# Invariant Global Motion Recognition in the Dorsal Visual System: A Unifying Theory

**Edmund T. Rolls**
*Edmund.Rolls@psy.ox.ac.uk*
**Simon M. Stringer**
*simon.stringer@psy.ox.ac.uk*
*Oxford University, Centre for Computational Neuroscience, Department of Experimental Psychology, Oxford OX1 3UD, England*

**The motion of an object (such as a wheel rotating) is seen as consistent independent of its position and size on the retina. Neurons in higher cortical visual areas respond to these global motion stimuli invariantly, but neurons in early cortical areas with small receptive fields cannot represent this motion, not only because of the aperture problem but also because they do not have invariant representations. In a unifying hypothesis with the design of the ventral cortical visual system, we propose that the dorsal visual system uses a hierarchical feedforward network architecture (V1, V2, MT, MSTd, parietal cortex) with training of the connections with a short-term memory trace associative synaptic modification rule to capture what is invariant at each stage. Simulations show that the proposal is computationally feasible, in that invariant representations of the motion flow fields produced by objects self-organize in the later layers of the architecture. The model produces invariant representations of the motion flow fields produced by global in-plane motion of an object, in-plane rotational motion, looming versus receding of the object, and object-based rotation about a principal axis. Thus, the dorsal and ventral visual systems may share some similar computational principles.**

## 1 Introduction

A key issue in understanding the cortical mechanisms that underlie motion perception is how we perceive the motion of objects such as a rotating wheel invariantly with respect to position on the retina, and size. For example, we perceive the wheel shown in Figures 1 and 4a rotating clockwise independent of its position on the retina. This occurs even though the local motion for the wheels in the different positions may be opposite (as indicated in the dashed box in Figure 1). How could this invariance of the visual motion perception of objects arise in the visual system? Invariant motion representations are known to be developed in the cortical dorsal visual system. Motion-sensitive neurons in V1 have small receptive fields

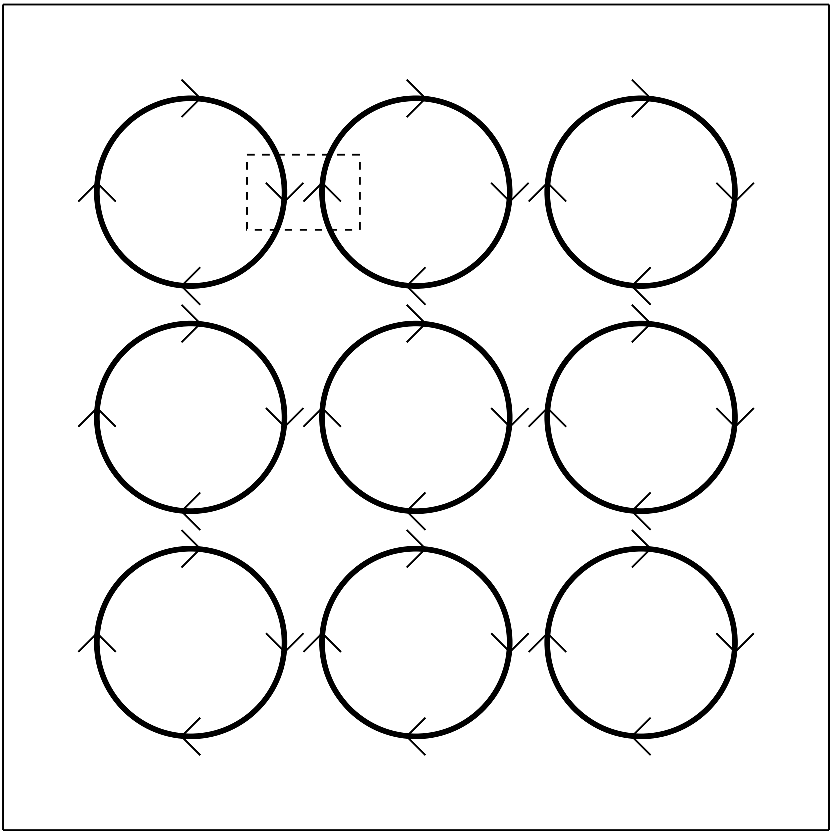


Figure 1: A wheel rotating clockwise at different locations on the retina. How can a network learn to represent the clockwise rotation independent of the location of the moving object? The dashed box shows that local motion cues available at the beginning of the visual system are ambiguous about the direction of rotation when the stimulus is seen in different locations. One rotating wheel is presented at any one time, but the need is to develop a representation of the fact that in the case shown, the rotating flow field is always clockwise, independent of the location of the flow field and even though the local motion cues may be ambiguous, as shown in the dashed box.

(in the range 1–2 degrees at the fovea), and therefore cannot detect global motion, and this is part of the aperture problem (Wurtz & Kandel, 2000). Neurons in MT, which receives inputs from V1 and V2, have larger receptive fields (e.g., 5 degrees at the fovea) and are able to respond to planar global motion, such as a field of small dots in which the majority (in practice, as little as 55%) move in one direction, or to the overall direction of a moving plaid, the orthogonal grating components of which have motion at 45 degrees to the overall motion (Movshon, Adelson, Gizzi, & Newsome, 1985; Newsome, Britten, & Movshon, 1989). Further on in the dorsal visual system, some neurons in macaque visual area MST (but not MT) respond to rotating flow fields or looming with considerable translation invariance (Graziano, Andersen, & Snowden, 1994; Geesaman & Andersen, 1996).

It is known that single neurons in the ventral visual system have translation, size, and even view-invariant representations of stationary objects (Rolls & Deco, 2002; Desimone, 1991; Tanaka, 1996; Logothetis & Sheinberg, 1996; Rolls, 1992, 2000, 2006). A theory that can account for this uses a feature hierarchy network (Fukushima, 1980; Rolls, 1992; Wallis & Rolls, 1997; Riesenhuber & Poggio, 1999) combined with an associative Hebb-like learning rule (in which the synaptic weights increase in proportion to the pre-and postsynaptic firing rates) with a short-term memory of, for example, 1 sec, to enable different instances of the stimulus to be associated together as the visual objects transform continuously from second to second in the world (Földiák, 1991; Rolls, 1992; Wallis & Rolls, 1997; Bartlett & Sejnowski, 1998; Rolls & Milward, 2000; Stringer & Rolls, 2000, 2002; Rolls & Deco, 2002).

In a unifying hypothesis, we propose here that the analysis of invariant motion in the dorsal visual system uses a similar architecture and learning rule, but in contrast utilizes as its inputs neurons that respond to local motion of the type found in the primary visual cortex, V1 (Wurtz & Kandel, 2000; Duffy, 2004; Bair & Movshon, 2004). A feature of the theory is that motion in the visual field is computed only once in V1 (by processes that take into account luminance changes across short times) and that the representations of motion that develop in the dorsal visual system require no further computation of time-delay-related firing to compute motion. The theory is of interest, for it proposes that some aspects of the computations in parts of the cerebral cortex that appear to be involved in different types of visual function, the dorsal and ventral visual systems, may in fact be performed by some similar organizational and computational principles.

