons that have been learned from full-body walker stimuli.

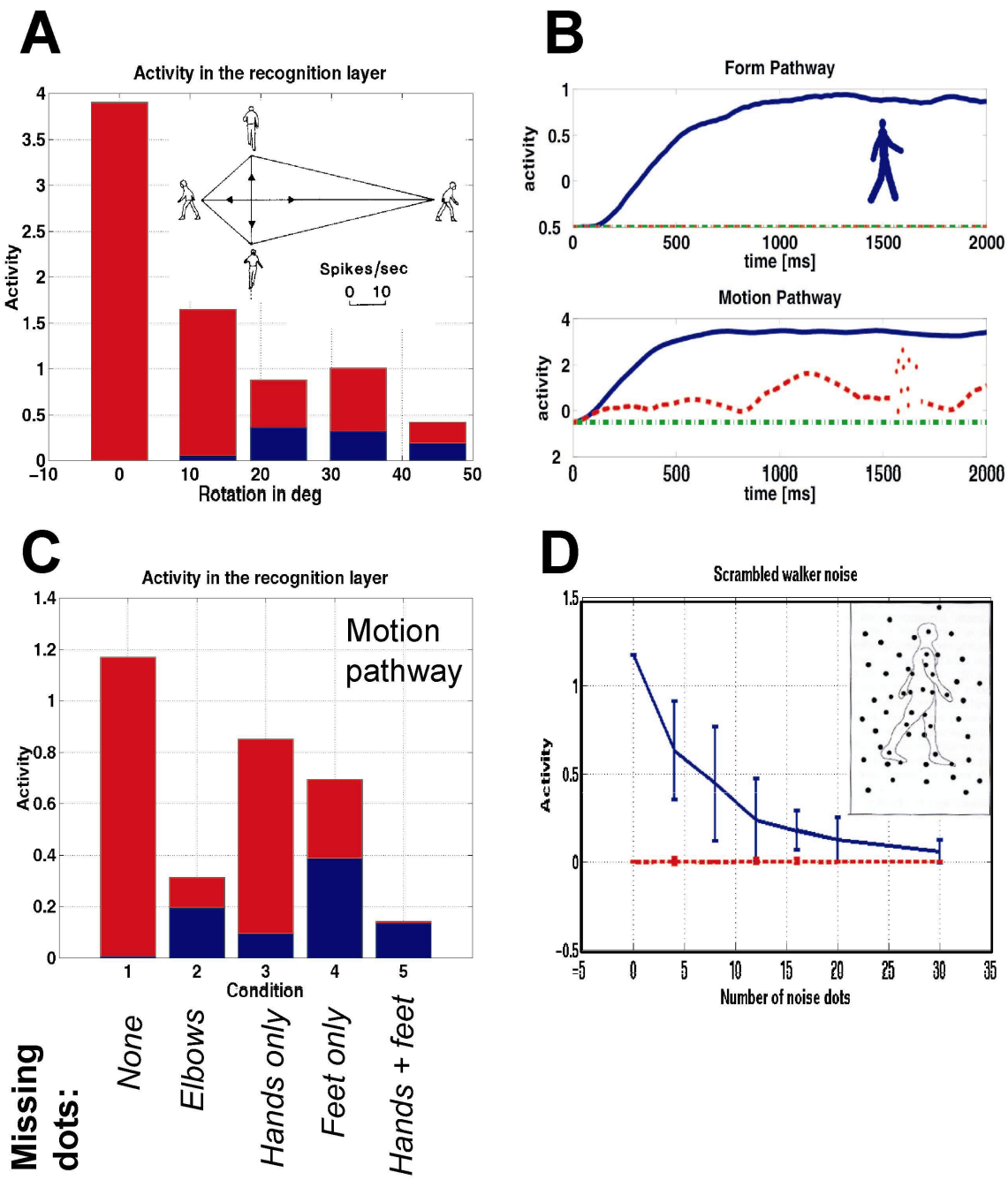
A
Activity in the recognition layer
Activity
Rotation in deg
Spikes/sec
0 10
B
Form Pathway
activity
time [ms]
Motion Pathway
activity
time [ms]
C
Activity in the recognition layer
Motion pathway
Activity
Condition
Missing dots:
None
Elbows
Hands only
Feet only
Hands + feet
D
Scrambled walker noise
Activity
Number of noise dots

***Figure 3A**: Activity in the neural representation for walking, in the form pathway, when the stimulus is rotated in depth against the learned view of the walking pattern. The neural activity decreases strongly with the rotation angle. The red bars indicate the activities of the motion pattern neuron for walking. The blue bars indicate the maximum of the activities of the other motion pattern neurons that are encoding running limping and marching. The inset shows the activity of a neuron from area TPO in the STS, measured electrophysiologically [reproduced with permission from (12)] that shows similar angle dependence. Equivalent results were obtained for the neurons in the motion pathway of the model.*

***Figure 3B**: Activities of the motion pattern neurons for walking in the form and motion pathway for three different stimuli: a full-body stimulus for walking (blue solid lines), walking presented as point light display (red dashed lines), and a full-body stimulus for running (green dashed-dotted lines). In the form pathway only the full-body stimulus for walking elicits high activity. The point light display and the distractor do not result in significant activation. In the motion pathway the point light stimulus induces significant activity that is much stronger than the activity induced by the distractor sequence.*

***Figure 3C**: Decay of the activity in the motion pattern neuron for "walking", in the motion pathway, for degraded point-light stimuli. The dots of different joints were removed from the stimulus. Red bars indicate the activities of the neural representation for walking and blue bars indicate the maximum of the activities of the motion pattern neurons for running, limping, and marching.*

***Figure 3D**: Activity in the motion pathway for the motion pattern neurons that encodes "walking to the right." The stimuli are point light walkers that are degraded by background noise. The noise was created by "scrambling" the sequence of a point light walker (see text). The solid blue line shows the activity for a walker walking to the right, and the broken red line the activities for a walker walking to the left. The error bars indicate the standard deviations over 20 repeated simulations. Discrimination between walking to the left and to the right side is possible if the two activities are significantly different. This is the case for up to 30 noise dots, a number that exceeds the number of the point light walker by a factor of three.*

Further experimental evidence for the high robustness of action recognition was obtained in masking experiments. In such experiments a large number of moving background dots can be added to point light stimuli without significantly impairing recognition (*48*). Such robust behavior can be observed even when the background dots are created by "scrambling" point light walkers. In this case the masking dots have exactly the same movements as the dots in the moving figure, but different average positions. High robustness against masking is predicted by our model. Figure 3D shows the activity of the motion pattern neurons trained with a standard "walking rightwards" stimulus and tested with point light stimuli moving rightwards and leftwards in presence of different amounts of background noise. Even if the number of noise dots exceeds the number of dots of the point light walker by a factor of three, the activation levels for the rightwards and leftwards walking are still significantly different. This implies that a right-left discrimination should be possible in spite of the presence of substantial motion clutter. The high robustness of the model might result, at least partially, from the suppression of clutter by the nonlinear pooling operation.