## 2 The Theory and Its Implementation in a Model

**2.1 The Theory.** We propose that the general architecture of the dorsal visual system areas we consider is a feedforward feature hierarchy network, the inputs to which are local motion-sensitive neurons of V1 with receptive fields of approximately 1 degree in diameter (see Figure 2). There is convergence from stage to stage, so that a neuron at any one stage need receive only a limited number of inputs from the preceding stage, yet by the end of the network, an effectively global computation that can take into account information derived from different parts of the retina can have been performed. Within each cortical layer of the architecture (or layer of the network), local lateral inhibition implemented by inhibitory feedback neurons implements competition between the neurons, in such a way that fast-firing neurons

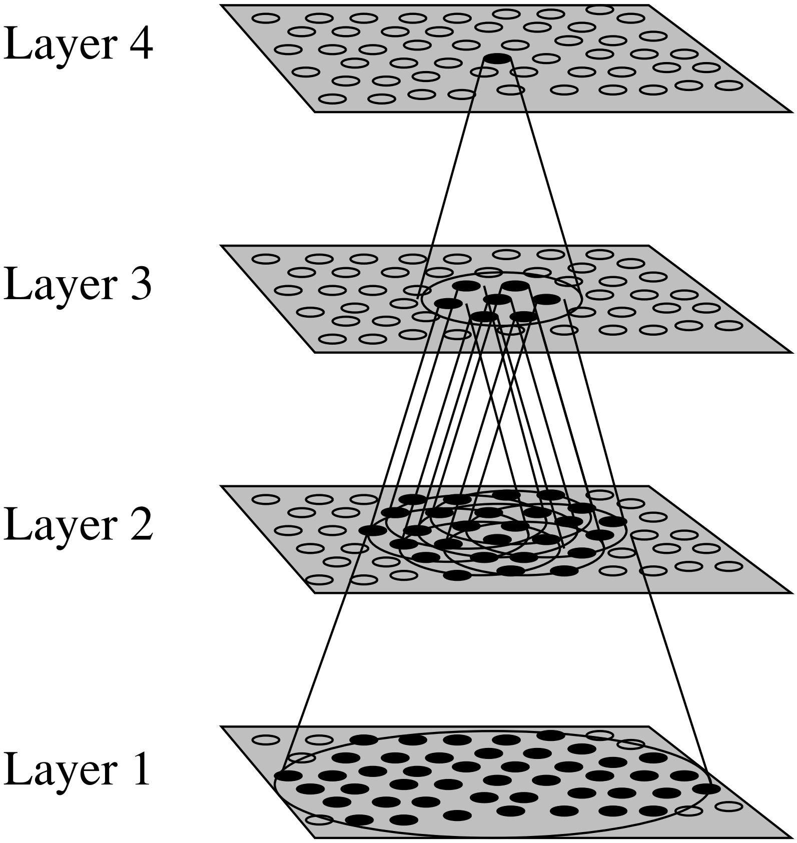


Figure 2: Stylized image of hierarchical organization in the dorsal as well as ventral visual system. The architecture is captured by the VisNet model, in which convergence through the network is designed to provide fourth-layer neurons with information from across the entire input retina.

inhibit other neurons in the vicinity, so that the overall activity within an area is kept within bounds. The competition may be nonlinear, due in part to the threshold nonlinearity of neurons, and this competition, helped by the diluted connectivity (i.e., the fact that only a low proportion of the neurons are connected), enables some neurons to respond to particular combinations of the inputs being received from the preceding area (Rolls & Deco, 2002; Deco & Rolls, 2005). These aspects of the architecture potentially enable single neurons at higher stages of the network to respond to combinations of the local motion inputs from V1 to the first layer of the network. These combinations, helped by the increasingly larger receptive fields, could include global motion to partly randomly moving dots (and to plaids) over areas as large as 5 degrees in MT (Wurtz & Kandel, 2000; Duffy & Wurtz, 1996). In the architecture shown in Figure 2, layer 1 might correspond to MT; layer 2 to MST, which has receptive fields of 15–65 degrees in diameter; and layers 3 and 4 to areas in the parietal cortex such as 7a and to areas in the cortex in the superior temporal sulcus, which receives from parietal visual areas where view-invariant object-based motion is represented (Hasselmo, Rolls, Baylis, & Nalwa, 1989; Sakata, Shibutani, Ito, & Tsurugai, 1986). The synaptic plasticity between the layers of neurons has a Hebbian associative component in order to enable the system to build reliable representations in which the same neurons are activated by particular stimuli on different occasions in what is effectively a hierarchical multilayer competitive network (Rolls, 1992; Wallis & Rolls, 1997; Rolls & Deco, 2002). Such processes might enable neurons in layer 2 of Figure 4a to respond to, for example, a wheel rotating clockwise in one position on the retina (e.g., neuron A in layer 2).

A key issue not addressed by the architecture described so far is how rotation (e.g., of a small wheel rotating clockwise) in one part of the retina activates the same neurons at the end of the network as when it is presented on a different part of the retina (see Figure 1). We propose that an associative synaptic learning rule with a short-term memory trace of neuronal activity is used between the layers to solve this problem. The idea is that if at a high level of the architecture (labeled layer 2/3 in Figure 4a) a wheel rotating clockwise is activating a neuron in one position on the retina, then the activated neurons remain active in a short delay period (of, e.g., 1 s) while the object moves to another location on the retina (e.g., the right position in Figure 4a). Then, with the postsynaptic neurons still active from the motion at the left position, the newly active synapses onto the layer 2/3 neuron (C) show associative modification, resulting in neuron C learning in an unsupervised way to respond to the wheel rotating clockwise in either the left or the right position on the retina. The idea is, just as for the ventral visual system (Rolls & Deco, 2002), that whatever the convergence allows to be learned at each stage of the hierarchy will be learned by this invariance algorithm, resulting in neurons higher in the hierarchy having higher- and higher-level invariance properties, including view-invariant object-based motion. More formally, the rule we propose is that identical to the one proposed for the ventral visual system (Földiák, 1991; Rolls, 1992; Wallis & Rolls, 1997; Rolls & Deco, 2002) as follows:

$$\Delta w_j = \alpha \overline{y}^{\tau-1} x_j^{\tau}, \tag{2.1}$$

where the trace $\overline{y}^{\tau}$ is updated according to

$$\overline{y}^{\tau} = (1-\eta) y^{\tau} + \eta \overline{y}^{\tau-1}, \tag{2.2}$$

and we have the following definitions:

$x_j$ : $j$th input to the neuron<br>
$\overline{y}^\tau$ : trace value of the output of the neuron at time step $\tau$<br>
$w_j$: synaptic weight between $j$th input and the neuron<br>
$y$ : output from the neuron<br>
$\alpha$ : learning rate; annealed between unity and zero<br>
$\eta$ : trace value; the optimal value varies with presentation sequence length

The parameter $\eta$ may be set anywhere in the interval [0, 1], and for the simulations described here, $\eta$ was set to 0.8, which works well with nine transforms for each object in the stimulus set (Wallis & Rolls, 1997). (A discussion of the good performance of this rule, and its relation to other versions of trace learning rules, including the point that the trace can be implemented in the presynaptic firing, is provided by Rolls & Milward, 2000, and Rolls & Stringer, 2001. We note that in the version of the rule used here (equation 2.1), the trace is calculated from the postsynaptic firing in the preceding time step ($y^{\tau-1}$) but not the current time step, but that analogous performance is obtained if the firing in the current time step is also included (Rolls & Milward, 2000; Rolls & Stringer, 2001).) The temporal trace in the brain could be implemented by a number of processes, as simple as continuing firing of neurons for several hundred ms after a stimulus has disappeared or moved (as shown to be present for at least inferior temporal neurons in masking experiments—Rolls & Tovee, 1994; Rolls, Tovee, Purcell, Stewart, & Azzopardi, 1994), or by the long time constant of NMDA receptors and the resulting entry of calcium to neurons. An important idea here is that the temporal properties of the biologically implemented learning mechanism are such that it is well suited to detecting the relevant continuities in the world of real motion of objects. The system uses the underlying continuity in the world to help itself learn the invariances of, for example, the motions that are typical of objects.

**2.2 The Network Architecture.** The model we used for the simulations was VisNet, which was developed as a model of hierarchical processing in the ventral visual system that uses a trace learning to develop invariant representations of stationary objects (Wallis & Rolls, 1997; Rolls & Milward, 2000; Rolls & Deco, 2002). The simulations performed here utilized the latest version of the VisNet model (VisNet2), with the same model parameters as used by Rolls and Milward (2000) for their investigations of the formation of invariant representations in the ventral visual system. These parameters were kept identical for all the simulations described here. The difference is that instead of using simple cell-like inputs to the model that respond to stationary-oriented bars and edges (with four spatial frequencies and four orientations), in the modeling described here we used motion-related

Table 1: VisNet Dimensions.

| | Dimensions | Number of Connections | Radius |
|---|---|---|---|
| Layer 4 | 32 × 32 | 100 | 12 |
| Layer 3 | 32 × 32 | 100 | 9 |
| Layer 2 | 32 × 32 | 100 | 6 |
| Layer 1 | 32 × 32 | 201 | 6 |
| Input layer | 128 × 128 × 8 | - | - |

Table 2: Lateral Inhibition Parameters.

| Layer | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Radius, $\sigma$ | 1.38 | 2.7 | 4.0 | 6.0 |
| Contrast, $\delta$ | 1.5 | 1.5 | 1.6 | 1.4 |

inputs that capture some of the relevant properties of neurons present in V1 as part of the primate magnocellular (M) system (Wurtz & Kandel, 2000; Duffy, 2004; Rolls & Deco, 2002).