Because the model tries to reproduce coarsely the structure of the visual cortex, another interesting set of predictions can be derived by reading out the time-averaged neural activity from the different hierarchy levels of the model's pathways. Such predictions seem (under some assumptions) appropriate for a comparison with data from functional imaging experiments. For this comparison, we loosely assigned the layers of the neural model to different areas in visual cortex (*49*). An obvious prediction that is consistent several experimental results (*16-18*). is that for stimuli of similar type (full-body or point light) the activity in the lower areas in the two pathways does not show clear discrimination between biological and non-biological motion (dots moving randomly). A differentiation between biological and non-biological

motion appears at the level of the model area STS. More specific predictions can be derived by comparing the activities for different types of point light stimuli in area STS. For instance, an upright point light walker stimulus induces significantly higher activity than an "inverted" point light walker, which has been rotated by 180 degrees. The activity elicited by an inverted walker exceeds the activity arising for a scrambled point light walker or a stimulus with the same number of randomly moving dots. This prediction is quantitatively consistent with the results of a recent FMRI study (*44*). Further details about these predictions are reported in (50).

Conclusions:

The recognition of complex movements is an important perceptual function, not only because it is fundamental for communication and recognition at a distance, but also for the learning of complex motor actions by imitation (*14*). An analysis of the neural basis of complex movement recognition seems timely. On one hand, there is a growing body of experimental results. On the other hand, recent work in computational vision on motion pattern and stationary objects recognition has increased our understanding of the underlying computational problems and plausible neural mechanisms required for their solution.

The current model shows that several principles that are central for the recognition of stationary objects might be important also for the recognition of complex motion patterns. The first principle is a representation in terms of learned prototypical patterns *(51)*. The model shows that the experimentally observed generalization properties of the perception of biological movements are quantitatively consistent with this hypothesis. The second principle is a neural architecture that consists of hierarchies of neural detectors with gradually increasing feature specificity and invariance. This architecture seems to be adequate to account for the invariance properties with respect to stimulus position, scaling, and speed that are characteristic for recognition of biological motion. An important additional assumption in our model is the existence of recurrent neural network structures that associate sequential information over time. This assumption leads to predictions that can be physiologically tested, such as the existence of asymmetric lateral connections between motion pattern selective neurons.

A key prediction of the model is that any sufficiently smooth complex movement pattern can be learned, independently of whether it is compatible with the motion of a biological organism or the physical rules of articulated moving objects. The only restriction is that the form and the optical flow features of the stimulus must be adequate for the activation of the neural detectors in the earlier levels of the ventral and the dorsal pathways. Another prediction that can be tested electrophysiologically, and with functional imaging methods, is that area IT, thought to be important for the recognition of stationary objects (*30*), may be also involved in the recognition of biological movements, potentially through the representation of „snapshots" from motion sequences (*52*). An experimental verification of this prediction might be possible by testing IT neurons with stationary images of body configurations that are embedded in image sequences that are compatible or incompatible with biological movements (*53*). The prediction that the recognition of biological movements is possible with the information from each pathway alone is consistent with clinical results showing that patients with lesions that include either only the human equivalent of are IT, or the MT/V5 complex are still able to recognize complex biological movements when the STS is spared. Only bilateral lesions of the STS have been reported to lead to severe deficits in the perception of biological movements (*54, 55*).

The model presented in this paper is only a first-order approximation with the explicit goal to consistently summarize most of the existing data. It serves to provide a plausibility proof that a relatively simple, biologically plausible architecture can account for the recognition of complex motion patterns. Most importantly, the model makes predictions that suggest new experiments. We are aware that the model makes a number of strong simplifications. For instance, there is experimental evidence for substantial

attentional effects on the recognition of biological motion *(56)*. Likewise, experimental results demonstrate top-down influences or cross talk between the two pathways *(53, 57)*. It seems almost surprising that the skeleton feed-forward architecture described here is already sufficient to account already for a variety of the known experimental results.

Appendix:

Details about the Neural Detectors

Tab. 1 lists the most important properties of the neural detectors in the two pathways. In the form pathway oriented contours are extracted using Gabor filters with 8 different orientations and two spatial scales that differed by factor two. Gabor filters are well established as models for V1 simple cells (*58*). The responses of these filters were given by

$g(x,y) = \exp(-\mathbf{d'Kd}) \cos(\kappa d_1)$, with $\mathbf{K} = \mathrm{diag}(0.5\,[\sigma_1^{-2}, \sigma_2^{-2}])$, and $\mathbf{d} = \mathbf{R}(\phi)\,[x - x_0, y - y_0]$

where $\mathbf{R}(\phi)$ is a two-dimensional rotation matrix, and where ϕ defines the preferred orientation. The vector $[x_0, y_0]$ defines the center of the receptive field. The receptive fields were arranged within an equidistant quadratic grid. The chosen parameters values were $\sigma_1 = 10$, $\sigma_2 = 7$, $\kappa = 0.35$ for the small spatial scale, and twice as large for the larger spatial scale. The output signals of the Gabor filters were thresholded using a linear threshold function before they were transmitted to the next higher hierarchy level.

To calculate the responses of the invariant bar detectors, the responses of the V1 neurons were pooled separately for each orientation within a limited spatial region, and over the two spatial scales using a maximum operation. The pooled responses model the responses of complex cells in areas V1, V2 and V4 that are known to be invariant against the spatial phase of the stimulus (*59*). The receptive field diameters of these neurons were about four times larger than the receptive fields for the neurons on the first hierarchy level, consistent with neurophysiological results from area V4 in macaques (*60*). In our simulations invariant bar detectors for different orientations were sufficient to discriminate accurately between the tested biological motion stimuli. An addition of detectors for more complex features, like corners or other orientation discontinuities, is easily possible on this level of the form pathway. In neural models for stationary object recognition such detectors increase the selectivity of the model (61). The responses of the bar detectors were also passed through a linear threshold function.

Consistent with other models that reproduce electrophysiological results from area IT (*26, 61,62*), the view-tuned neurons in the form pathway are modeled by radial basis functions. The input signals of these neurons are derived from the invariant bar detectors whose responses show significant variation over time, and between the different training patterns. The criterion for significant variation was that the variance of the detector response exceeded a critical threshold. The radial basis function units are embedded in a recurrent neural network that is described in greater detail below. The feed-forward input of the view-tuned neurons is then given by:

$$G(\mathbf{u}) = \exp(-[\mathbf{u} - \mathbf{u}_0]'\ \mathbf{C}\ [\mathbf{u} - \mathbf{u}_0]) \qquad (1)$$

where $\mathbf{u}$ is the vector of the responses of the (significant) invariant bar detectors, and where $\mathbf{u}_0$ signifies the preferred input pattern of the neuron. $\mathbf{C}$ is a diagonal matrix with the elements C_{ll} that are inversely proportional to the variance of the l-th component of $\mathbf{u}$ in the training set. The centers of the radial basis functions were learned from key frames of the prototypical image sequence. Each individual biological motion pattern was represented by 21 key frames. Consistent with neurophysiological data from the area IT in monkeys (*30*), the view-tuned model neurons have large receptive fields that include the whole visual area of the model that had a diameter of approximately to 10 deg visual angle.

Model neurons	**Areas**	**Number of neurons**	**RF Size**	**References**
FORM PATHWAY				
Simple cells	V1	1010	0.6 / 1.2 deg	(*58, 59*)
Complex cells	V1, V2, V4	128	4 deg	(59, *60*)
View-tuned cells	IT, STS	1050	>8 deg	(*30*)
Motion pattern neurons	IT, STS, F5	3…40	> 8 deg	(*12, 17*)
MOTION PATHWAY				
Local motion detectors	V1,V2, (MT)	1147	0.9 deg	(*64*)
Local OF pattern detectors	MT, MST MST, KO	72 (translation) 2 x 50 (expansion / contraction)	2 deg	(*66, 67, 68, 72*) (17, 34, *69, 71*)
Complex OF pattern detectors	STS	1324	>8 deg	(*12*)
Motion pattern neurons	STS, F5	3…40	> 8 deg	(*12, 17*)

Tab 1: *Most important properties of the different model neurons in the form and the motion pathway. It is assumed that the moving figure covers a region with a height of about 6 deg visual angle.*

The highest level of the form pathway contains neurons that sum the output activities of all view-tuned units that represent the same biological motion pattern and smooth them over time. The activities of these motion pattern neurons are shown in the simulation results. Let $H_k^l(t)$ signify the activity of the view-tuned neuron that encodes keyframe (or "snapshot") k of motion pattern l. The dynamics of the response $P^l(t)$ of the motion pattern neuron for pattern l is given by (with the integration time constant $\tau_s \approx 150$ ms):

$$\tau_s \dot{P}^l(t) = -P^l(t) + \sum_l H_k^l(t) \qquad (2)$$

The first level of the motion pathway extracts the local motion information from the stimulus sequence. We did not model the extraction of the local motion energy in detail since a variety of neural models for low-level motion perception have already been proposed (*63*). Instead we calculated the optic flow fields directly from a stick figure model that was animated using two-dimensional tracking data from video sequences. This simplification is based on the assumption that the visual system is able to extract local motion information relatively accurately from articulated motion stimuli. This assumption seems consistent with perceptual experience. In addition, our simulations show that motion pattern recognition is possible even with strongly impoverished optic flow information. This implies that the performance of the model should not depend strongly on the accuracy of the estimation of the local motion energy. Neurophysiological experiments suggest that in the macaque cortical areas V1, V2 and MT are most important for the extraction of local motion (*64*). The receptive field size of the local motion energy detectors in the model is in the range of foveal direction-selective neurons in the primary visual cortex of monkeys (*65*).

The neurons on the next hierarchy level of the motion pathway extract specific local optic flow patterns. The first class of neurons responds selectively for translatory motion and has receptive field sizes that are consistent with the receptive fields of neurons in area MT in monkeys (*66*). In accordance with neurophysiological data, we assume approximately a width of about 90^o for the direction tuning (*67*), and two classes of motion detectors tuned for low and medium speeds (*68*). The model contains detectors for 8 different preferred directions. A second class of neural detectors on this hierarchy level is selective for expansion and contraction flow, or motion edges along horizontal or vertical lines. Neurons that are sensitive for such optic flow features have been found in areas MSTd (*69*). The large receptive fields of many MSTd neurons, however, make it disputable whether this area is involved in the analysis of object motion (*70*). Neurons that are selective for smaller stimuli, but have also large receptive fields have been reported in area MSTl (*71*). Neurons sensitive for local discontinuities of the optic flow have also been reported in macaque area MT (*72*). Also the kinetic occipital region (KO), which has been described recently in humans (*34*), might be important for the extraction of motion discontinuities. This area shows strong selective activation during the recognition of biological motion stimuli (*17*).

The neurons on the next-higher level of the motion pathway are selective for complex instantaneous optic flow patterns. Such "optic flow pattern neurons" are modeled by laterally coupled radial basis function units, like the view-tuned units in the form pathway. The units are trained with complex optic flow field patterns that are derived from the training video sequences. Like in the form pathway, only neurons from the previous level with significant response variation over time, or between different movement patterns contribute input signals to this hierarchy level. Such neurons that are selective for complex global optic flow patterns might be found in the different areas of the STS that have been reported to be selective for biological motion (*12*).

The highest level of the motion pathway consists of motion pattern neurons that summate and temporally smooth the activities of the optic flow pattern neurons from the previous layer. They are modeled in the same way as the motion pattern neurons in the form pathway. On this level of the visual system the ventral and the dorsal pathway may be already fused. In this case, a single set of motion pattern neurons

would integrate the signals from both pathways. Anatomical sites of such motion pattern neurons in monkeys are likely the mentioned regions in STS, but possibly also the action sensitive area F5 in the premotor cortex (*14*).

Neural Recognition Dynamics

To account for temporal order selectivity, we assumed that view-tuned and optic flow pattern-selective neurons are embedded in recurrent neural networks with asymmetric connections. Assume that $G_k^l(t)$ is the feed-forward input of a view-tuned (or optic flow pattern) neuron calculated by equation (1). If $H_k^l(t)$ is the output signal, the activity dynamics of the network is given by:

$$\tau \dot{H}_k^l(t) = -H_k^l(t) + \sum_m w(k-m) f(H_m^l(t)) + G_k^l(t)$$

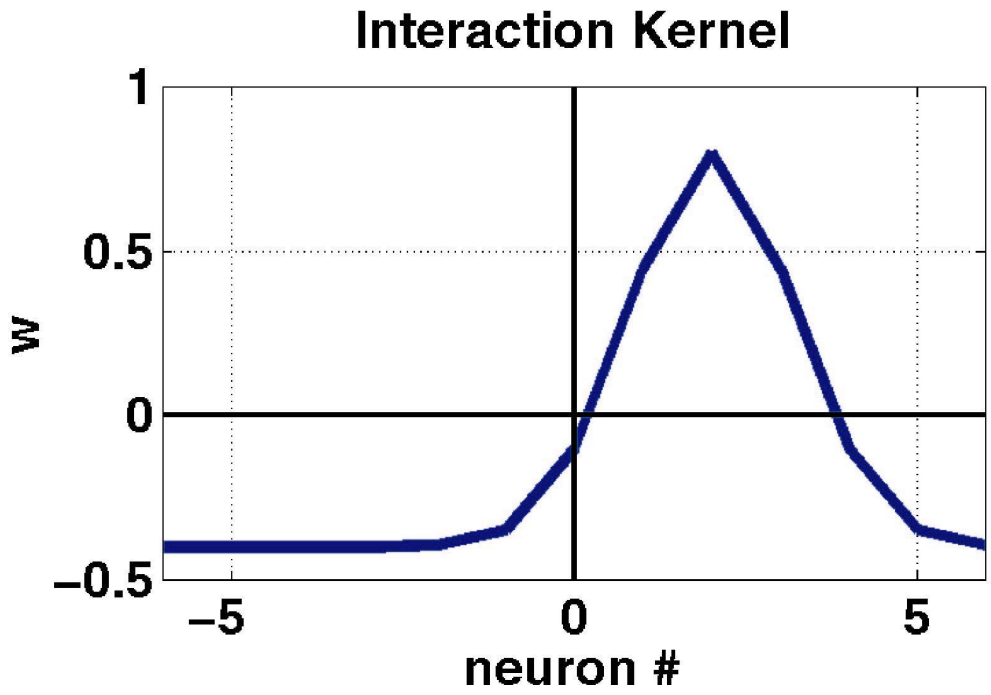

Fig. 6: *Asymmetric interaction function used in the model.*

The function $f(.)$ is a monotonic nonlinear threshold function. In the simulations a step threshold was used. Sequence selectivity is achieved because of the asymmetric interaction kernel $w(k)$ that is depicted in Fig 6. A detailed mathematical analysis of the network dynamics has been presented in (*73*). In presence of a moving form-constant input peak, given by the signals $G_k(t)$, a stable traveling pulse solution arises in the network. Outside a certain speed regime of the input peak this stable solution breaks down, giving rise to another class of solutions with much smaller amplitude. The same behavior can be shown for networks with linear threshold neurons (*73*). The time constant τ was 150 ms.

Stimulus Generation

The stimuli were created from video sequences showing actors performing different complex full-body movements. All these movements were periodic. In the movies 12 joint positions were marked by hand in about 20-30 frames per movement cycle. The tracked trajectories were smoothed by fitting them with second order Fourier series and normalized to 21 frames per movement cycle. The smoothed trajectories were used to animate a stick figure that had approximately the same outline as the moving actor. The stick figure was used to create the input images (pixel maps) for the form pathway, and also for the direct calculation of the instantaneous optic flow fields. The articulated body motion leads to highly discontinuous optic flow fields that prohibit an application of standard optic flow algorithms. Also we did not want to model the extraction of local motion energy in detail. The flow field was calculated from the stick figure using the shifts between subsequent frames of corresponding points of the articulated figure.

Invariance with Respect to Translation, Scaling, and Speed

The model was trained with stimuli with one specific stimulus size and speed. To test the efficiency of the invariance mechanism that is based on nonlinear pooling over multiple scales and positions we quantified first the translation invariance of the model. The result is shown in Fig. 7A. Significant responses in the neural representation arise for translated stimuli, as long as the shifts do not exceed about half of the width of the walking figure (about 2 deg of visual angle). Similar translation invariance has been found in psychophysical experiments in which subjects had to detect changes in a point light walker that was translated during saccades (*74*). We tested also the scaling invariance by presenting stimuli that were increased or decreased in size relative to the training patterns. The simulation results are shown in Fig. 7B. A scaling invariance of about 1.3 octaves is achieved. This size invariance is in the regime that has been reported for neurons in anterior area IT in monkeys (*75*).

We finally tested also the invariance of the stimulus with respect to speed changes. Complex movements can be recognized easily from movies with increased or decreased in speed. We tested the model with patterns that were slowed down or speeded up relative to the training sequences. Fig. 7C shows that the neural activity stays high for speed changes as high as factor two. Similar robustness against changes in speed has been observed in psychophysical experiments (*76*). The robustness against changes in speed is explained by the stability of the traveling pulse solution of the recurrent neural network that produces solutions with substantial amplitudes in a whole regime of stimulus speeds (*73*).