VisNet is a four-layer feedforward network with unsupervised competitive learning at each layer. For each layer, the forward connections to individual cells are derived from a topologically corresponding region of the preceding layer, with connection probabilities based on a gaussian distribution (see Figure 2). These distributions are defined by a radius that will contain approximately 67% of the connections from the preceding layer. Typical values are given in Table 1.

Within each layer there is competition between neurons, which is graded rather than winner-take-all, and is implemented in two stages. First, to implement lateral inhibition, the firing rates of the neurons (calculated as the dot product of the vector of presynaptic firing rates and the synaptic weight vector on a neuron, followed by a linear activation function to produce a firing rate) within a layer are convolved with a spatial filter, $I$, where $\delta$ controls the contrast and $\sigma$ controls the width, and $a$ and $b$ index the distance away from the center of the filter:

$$I_{a,b} = \begin{cases} -\delta e^{-\frac{a^2+b^2}{\sigma^2}} & \text{if } a \neq 0 \quad \text{or} \quad b \neq 0, \\ 1 - \sum_{\substack{a\neq 0 \\ b\neq 0}} I_{a,b} & \text{if } a = 0 \quad \text{and} \quad b = 0. \end{cases} \tag{2.3}$$

Typical lateral inhibition parameters are given in Table 2.

Next, contrast enhancement is applied by means of a sigmoid function

$$y = f^{sigmoid}(r) = \frac{1}{1 + e^{-2\beta(r-\alpha)}}, \tag{2.4}$$

Table 3: Sigmoid Parameters.

| Layer | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Percentile | 99.2 | 98 | 88 | 91 |
| Slope $\beta$ | 190 | 40 | 75 | 26 |

where $r$ is the firing rate after lateral inhibition, $y$ is the firing rate after contrast enhancement, and $\alpha$ and $\beta$ are the sigmoid threshold and slope, respectively. The parameters $\alpha$ and $\beta$ are constant within each layer, although $\alpha$ is adjusted to control the sparseness of the firing rates. For example, to set the sparseness to, say, 5%, the threshold is set to the value of the 95th percentile point of the firing rates $r$ within the layer. Typical parameters for the sigmoid function are shown in Table 3.

The trace learning rule (Földiák, 1991; Rolls, 1992; Wallis & Rolls, 1997; Rolls & Milward, 2000; Rolls & Stringer, 2001; Rolls & Deco, 2002) is that shown in equation 2.1 and encourages neurons to develop invariant responses to input patterns that tend to occur close together in time, because these are likely to be from the same moving object.

**2.3 The Motion Inputs to the Network.** The images presented to the network represent local motion signals with small receptive fields. These local visual motion (or local optic flow) input signals are similar to those of neurons in V1 in that they have small receptive fields and cannot detect global motion because of the aperture problem (Wurtz & Kandel, 2000). At each pixel coordinate in the $128 \times 128$ image, a direction of local motion/optic flow is defined. The global optic flow patterns used in the different experiments occupied part of this $128 \times 128$ image, as described for each experiment below. At each coordinate, there are eight cells, where the optimal response is defined by flows 45 degrees apart. That is, the cells are tuned to local optic flow directions of 0, 45, 90, . . . , 315 degrees. The firing rate of each cell is set equal to a gaussian function of the difference between the cell's preferred direction and the actual direction of local optic flow. The standard deviation of this gaussian was 20 degrees. The number of inputs from the arrays of motion sensitive cells to each cell in the first layer of the network is 201, selected probabilistically as a gaussian function of distance as described above and in more detail elsewhere (Rolls & Milward, 2000). The local motion signals are given to the network, and not computed in the simulations, because the aim of the simulations is to test the theory that (given that local motion inputs that are known to be present in early cortical processing; Wurtz & Kandel, 2000) the trace learning mechanism described can in a hierarchical network account for a range of the types of global motion neuron that are found in the dorsal stream visual cortical areas.

**2.4 Training and Test Procedure.** To train the network, each stimulus is presented to VisNet in a randomized sequence of locations or orientations with respect to VisNet's input retina. The different locations were spaced 32 pixels apart on the 128 × 128 retina. At each stimulus presentation, the activation of individual neurons is calculated, then the neuronal firing rates are calculated, and then the synaptic weights are updated. Each time a stimulus has been presented in all the training locations or orientations, a new stimulus is chosen at random and the process repeated. The presentation of all the stimuli through all locations or orientations constitutes one epoch of training. In this manner, the network is trained one layer at a time starting with layer 1 and finishing with layer 4. In the investigations described here, the numbers of training epochs for layers 1 to 4 were 50, 100, 100, and 75, respectively, as these have been shown in previous work to provide good performance (Wallis & Rolls, 1997; Rolls & Milward, 2000). The learning rates $\alpha$ in equation 2.1 for layers 1 to 4 were 0.09, 0.067, 0.05, and 0.04.

Two measures of performance were used to assess the ability of the output layer of the network to develop neurons that are able to respond with view invariance to individual stimuli or objects (see Rolls & Milward, 2000). A single cell information measure was applied to individual cells in layer 4 and measures how much information is available from the response of a single cell about which stimulus was shown independent of view. The measure was the stimulus-specific information or surprise, $I(s, R)$, which is the amount of information the set of responses, $R$, has about a specific stimulus, $s$. (The mutual information between the whole set of stimuli $S$ and of responses $R$ is the average across stimuli of this stimulus-specific information.) (Note that $r$ is an individual response from the set of responses $R$.)

$$I(s, R) = \sum_{r \in R} P(r|s) \log_2 \frac{P(r|s)}{P(r)} \tag{2.5}$$

The calculation procedure was identical to that described by Rolls, Treves, Tovee, and Panzeri (1997) with the following exceptions. First, no correction was made for the limited number of trials because, in VisNet, each measurement of a response is exact, with no variation due to sampling on different trials. Second, the binning procedure was to use equispaced rather than equipopulated bins. This small modification was useful because the data provided by VisNet can produce perfectly discriminating responses with little trial-to-trial variability. Because the cells in VisNet can have bimodally distributed responses, equipopulated bins could fail to separate the two modes perfectly. (This is because one of the equipopulated bins might contain responses from both of the modes.) The number of bins used was equal to or less than the number of trials per stimulus, that is, for VisNet the number of positions on the retina (Rolls et al., 1997). Because

VisNet operates as a form of competitive net to perform categorization of the inputs received, good performance of a neuron will be characterized by large responses to one or a few stimuli regardless of their position on the retina (or other transform) and small responses to the other stimuli. We are thus interested in the maximum amount of information that a neuron provides about any of the stimuli rather than the average amount of information it conveys about the whole set $S$ of stimuli (known as the mutual information). Thus, for each cell, the performance measure was the maximum amount of information a cell conveyed about any one stimulus (with a check, in practice always satisfied, that the cell had a large response to that stimulus, as a large response is what a correctly operating competitive net should produce to an identified category). In many of the graphs in this article, the amount of information each of the 50 most informative cells had about any stimulus is shown.

A multiple cell information measure, the average amount of information that is obtained about which stimulus was shown from a single presentation of a stimulus from the responses of all the cells, enabled measurement of whether across a population of cells, information about every object in the set was provided. Procedures for calculating the multiple cell information measure are given by Rolls, Treves, and Tovee (1997) and Rolls and Milward (2000). The multiple cell information measure is the mutual information $I(S, \mathbf{R})$, that is, the average amount of information that is obtained from a single presentation of a stimulus about the set of stimuli $S$ from the responses of all the cells. For multiple cell analysis, the set of responses, $\mathbf{R}$, consists of response vectors comprising the responses from each cell. Ideally, we would like to calculate

$$I(S, \mathbf{R}) = \sum_{s \in S} P(s) I(s, \mathbf{R}). \tag{2.6}$$

However, the information cannot be measured directly from the probability table $P(\mathbf{r}, s)$ embodying the relationship between a stimulus $s$ and the response rate vector $\mathbf{r}$ provided by the firing of the set of neurons to a presentation of that stimulus. (Note that "stimulus" refers to an individual object that can occur with different transforms, e.g., translation or size; see Wallis & Rolls, 1997.) This is because the dimensionality of the response vectors is too large to be adequately sampled by trials. Therefore, a decoding procedure is used, in which the stimulus $s'$ that gave rise to the particular firing-rate response vector on each trial is estimated. This involves, for example, maximum likelihood estimation or dot product decoding. For example, given a response vector $\mathbf{r}$ to a single presentation of a stimulus, its similarity to the average response vector of each neuron to each stimulus is used to estimate using a dot product comparison which stimulus was shown. The probabilities of it being each of the stimuli can be estimated in this way. Details are provided by Rolls et al. (1997). A probability table is then constructed of the real stimuli $s$ and the decoded stimuli $s'$. From this probability table, the mutual information is calculated as

$$I(S, S') = \sum_{s,s'} P(s, s') \log_2 \frac{P(s, s')}{P(s)P(s')}. \tag{2.7}$$

The multiple cell information was calculated using the five cells for each stimulus with high information values for that stimulus. Thus, in this letter, 10 cells were used in the multiple cell information analysis.

## 3 Simulation Results

We now describe simulations with the neural network model described in section 2 that enabled us to test this theory.

**3.1 Experiment 1: Global Planar Motion.** Motion-sensitive neurons in V1 have small receptive fields (in the range 1–2 deg at the fovea) and therefore cannot detect global motion, and this is part of the aperture problem (Wurtz & Kandel, 2000). As described in section 1, neurons in MT have larger receptive fields and are able to respond to planar global motion (Movshon et al., 1985; Newsome et al., 1989). Here we show that the hierarchical feature network we propose can solve this global planar motion problem and, moreover, that the performance is improved by using a trace rather than a purely associative synaptic modification rule. Invariance is addressed in later simulations.

The network was trained on two 100 × 100 stimuli representing noisy left and right global planar motion (see Figure 3a). During the training, cells developed that responded to either left or right global motion but not to both (see Figure 3), with 1 bit of information representing perfect discrimination of left from right. The untrained network with initial random synaptic weights tested as a control showed much poorer performance, as shown in Figure 3.

It might be expected that some global planar motion sensitivity would be developed by a purely Hebbian learning rule, and indeed this has been demonstrated (under somewhat different training conditions) by Sereno (1989) and Sereno and Sereno (1991). This occurs because on any single trial with one average direction of global motion, neurons at intermediate layers will tend to receive on average inputs that reflect the current average global planar motion and will thus learn to respond optimally to the current inputs that represent that motion direction. We showed that the trace learning rule used here performed better than a Hebb rule (which produced only neurons with 0.0 bits given that the motion stimulus patches presented in our simulations were in nonoverlapping locations, as

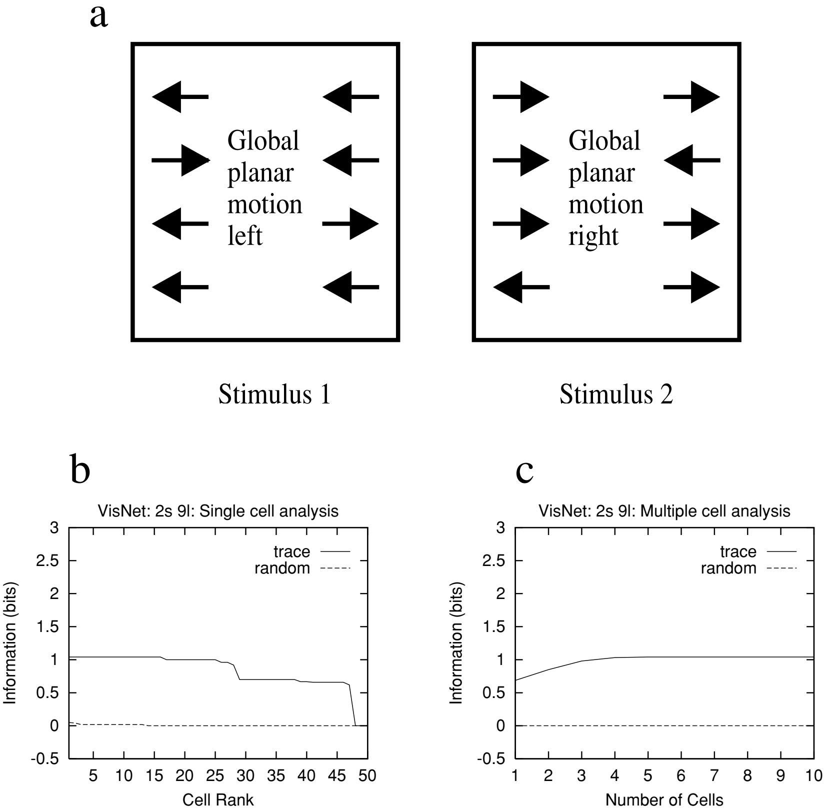


Figure 3: Experiment 1. (a) The two motion stimuli used in experiment 1 were noisy global planar motion left (left) and noisy global planar motion right (right), present throughout the 128 × 128 retina. Each arrow in this and subsequent figures represents the local direction of optic flow. The size of the optic flow pattern was 100 × 100 pixels, not the 2 × 4 shown in the diagram. The noise was introduced into each image stimulus by inverting the direction of optic flow at a random set of 45% of the image nodes. This meant that it would not be possible to determine the directional bias of the flow field by examining the optic flow over local regions of the retina. Instead, the overall directional bias could be determined only by analyzing the whole image. (b) When trained with the trace rule, equation 2.1, some single cells in layer 4 conveyed 1 bit of information about whether the global motion was left or right, and this is perfect performance. (The single cell information is shown for the 50 most selective cells.) (c) The multiple cell information measures, used to show that different neurons are tuned to different stimuli (see section 2.4), indicate that over a set of neurons, information about the whole stimulus set was present. (The information values for one cell are the average of 10 cells selected from the 50 most selective cells, and hence the value is not exactly 1 bit.)

illustrated in Figure 1). A further reason for the better performance of the trace rule is that on successive trials, the average global motion identifiable by a single intermediate-layer neuron from the probabilistic inputs will be a better estimate (a temporal average) of the true global motion, and this will be utilized in the learning.

These results show that the network architecture is able to develop global motion representations of the noisy local motion patterns. Indeed, it is emphasized that neurons in the input to VisNet had only local but not global motion information, as shown by the fact that the average amount of information the 50 most selective input cells had about the global motion was 0.0 bits.

**3.2 Experiment 2: Rotating Wheel.** Neurons in MST, but not MT, are responsive to rotation with considerable translation invariance (Graziano et al., 1994). The aim of this simulation was to determine whether layer 4 cells in our network develop position-invariant representations of wheels rotating clockwise (as shown in Figure 4a) versus anticlockwise. The stimuli consist only of optic flow fields around the rim of a geometric circle with radius 16 unless otherwise stated. The local motion inputs from the wheel in the two positions shown are ambiguous where the wheels are close to each other in Figure 4a. The network was expected to solve the problem as illustrated in Figure 4a.

The results in Figures 4b to 4d show perfect performance on position invariance when trained with the trace rule but not when untrained. The perfect performance is shown by the neurons that responded to, for example, clockwise but not anticlockwise rotation, and did this for each of the nine training positions.

Figure 4e shows perfect size invariance for some layer 4 cells when the network was trained with the trace rule with three different radii of the wheels: 10, 16, and 22.

These results show that the network architecture is able to develop location- and size-invariant representations of the global, rotating wheel, motion patterns even though the neurons in the input layer receive information from only a small local region of the retina.