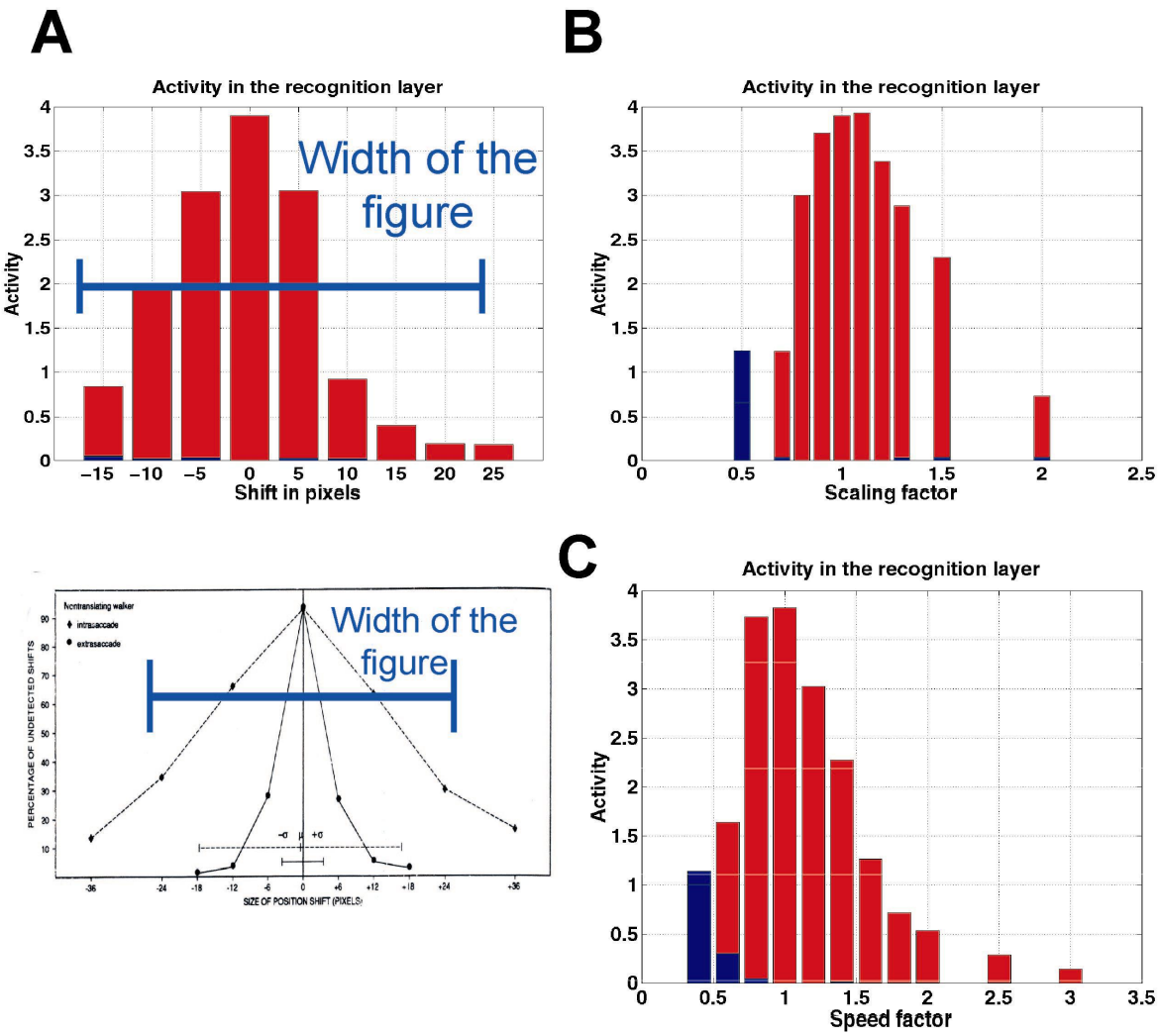

Fig. 7: *Simulation results for translation, scaling and speed invariance. The activity of the motion pattern neuron for walking in the motion pathway is shown. The dark bars indicate the maximum of the response of the other motion pattern neurons that are not selective for walking.* ***Fig. A****: Translation invariance is achieved in a regime of about half of the width of the walking figure (upper panel). This is consistent with psychophysical data obtained by translating biological motion stimuli during saccades (lower panel) [reproduced with permission from (22)].* ***Fig. B****: Activity in the motion pattern neurons for the motion pathway when the stimulus is scaled compared to the training view. Scaling invariance of about 1.3 octaves is achieved.* ***Fig. C****: Activity for changing the stimulus speed. In accordance with psychophysical experiments the speed can be changed by factor two without abolishing recognition. The form pathway shows similar invariance properties.*

Generalization over Different Walkers and Recognition of People by their Gait

Our model postulates that complex movement patterns are encoded by learning prototypical example patterns. This account is only computationally feasible if it is possible to recognize with a single prototype the biological movements executed by different people. This implies that the model after training with the locomotion pattern of one actor should generalize to patterns showing the same locomotion pattern executed by a different actor. We tested if the neural model fulfills this requirement by testing the model with the training locomotion pattern executed by nine other actors. Fig. 8A shows that reasonable generalization is achieved in both pathways. It seems possible to code for motion patterns executed by different individuals by a single learned prototype.

A number of psychophysical studies have shown that subjects are able to recognize other people by their movements, or at least can exploit movement information to improve recognition (*77*). It seems a legitimate question to test if the proposed neural mechanism is sufficiently selective to recognize individuals by their movements. We trained the model with the locomotion pattern „walking“ executed by four actors (two males and two females). From each actor we recoded five repetitions of the same locomotion pattern. Only the first repetition was used for training. Fig. 8B shows the activities of the motion pattern selective neurons. In all cases the neuron that was trained with the movement of the same actor shows the highest response. This implies that the identity of the walker is easily possible on the basis of the activities of the motion pattern selective neurons. A similar result for the motion pathway is obtained using point light stimuli. (The neurons in the form pathway are silent for such stimuli.)

Our computational analysis suggests that complex motion patterns might be encoded by relatively small populations of neurons. Dependent on the task (categorization of different gaits or identification of the identity of a walker) these populations may be read out as population code in a flexible way. Alternatively, multiple neural populations might exist that code for the characteristic gait patterns of individual people and general classes of locomotion patterns, like „walking“, „running“, etc.

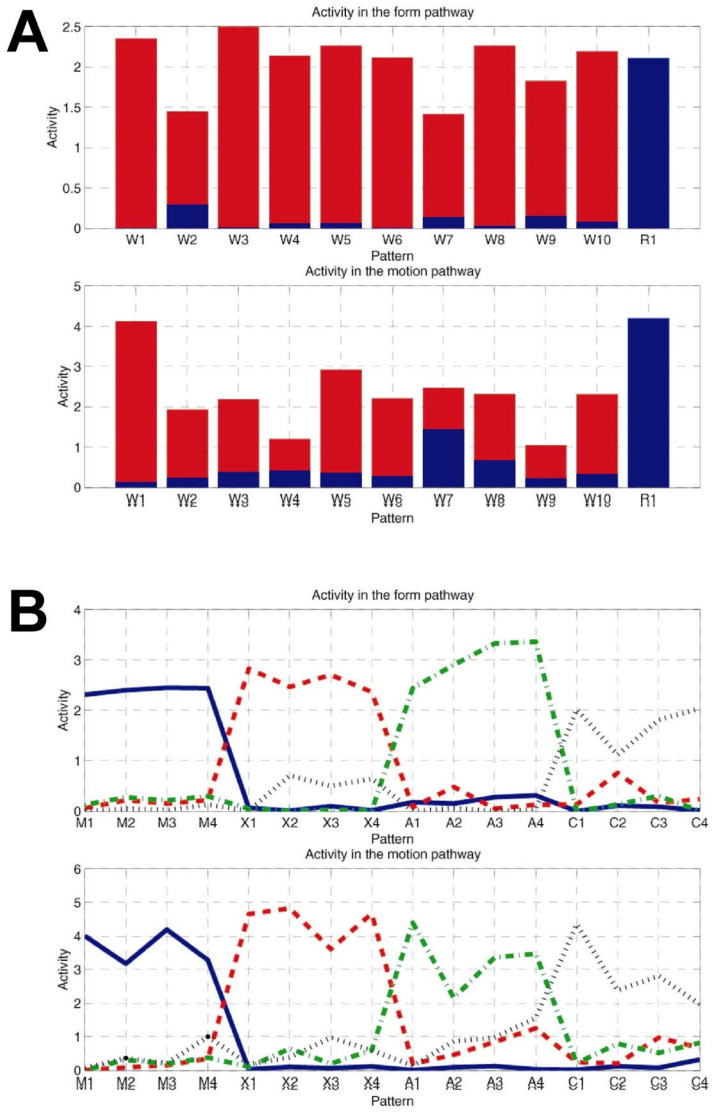

Fig. 8: *Testing the model with locomotion patterns from multiple people.* ***Fig. A****: The model was trained with walking, running, and marching executed by one actor and tested with walking executed by nine other actors. The activities of the motion pattern selective neurons are shown. Dark bars show the maximum activity of the neurons that code for motion patterns different from „walking". W1 and R1 indicate the training patterns „walking" and „running" of the actor "M" whose locomotion patterns were used during as training. W2 ... W10 correspond to „walking" executed by other actors. Substantial generalization over different actors is achieved in both pathways.* ***Fig. B****: Individual motion pattern neurons of the model are trained with walking executed by four different actors, signified by M, X, A and C. The numbers indicate the repetitions of the same gait pattern for each actor (e.g. M4 indicating the fourth recorded pattern for actor M). Only the patterns M1, ..., C1 were used for training. The activities of the motion pattern neurons are indicated by different line styles. Each neuron shows high activity only for test patterns from the same actor whose movement was used to train the neuron. From these neural activities the identities of the walkers can be easily recovered.*

References and Notes:

1. J.P. Ewert, *Behav. Brain Sci.* **10**, 337 (1989); D. Morris, *Behavior* **7**, 1 (1954).
2. C. Darwin, *The Expressions of the Emotions in Man and Animals* (Murray, London, 1872); R.J. Andrew, *Science* **142**, 1034 (1963).
3. R. Adolphs, *Curr. Op. Neurobiol*. **11**, 231 (2001).
4. G. Johansson, *Perc. Psychophys*. **14**, 201 (1973).
5. W.H. Diettrich, *Perception* **22**, 15 (1993); H. Poizner, U. Bellugi, V. Lutes-Discroll, *J. Exp. Psych. Human Perc. Perf.* **7**, 430 (1981); J.N. Bassili, *J. Exp. Psych. Human Perc. Perf.* **4**, 373 (1978).
6. J.E. Cutting, L.T. Kozlowski, *Bull. Psychon. Soc.* **9**, 353 (1977); L.T. Kozlowski, J.E. Cutting, *Perc. Psychophys.* **21**, 575 (1977); H. Hill, F.E. Pollick, *Psychol. Sci*. **11**, 223 (2000).
7. A.F. Bobick, *Phil. Trans. R. Soc. Lond. B Biol. Sci*. **352**, 1257 (1997).
8. J.T. Todd, *J. Exp. Psych. Human Perc. Perf.*. **9**, 31 (1993); M. Lappe, J.P Beintema, P. Schuchert , *Soc. Neurosc. Abst*. **26,** 1503 (2000).
9. A review of the large variety of computer vision approaches to motion recognition exceeds the scope of this paper. See for instance [C..Cedras, M. Shah, *Img. Vis. Comp*. **13**, 129 (1995); J.K. Aggarwal, Q. Cai, *Comp. Vis. Img. Underst,* **73**, 428 (1999); D.M. Gavrila, *Comp. Vis. Image Underst*. **73**, 82 (1999)].
10. I. Essa, A.P. Pentland, *IEEE Trans. Patt. Anal. Mach. Intell.* **13**, 730 (1991); S. Ju, M.J. Black, Y.Yacoob, *Proc. 2nd Int. Conf. Autom. Face and Gest. Recogn*., Killington, VT TX, 38 (IEEE, Los Alamitos, 1996).
11. N.H. Goddard, *Proc. Workshop on Vis. Motion* 212 (IEEE, Los Alamitos, 1989); N.H. Goddard, thesis, Univ. of Rochester (1992).
12. D.I. Perrett *et al., Behav. Brain Res.* **16**, 153 (1985); M.W. Oram, D.I. Perrett, *J. Cogn. Neurosc.* **6**, 99 (1994); M.W. Oram, D.I. Perrett, *J. Neurophys.* **76** 109 (1996).
13. D.I. Perrett *et al., J. Exp. Biol.* **146**, 87 (1989).
14. G. Rizzolatti *et al., Exp. Brain Res.* **71**, 491 (1988); G. Di Pelligrino *et al., Exp. Brain Res.* **91**, 176 (1992); G. Rizzolatti, M.A. Arbib*, TINS* **21**, 188 (1998).
15. J. Decety, J. Grafton, *TICS* **3** 172 (1999).
16. E. Bonda *et al., J. Neurosc*. **16**, 3737 (1996); R.J. Howard *et al*., *Curr. Biol*. **6**, 1015 (1996); E. Grossman *et al., J. Cog. Neurosc*. **12**, 711 (2000)
17. L.M. Vaina *et al., Proc. Natl. Acad. Sci USA*, **98**, 11656 (2001).
18. A. Puce *et al., J. Neurosc*. **18**, 2188 (1998); J.V. Haxby, E.A. Hoffman, M. I Gobbini, *TICS* **4**, 223 (2000); T. Allison, A. Puce, G. McCarthy, *TICS* **4**, 267 (2000).
19. S.T.Grafton *et al., Exp. Brain Res*. **112**, 103 (1996); G. Rizzolatti *et al*., *Exp. Brain Res*. **111**, 246 (1996); J. Grèzes, N. Costes, J. Decety*, Cogn. Neuropsych*. **15**, 553 (1998); Stevens *et al., Neuroreport* **11**, 109 (2000).
20. J. Decety, J. Grèzes, *Trends Cogn. Sci*. **3**, 172 (1999).
21. T. Poggio, S. Edelman, *Nature* **343**, 263 (1990); for reviews see M.J.Tarr, H.H. Bülthoff, *Cognition* **1-2**, 1 (1998); M.J. Tarr, *Nat. Neurosc*. **2**, 932 (1999); M. Riesenhuber, T. Poggio, *Nat. Neurosc*. **3** (Suppl.), 1199 (2000).
22. L.G. Ungerleider, M. Mishkin, in *The Analysis of Visual Behavior*, D.J. Ingle, R.J.W. Mansfield, M.S. Goodale, Eds. (MIT Press, Cambridge, 1982), pp. 549-586; D.J. Felleman, D.C. van Essen, *Cereb. Cortex* **1**, 1 (1991); A.D. Milner, M.A. Goodale, *Prog. Brain Res*. **95**, 317 (1993).
23. Evidence for the coupling of the ventral and the dorsal pathway exists already at the level of areas V2 and MT (*22*). A convergence of motion and form information in monkeys cortex occurs very likely at the level of the *superior parietal sulcus* [K.S. Saleem, W. Suzuki, K. Tanaka, T. Hashikawa, *J. Neurosc.* **20**, 5083 (2000)]. In fact a complete coupling might be present only in some STS neurons. For instance, a substantial fraction of neurons in the STS of monkeys have been observed to respond also to rigidly translating photographs of the human body, whereas other neurons in STS respond only in presence of articulated movements [D.I. Perrett, *et al*., in *AI and the Eye*, A. Blake, T. Troscianko, Eds. (John Wiley & Sons Ltd., New York, 1990)].