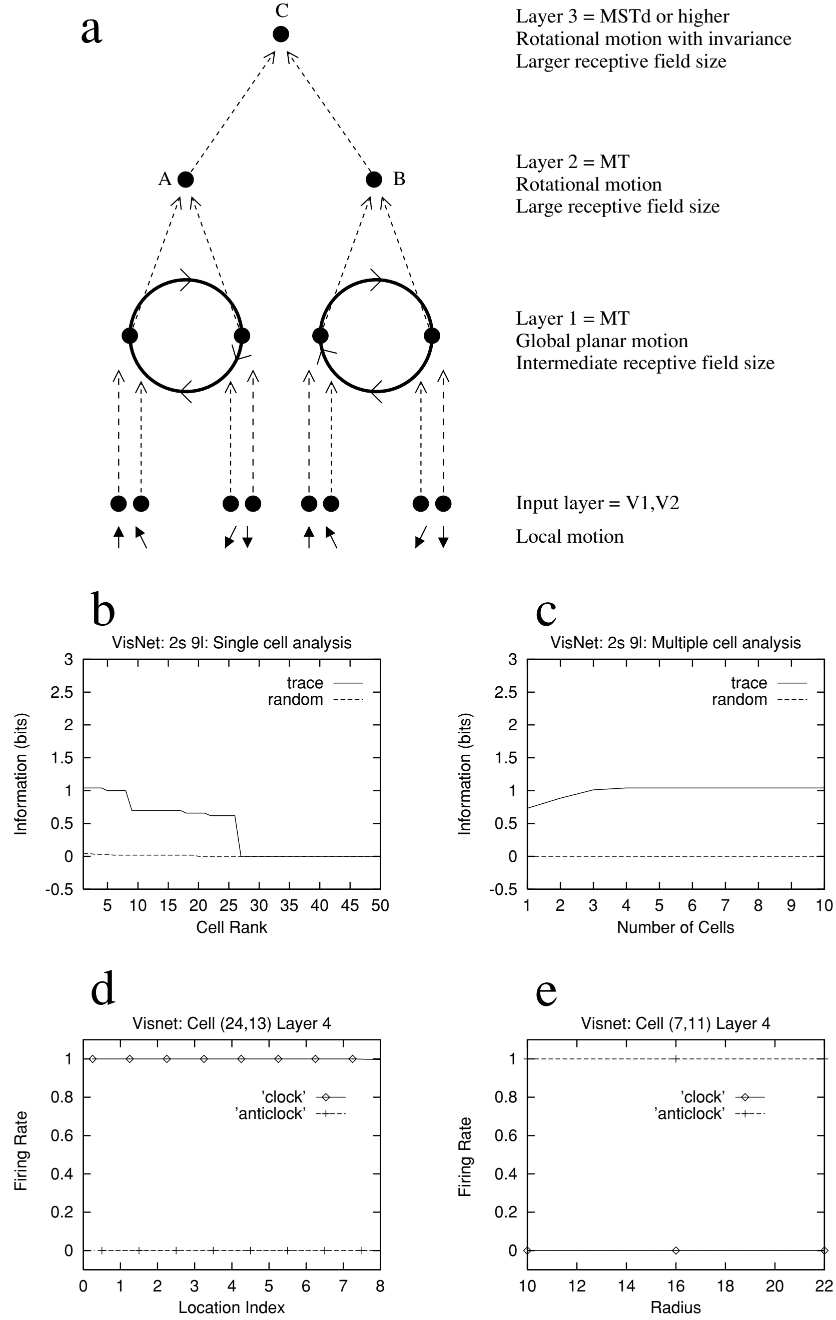
We note that the position-invariant global motion results shown in Figure 4 were not due to chance mappings of the two stimuli through the network and were a result of the training, in that the position-invariant information about whether the global motion was clockwise or anticlockwise was 0.0 bits for both the single and the multiple cell information in the untrained ("random") network. Corresponding differences between the trained and the untrained networks were found in all the other experiments described in this article.

**3.3 Experiment 3: Looming.** Neurons in macaque dorsal stream visual area MSTd respond to looming stimuli with considerable translation invariance (Graziano et al., 1994; Geesaman & Andersen, 1996). We tested whether the network could learn to respond to small patches of looming versus contracting motion typically generated by objects as they are seen successively on different locations on the retina. The network was trained on two circular flow patterns representing looming toward and looming away, as shown in Figure 5a. The stimuli are circular optic flow fields, with the direction of flow either away from (left) or toward (right) the center of the circle and with radius 16 unless otherwise stated.

The results shown in Figures 5b to 5d show perfect performance on position invariance when trained with the trace rule but not when untrained. The perfect performance is shown by the neurons that responded to, for example, looming toward but not movement away, and did this for each of the nine training positions.

Simulations were run for various optic flow field diameters to test the robustness of the results, and in all cases tested (which included radii of

Figure 4: Experiment 2. (a) Two rotating wheels at different locations rotating in opposite directions. The local flow field is ambiguous. Clockwise or counterclockwise rotation can be diagnosed only by a global flow computation, and it is shown how the network is expected to solve the problem to produce position-invariant global-motion-sensitive neurons. One rotating wheel is presented at any one time, but the need is to develop a representation of the fact that in the case shown, the rotating flow field is always clockwise, independent of the location of the flow field. (b) Single cell information measures showing that some layer 4 neurons have perfect performance of 1 bit (clockwise versus anticlockwise) after training with the trace rule, but not with random initial synaptic weights in the untrained control condition. (c) The multiple cell information measures show that small groups of neurons have perfect performance. (d) Position invariance illustrated for a single cell from layer 4, which responded only to the clockwise rotation, and for every one of the nine positions. (e) Size invariance illustrated for a single cell from layer 4, which after training with three different radii of rotating wheel, responded only to anticlockwise rotation, independent of the size of the rotating wheels. (For the position-invariant simulations, the wheel rims overlapped, but are shown slightly separated in Figure 1 for clarity.) The training grid spacing was 32 pixels, and the radii of the wheels were 16 pixels. This ensured the rims of the wheels in adjacent training grid locations overlapped. One wheel was shown on any one trial. On successive trials, the wheel rotating clockwise was shown in each of the nine locations, allowing the trace learning rule to build location-invariant representations of the wheel rotating in one direction. In the next set of training trials, the wheel was shown rotating in the opposite direction in each of the nine locations. For the size-invariant simulations, the network was trained and tested with the set of clockwise versus anticlockwise rotating wheels presented in three different sizes.


a
C
A
B
Layer 3 = MSTd or higher
Rotational motion with invariance
Larger receptive field size
Layer 2 = MT
Rotational motion
Large receptive field size
Layer 1 = MT
Global planar motion
Intermediate receptive field size
Input layer = V1,V2
Local motion
b
VisNet: 2s 9l: Single cell analysis
Information (bits)
3
2.5
2
1.5
1
0.5
0
-0.5
5 10 15 20 25 30 35 40 45 50
Cell Rank
trace
random
c
VisNet: 2s 9l: Multiple cell analysis
Information (bits)
3
2.5
2
1.5
1
0.5
0
-0.5
1 2 3 4 5 6 7 8 9 10
Number of Cells
trace
random
d
Visnet: Cell (24,13) Layer 4
Firing Rate
1
0.8
0.6
0.4
0.2
0
0 1 2 3 4 5 6 7 8
Location Index
'clock'
'anticlock'
e
Visnet: Cell (7,11) Layer 4
Firing Rate
1
0.8
0.6
0.4
0.2
0
10 12 14 16 18 20 22
Radius
'clock'
'anticlock'

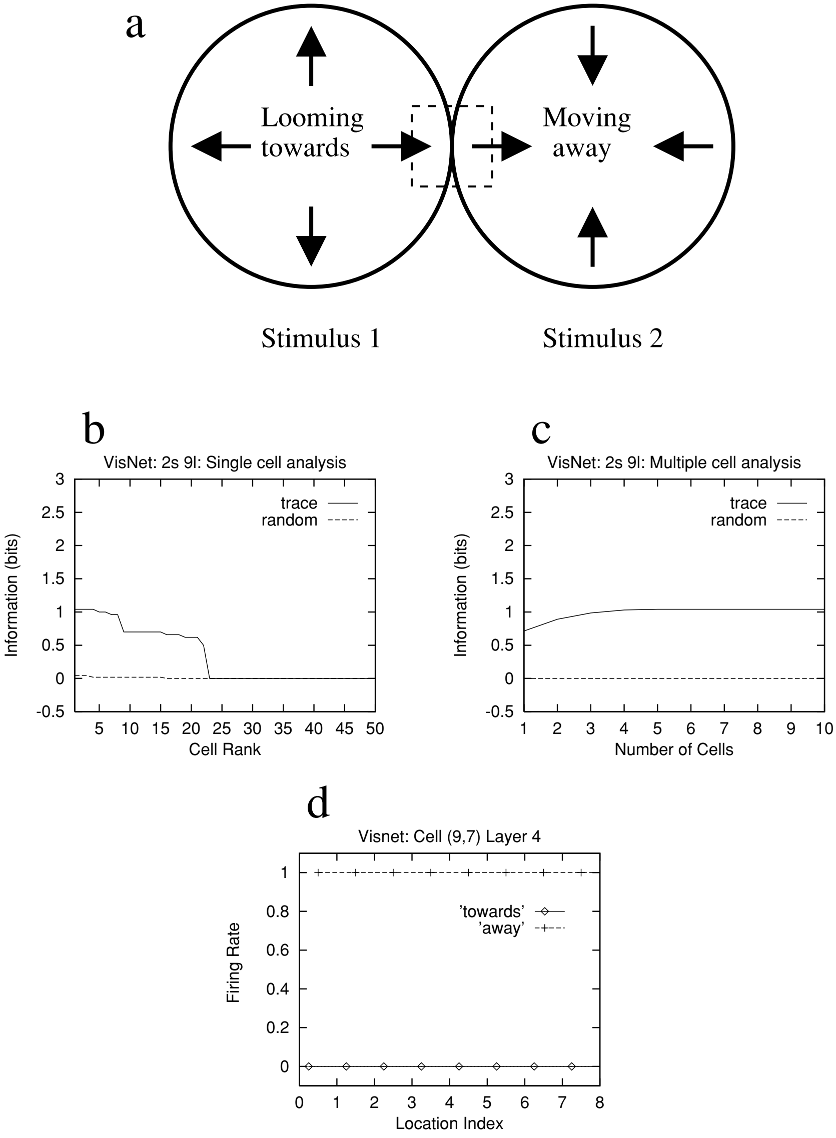
10 and 20 as well as intermediate values), cells developed a transform (location) invariant representation in the output layer.

These results show that the network architecture is able to develop invariant representations of the global looming motion patterns, even though the neurons in the input layer receive information from only a small local region of the retina.

**3.4 Experiment 4: Rotating Cylinder.** Some neurons in the macaque cortex in the anterior part of the superior temporal sulcus (which receives inputs from both the dorsal and ventral visual streams; Ungerleider & Mishkin, 1982; Seltzer & Pandya, 1978; Rolls & Deco, 2002) respond to a head when it is rotating clockwise about its own axis but not counterclockwise, regardless of whether it is upright or inverted (Hasselmo et al., 1989). The result of the inversion experiment shows that these neurons are not just responding to global flow across the visual field, but are taking into account information about the shape and features of the object. Some neurons in the parietal cortex may also respond to motion of an object about one of its axes in an object-based way (Sakata et al., 1986). In experiment 4, we tested whether the network could self-organize to form neurons that represent global motion in an object-based coordinate frame.

The network was trained on two stimuli, with four transforms of each. Figure 6a shows stimulus 1, which is a cylinder with shading at the top rotating clockwise about its own (top-defined) axis. Stimulus 1 is shown in its upright and inverted transforms. Stimulus 2 is the same cylinder with shading at the top, but rotating anticlockwise about its own vertical axis. The stimuli were presented in a single location, but to solve the problem,

Figure 5: Experiment 3. (a) The two motion stimuli were flow fields looming toward (left) and looming away (right). The stimuli are circular optic flow fields, with the direction of flow either away from (left) or toward (right) the center of the circle. Local motion cells near, for example, the intersection of the two stimuli cannot distinguish between the two global motion patterns. Location-invariant representations (for nine different locations) of stimuli looming toward or moving away from the observer were learned, as shown by the single cell information measures (b), and multiple cell information measures (c) (using the same conventions as in Figure 3) were formed if the network was trained with the trace rule but not if it was untrained. (d) Position invariance illustrated for a single cell from layer 4, which responded only to moving away, and for every one of the nine positions. (The network was trained and tested with the stimuli presented in a 3 × 3 grid of nine retinal locations, as in experiment 1. The training grid spacing was 32 pixels, and the radii of the circular looming stimuli were 16 pixels. This ensured that the edges of the looming stimuli in adjacent training grid locations overlapped, as shown in the dashed box of Figure 5a.)


a
Looming towards
Moving away
Stimulus 1
Stimulus 2
b
VisNet: 2s 9l: Single cell analysis
Information (bits)
Cell Rank
trace
random
3
2.5
2
1.5
1
0.5
0
-0.5
5 10 15 20 25 30 35 40 45 50
c
VisNet: 2s 9l: Multiple cell analysis
Information (bits)
Number of Cells
trace
random
3
2.5
2
1.5
1
0.5
0
-0.5
1 2 3 4 5 6 7 8 9 10
d
Visnet: Cell (9,7) Layer 4
Firing Rate
Location Index
'towards'
'away'
1
0.8
0.6
0.4
0.2
0
0 1 2 3 4 5 6 7 8

the network must form some neurons that respond to the clockwise rotation of the shaded cylinder independent of the four transforms of each, which were upright (0 degrees), 90 degrees, inverted (180 degrees) and 270 degrees. Other neurons should self-organize to respond to view invariant counterclockwise rotation.

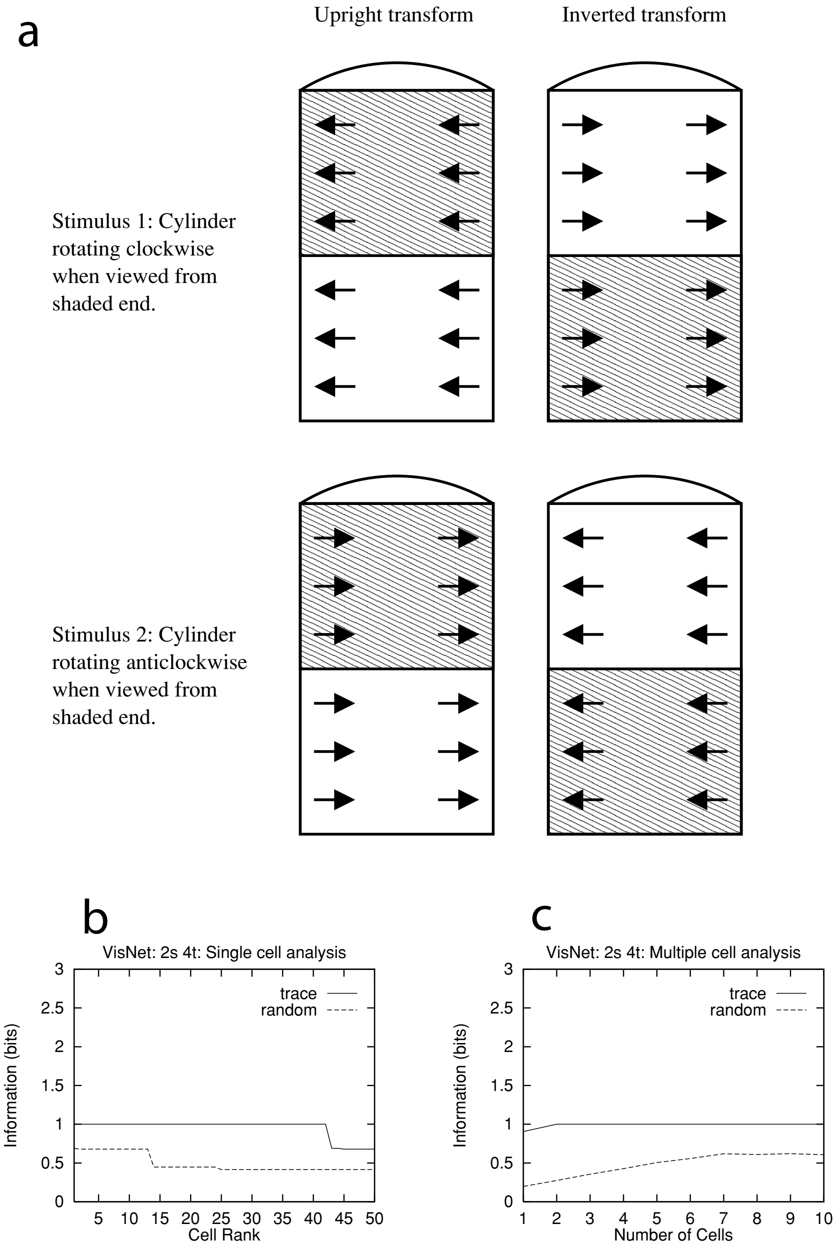
For this experiment, additional information about surface luminance must be fed into the first layer of the network in order for the network to be able to distinguish between the clockwise and anticlockwise rotating cylinders. Additional retinal inputs to the first layer of the network came from a $128 \times 128$ array of luminance-sensitive cells. The cells within the luminance array are maximally activated for the shaded region of the cylinder image. Elsewhere the luminance inputs are zero. The number of inputs from the array of luminance sensitive cells to each cell in the first layer of the network was 50.