24. M.A. Giese, L.M. Vaina, *Perception (Suppl.)* **30**, 116 (2001).
25. K. Fukushima, *Biol. Cybern*. **36**, 793 (1980); D. Perrett, M. Oram, *Img. Vis. Comp*. **11**, 317-333 (1993); B. Mel, *Neural Comp*. **9**, 777 (1997); G. Wallis, E. Rolls, *Prog. Neurobiol*. **51**, 167 (1997).
26. M. Riesenhuber, T. Poggio, *Nat. Neurosc*. **2**, 1019 (1999).
27. A.J. Yu, M.A. Giese, T. Poggio, *CBCL Paper* **207** / *A.I. Memo* **2001-022**, M.I.T. (2001).
28. K. Sakai, S. Tanaka, *Vis. Res.* **40**, 855 (2000); I. Lampl, M. Riesenhuber, T. Poggio, D. Ferster, *Soc. Neurosc. Abst*. (2001).
29. J.L. Gallant, J. Braun, D.C. van Essen, *Science* **259**, 100 (1993); A. Pasupathy, C.E. Connor, *J. Neurophysiol*. **82**, 2490 (1999).
30. N.K. Logothetis, J. Pauls, T. Poggio, *Curr. Biol*. **5**, 552 (1995); N.K. Logothetis, D.L. Sheinberg, *Ann. Rev. Neurosc*. **19**, 577 (1996); K. Tanaka, *Ann. Rev. Neurosc*. **19**, 109 (1996).
31. At least for the representation of complex shapes in the infratemporal cortex experimental results suggest a sparse neural encoding [M. Young, S. Yamane, *Science* **256**, 1327 (1992); E.T. Rolls, M.J. Tovee, *Exp. Brain Res.* **73**, 713 (1995); E. Thomas, M.M. van Hulle, R. Vogels, *J. Cogn. Neurosci*. **13**, 190 (2001)] that might be read out in a flexible way, dependent of the categorization task [M. Riesenhuber, T. Poggio, *A.I. Memo No. 1679*, Artificial Intelligence Laboratory, M.I.T. (1999)].
32. R. Dubner, S.M. Zeki, *Brain Res.* **35**, 528 (1971); J.H. Maunsell, D.C. van Essen, J. Neurophysiol. **49**, 1127 (1983); J.A. Movshon, E.H. Adelson, M.S. Gizzi and W.T. Newsome, *Exp. Brain Res*. **Suppl. 11**, 117 (1986); C. Galletti, P.P. Battaglini , P. Fattori, *Exp. Brain Res*. **82**, 67 (1990); H.R. Rodman, T.D. Albright, Exp. Brain Res. **75**, 53 (1989); N.K. Logothetis, in *Visual Detection of Motion*, A.T. Smith, R.J. Snowden, Eds. (Academic Press, London, UK), pp. 177-216. Motion energy models of motion are, under certain assumptions, equivalent to correlation models [G. Sperling, J.P. van Santen, P.J. Burt P, *Spa.t Vis.* **1**, 47 (1985)].
33. Many neurons in area MSTd show selectivity for local expansion [K.Tanaka, H. Saito, *J. Neuroph*. **62**, 626 (1989); B.J. Geesaman, R.A. Andersen, *J. Neurosc*. **16**, 4716 (1996)]. Their involvement in the processing of object movement has, however, been disputed because of the typically large size of their receptive fields [G. Orban, L. Lagae, S. Raiguel, D. Xiao, H. Maes, *Perception* **24**, 269 (1995)]. Neurons with smaller receptive fields and specific center surround interactions have been reported in area MSTl [G.H. Recanzone, R.H. Wurtz, U. Schwarz, *J. Neuroph*. **78**, 2904 (1997); S. Eifuku, R.H. Wurtz, *J. Neuroph*. **80**, 282 (1998)]. Neurons sensitive to motion boundaries in area MT have been reported in several studies [e.g. J. Allman, F. Miezin, E. Mc Guinness, *Perception* **14**, 105 (1985); R.T. Born, R.B. Tootell, *Nature* **357**, 497 (1992); R.T. Born, *J. Neurophysiol.* **84**, 2658 (2000)]. In principle other detectors for the local optic flow field structrure could be added at this level, e.g. detectors for rotational flow or spiral motion.
34. The kinetic occipital area (KO) was described in [G. Orban et al., *Proc. Nat. Acad. Sci. USA* **92**, 993 (1995); P. Dupont *et al*., *Cereb. Cortex* **7**, 1047 (1997)]. A strong involvement of this area in biological motion perception is suggested by recent fMRI data (17).
35. K. Lander, V. Bruce, *Ecol. Psych*. **12**, 259 (2000); K. Verfaillie, *Brain Cogn*. **44**, 192 (2000).
36. A similar mechanism was proposed to account for direction selectivity in area V1 [P. Mineiro, D. Zipser, *Neural Comp*. **10**, 353 (1998)]; alternative mechanisms rely on synaptic delays (*11*), or exploit Reichardt or equivalent motion energy detectors that receive input signals from the snapshot and optic flow pattern neurons.
37. T.J. Sejnowski, *Neuron*, **24**, 773 (1999); R.P.N. Rao, T.J. Sejnowski, *Neural Comp.* **13**, 2221 (2001). Similar time-dependent associative learning rules have been postulated as basis for the learning of the spatial relationships between different views of the same object from the temporal coincidence of images in natural visual scenes [M.P. Stryker, *Nature*, **354**, 108 (1991); P. Foeldiak, *Neural Comp*. **3**, 194 (1991); G. Wallis, E.T. Rolls, *Progr. Neurobiol*. **51**, 167 (1997)].
38. Y. Miyashita, H.S. Chang, *Nature* **331**, 68 (1988); Y. Miyashita, *Nature* **335**, 817 (1988); D.J. Amit, S. Fusi, V. Yakovlev, *Neural Comp*. **9**, 1071 (1997); V.Yakovlev, S. Fusi, E. Berman, E. Zohary, Nat. Neurosc., **1**, 310 (1998).
39. G. Johansson, *Psychol. Res.* **38**, 379 (1976).
40. V.S. Ramachandran, C. Armel, C. Foster, R. Stoddard, *Nature* **395**, 852 (1998).

41. The technique of spatio-temporal morphable models [M.A. Giese, T. Poggio, *Int. J. Comp. Vis*, **38**, 59 (2000)] linearly combines complex motion patterns in space-time. The patterns are defined by trajectories of the joint positions. A class of new movement patterns can be generated in the following way, for instance from the prototypes walking, running and limping:

$$\text{new locomotion pattern} = \alpha_1 \text{ walking} + \alpha_2 \text{ running} + \alpha_3 \text{ limping}$$