The results shown in Figures 6b to 6c show perfect performance for many single cells, and across multiple cells, in representing the direction of rotation of the shaded cylinder about its own axis regardless of which of the four transforms was shown, when trained with the trace rule but not when untrained.

Simulations were run for various sizes of the cylinders, including height $= 40$ and diameter $= 20$. For all simulations, cells developed a transform (e.g., upright, inverted) invariant representation in the output layer. That is, some cells responded to one of the stimuli in all of its four transformations (i.e., orientations) but not to the other stimulus.

These results show that the network architecture is able to develop object-centered view-invariant representations of the global motion patterns representing the two rotating cylinders, even though the neurons in the input layer receive information from only a small, local region of the retina.

Figure 6: Experiment 4. (a) Stimulus 1, which is a cylinder with shading at the top rotating clockwise about its own (top-defined) axis. Stimulus 1 is shown in its upright and inverted transforms. Stimulus 2 is the same cylinder with shading at the top, but rotating anticlockwise about its own axis. Invariant representations were formed, with some cells coding for the object rotating clockwise about its own axis and other cells coding for the object rotating anticlockwise, invariantly with respect to whether which of the four transforms (0 degrees = upright, 90 degrees, 180 degrees = inverted, and 270 degrees) was viewed, as shown by the single cell information measures (b) and multiple cell information measures (c) (using the same conventions as in Figure 3). Because only eight images in one location form the training set, some single cells by chance with the random untrained connectivity had some information about which stimulus was shown, but cells performed the correct mapping only if the network was trained with the trace rule.


a
Upright transform
Inverted transform
Stimulus 1: Cylinder rotating clockwise when viewed from shaded end.
Stimulus 2: Cylinder rotating anticlockwise when viewed from shaded end.
b
VisNet: 2s 4t: Single cell analysis
Information (bits)
Cell Rank
trace
random
0
0.5
1
1.5
2
2.5
3
5
10
15
20
25
30
35
40
45
50
c
VisNet: 2s 4t: Multiple cell analysis
Information (bits)
Number of Cells
trace
random
0
0.5
1
1.5
2
2.5
3
1
2
3
4
5
6
7
8
9
10

**3.5 Experiment 5: Optic Flow Analysis of Real Images: Translation Invariance.** In experiments 5 and 6, we extend this research by testing the operation of the model when the optic flow inputs to the network are extracted by a motion analysis algorithm operating on the successive images generated by moving objects.

The optic flow fields generated by a moving object were calculated as described next and were used to set the firing of the motion-selective cells, the properties of which are described in section 2.3. These optic flow algorithms use an image gradient-based method, which exploits the relationship between the spatial and temporal gradients of intensity, to compute the local optic flow throughout the image. The image flow constraint equation $I_x U + I_y V + I_t = 0$ is approximated at each pixel location by algebraic finite difference approximations in space and time (Horn & Schunk, 1981). Systems of these finite difference equations are then solved for the local image velocity $(U, V)$ within each $4 \times 4$ pixel block within the image. The images of the rotating objects were generated using OpenGL.

In experiment 5, we investigated the learning of translation-invariant representations of the optic flow vector fields generated by clockwise versus anticlockwise rotation of the tetrahedron stimulus illustrated in Figure 7a. The network was trained with the two optic flow patterns generated in nine different locations, as in experiments 2 and 3. The flow fields used to train the network were generated by the object rotating through one degree of angle. The single cell information measures (see Figure 7b) and multiple cell information measures (see Figure 7c) (using the same conventions as in Figure 3) show that the maximal information, one bit, was reached by single cells and with the multiple cell information measure. The dashed line shows the control condition of a network with random untrained connectivity.

This experiment shows that the model can operate well and learn translation-invariant representations with motion flow fields actually extracted from the successive images produced by a rotating object.

**3.6 Experiment 6: Optic Flow Analysis of Real Images: Rotation Invariance.** In experiment 6 we investigated the learning of rotation-invariant representations of the optic flow vector fields generated by clockwise versus anticlockwise rotation of the spoked wheel stimulus illustrated in Figure 8a. (The algorithm for generating the optic flow field is described in section 3.5.) The radius of the spoked wheel was 50 pixels on the $128 \times 128$ background. The rotation was in-plane, and the optic flow fields used as an input to the network were extracted from the changing images, each separated by one degree of the object as it rotated through 360 degrees. The single cell information measures (see Figure 8b) and multiple cell information measures (see Figure 8c) (using the same conventions as in Figure 3) show that the maximal information, one bit, was almost reached by single cells and by the multiple cell information measure. The dashed

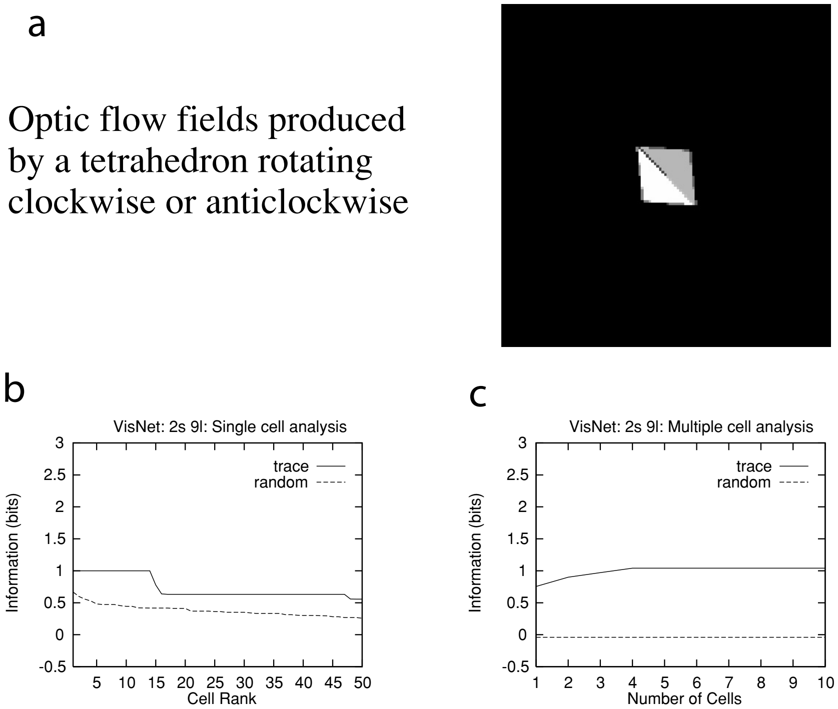


Figure 7: Experiment 5. Translation-invariant representations of the optic flow vector fields generated by clockwise versus anticlockwise rotation of the tetrahedron stimulus illustrated. The optic flow field used as an input to the network was extracted from the changing images of the object as it rotated. The single cell information measures (b) and multiple cell information measures (c) (using the same conventions as in Figure 3) show that the maximal information, 1 bit, was reached by both single cells and in the multiple cell information measure. The dashed line shows the control condition of a network with random untrained connectivity.

line shows the control condition of a network with random untrained connectivity.