The weights $\alpha_1, \alpha_2, \alpha_3$ determine the individual prototypes' contributions to the linear combination. By varying these coefficients a continuous class of similar motion patterns can be generated. At the same time, the coefficient vectors define a metric space over this class. The spatio-temporal dissimilarities of the morphs are given by the Euclidian distances between the associated weight vectors ($\alpha_1, \alpha_2, \alpha_3$). The weights were always non-negative and summed to one.
42. M.A. Giese, M. Lappe, *Vis. Res.* **42**,1847 (2002).
43. S. Sumi, *Perception* **13**, 283 (1984); V. Ahlström, R. Blake, U. Ahlström, *Perception* **26**, 1539 (1997); M. Pavlova, A. Sokolov, *Percept. Psychophys*. **62**, 889 (2000); I. Bülthoff, H.H. Bülthoff, P. Sinha, *Nat. Neurosc*. **1**, 254 (1998).
44. E.D. Grossman, R. Blake, *Vis. Res*. **41**, 1475 (2001).
45. R. Blake, *Psycholog. Sci*. **4**, 54 (1993).
46. P. Neri, C. Morrone, D.C. Burr, *Nature* **395**, 894 (1998).
47. G. Mather, K. Redford, S.West, *Proc. Roy. Soc Lond. Ser. B* **214**, 501 (1992); J. Pinto, M. Shiffrar, *Acta Psych.* **102**, 293 (1999).
48. J.E. Cutting, C. Moore, R. Morrison, *Percept. Psychop*. **44**, 339 (1988); B.I. Bertenthal, J. Pinto, *Psychol. Sci.* **5**, 221 (1994); I.M. Thornton, J. Pinto, M. Shiffrar, *Cogn. Neuropsych*. **15**, (1998).
49. Such assignments are a strong simplification and have to be interpreted with caution for a number of reasons: First, the homology between the different cortical areas in humans und monkeys is not yet fully established. Secondly, some functional layers of the model might correspond to multiple cortical areas. For instance, the extraction of local motion might involve the primary visual cortex and area MT/V5. Also we neglected the activity of multiple areas outside the visual system, which has been described in a number of studies (*20*). Another problem is the nontrivial relationship between the neural activity and the BOLD signal. In some experiments a linear relationship has been found [G. Rees, K. Friston, C. Koch, *Nat. Neurosc.* **3**, 716 (2000); Logothetis *et al., Nature* **412**, 150 (2001)]. Here we assume that the BOLD signal varies at least monotonically with cortical neural activity.
50. M.A Giese, T. Poggio, *Nat. Neurosc*., submitted.
51. T. Poggio, *Proceedings of Cold Spring Harbor Symposia on Quantitative Biology* **4**, 899-910 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1990).
52. Interesting in this context seems also the finding that a region in the human occipitotemporal cortex shows selective activation for pictures and movies of the human body and body parts [P. Downing, Y. Jiang, M. Shuman, N. Kanwisher, *Science* **293**, 2470, 2001].
53. S.H. Chatterjee, J.J. Freyd, M. Shiffrar, *J. Exp. Psychol. Hum. Percept. Perform*. **22**, 916 (1996).
54. L.M. Vaina, M. Lemay, D.C. Bienfang, A.Y. Choi, *Vis. Neurosci*. **5**, 353 (1990); McLeod et al., *Vis. Cogn*. **3**, 363 (1996); T. Schenk, J. Zihl, *Neuropsychologia* **35**, 1299 (1997).
55. With a fusion of the pathways on the level of the STS the model predicts correctly the characteristic deficits of such patients (*24*).
56. I.M. Thornton, R.A. Rensink, M. Shiffrar*, Perception*, (submitted).; P. Cavanagh, A.T. Labianca, I.M. Thornton, *Cognition* **80**, 47 (2001).
57. Z. Kourtzi, N. Kanwisher, *J. Cog. Neurosc*. **2**, 48 (2000); P.U. Tse, P. Cavanagh, *Cognition* **74**, B27 (2000).
58. J.P. Jones, L.A. Palmer, *J. Neurophys*. **58**, 1233 (1987).
59. D.H. Hubel, T.N. Wiesel, *J. Physiol (Lond.)* **160**, 106 (1962); B.M. Dow, A.Z. Snyder, R.G. Vautin, R. Bauer, *Exp. Brain Res*. **44**, 213 (1981).
60. E. Kobatke, K. Tanaka, *J. Neurophys*. **71**, 856 (1995).

61. M.K. Riesenhuber, T. Poggio, in *Advances of Neural Information Processing Systems,* M.I. Jordan, M.J. Kaerns, S.A. Solla (Ed.) **10**, (MIT Press, Cambridge, USA, 1998), pp. 215-221.
62. T. Poggio, S. Edelman, *Nature* **343**, 263 (1990).
63. M.E. Sereno, Neural Computation of Pattern Motion (MIT Press, Cambridge, USA, 1993); S.J. Nowlan, T.J. Sejnowski, *J. Neurosci.* **15**, 1195 (1995); E.P. Simoncelli, D.J. Heeger, Vis. Res. **38**, 743 (1998); J. Chey, S. Grossberg, E. Mingolla, *Vis. Res.* **38**, 2769 (1998).
64. J.A. Movshon, E.H. Adselson, M. Gizzi, W.T. Newsome, in *Study Group on Pattern Recognition Mechanisms*, C. Chagass, C.G. Gross Ed., (Ponti. Acad. Scient., Vatican City, Italy, 1985), pp. 117-151; H.R. Rodman, T.D. Albright, *Exp. Brain Res*. **75**, 53 (1989); N.K. Logothetis, in *Visual Detection of Motion*, A.T. Smith, R.J. Snowden (Ed.) (Academic Press, London, UK), pp. 177-216.
65. A. Mikami, W.T. Newsome, R.H. Wurtz, *J. Neurophysiol*. **55**, 1328 (1986).
66. T.D. Albright, R. Desimone, *Exp. Brain Res*. **65**, 582 (1987).
67. H.R. Rodman, T.D. Albright, *Vis. Res*. **27**, 2035 (1987).
68. G.A. Orban, *Neuronal Operations in the Visual Cortex*, Springer-Verlag (1984, Berlin, Germany); L. Lagae, S.Raiguel, G.A. Orban, *J. Neurophysiol*. **69**, 19 (1993).
69. H. Saito *et al*., *J. Neurosci*. **6**, 145 (1986); K. Tanaka, H. Saito et al., *J. Neurophysoiol*. **62**, 626 (1989);
70. G.A. Orban, L. Lagae, S. Raiguel, D. Xiao, H. Maes, *Perception* **24**, 269 (1995); B.J. Geesaman, R.A. Andersen, *J. Neurosc*., **16** 4716 (1996).
71. K. Tanaka, Y. Sugita, M. Morita, J. Saito, *J. Neurophysiol*. **69**, 128 (1993); H. Kumatsu, R.H. Wurtz, J. Neurophysiol., 60, 80 (1988); S. Eifuku, R.H. Wurtz, *J. Neurophysiol*. **80**, 282 (1998).
72. J. Allman, F. Miezin, E. McGuinness, *Perception*, 14, 105 (1985); D.K. Xiao, S. Raiguel, V. Marcar, J.J. Koenderink, G.A. Orban, *Proc. Natl. Acad. Sci. USA* **92**, 11303 (1995); R.A. Anderson, S. Treue, *Vis. Neurosci*. 13, 797 (1996); R.T. Born, *J. Neurophysiol.* **84**, 2658 (2000).
73. M.A.Giese, X. Xie, *Neurocomputing* (in press); X. Xie, M.A.Giese, in *Advances in Neural Information Processing Systems* **14**, (in press).
74. K. Verfaillie, A. De Troy, J. Van Rensbergen, *J. Exp. Psychol. Learn. Mem. Cogn*. **20**, 649 (1994).
75. E.L. Schwartz, R. Desimone, T.D. Albright, C.G. Gross, *Proc. Natl. Acad. Sci. USA* **80**, 5776 (1983); M. Ito, H. Tamura, I. Fujita, K. Tanaka, *J. Neurophysiol*. **73**, 218 (1995); N.K. Logothetis, J. Pauls, T. Poggio, *Curr. Biol*. **5**, 552 (1995).
76. M. Pavlova, S. Sokolov, *Percept. Psychophys*. **62**, 889 (2000).
77. J.E. Cutting, L.T Kozlowski, *Bull. Psychonom. Soc.* **9**, 353 (1977); B. Knight, A. Johnston, *Vis. Cogn.* **4**, 265 (1997); F. Christie, V. Bruce, *Mem. Cogn*. **26**, 780 (1998); H. Hill, A. Johnston, *Curr. Biol*. **11**, 880 (2001).