This experiment shows that the model can operate well and learn rotation-invariant representations with motion flow fields actually extracted from a very large number of the successive images produced by a rotating object. Because of the large number of closely spaced training images used in this simulation, it is likely that the crucial type of learning was continuous transformation learning (Stringer, Perry, Rolls, & Proske, 2006). Consistent with this, the learning rate was set to the lower value of $7.2 \times 10^{-5}$ for all layers for experiment 6 (cf. Stringer et al., 2006).

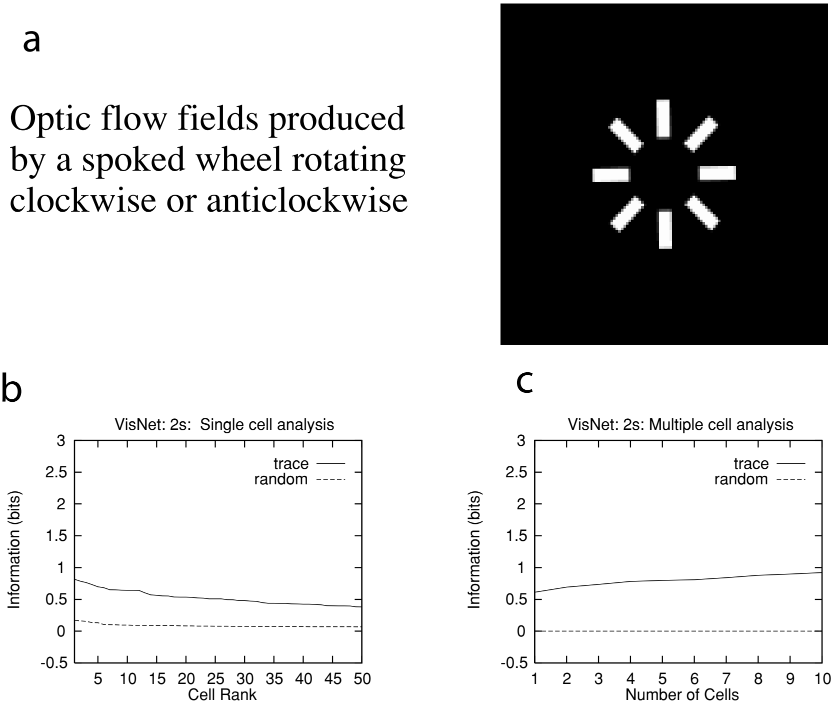


Figure 8: Experiment 6. In-plane rotation-invariant representations of the optic flow vector fields generated by a spoked wheel rotating clockwise or anticlockwise. The optic flow field used as an input to the network was extracted from the changing images of the object as it rotated through 360 degrees, each separated by 1 degree. The single cell information measures (b) and multiple cell information measures (c) (using the same conventions as in Figure 3) show that the maximal information, 1 bit, was reached by both single cells and in the multiple cell information measure. The dashed line shows the control condition of a network with random untrained connectivity.

**3.7 Experiment 7: Generalization to Untrained Images.** To investigate whether the representations of object-based motion such as circular rotation learned with the approach introduced in this article would generalize usefully to the flow fields generated by other objects moving in the same way, we trained the network on the optic flow vector fields generated by clockwise versus anticlockwise rotation of the spoked wheel stimulus illustrated in Figure 8. The training images rotated through 90 degrees in 1 degree steps. We then tested generalization to the new, untrained image shown in Figure 9a. The single and multiple cell information plots in Figure 9b show that information was available about the direction of

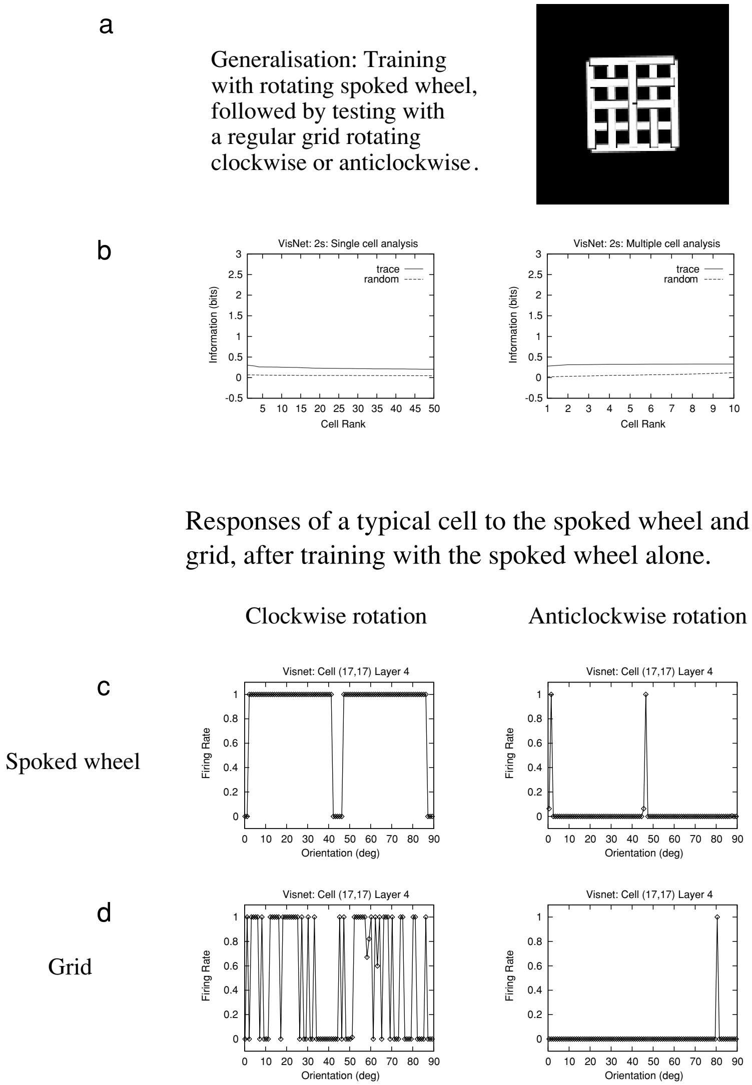


Figure 9: Experiment 7. Generalization to untrained images. The network was trained on the optic flow vector fields generated by the spoked wheel stimulus illustrated in Figure 8 rotating clockwise or anticlockwise. (a) Generalization to the new untrained image shown at the top right of the Figure was then tested. (b) The single and multiple cell information plots show that information was available about the direction of rotation (clockwise versus anticlockwise) of the untrained test images. (c) The firing rate of a fourth layer cell to the clockwise and anticlockwise rotations of the trained image illustrated in Figure 8. (d) The firing rate of the same fourth layer cell to the clockwise and anticlockwise rotations of the untrained image illustrated in Figure 9a.

rotation (clockwise versus anticlockwise) of the untrained test images. Although the information was not as high as 1 bit, which would have indicated perfect generalization, individual cells did generalize usefully to the new images, as shown in Figures 9c and 9d. For example, Figure 9c shows the firing rate of a fourth layer cell to the clockwise and anticlockwise rotations of the trained image illustrated in Figure 8. Figure 9d shows the firing rate of the same fourth-layer cell to the clockwise and anticlockwise rotations of the untrained image illustrated in Figure 9a. The neuron responded correctly to almost all the anticlockwise rotation shifts, and correctly to many of the clockwise rotation shifts, though some noise was evident in the responses of the neuron to the untrained images. Overall, the results demonstrate useful generalization after training with one object to testing with an untrained, different, object on the ability to represent rotation.

## 4 Discussion

We have presented a hierarchical feature analysis theory of the operation of parts of the dorsal visual system, which provides a computational account for how transform-invariant representations of the flow fields generated by moving objects could be formed in the cerebral cortex. The theory uses a modified Hebb rule with a short-term temporal trace of preceding activity to enable whatever is invariant at any stage of the dorsal motion system across short time intervals to be associated together. The theory can account for many types of invariance and has been tested by simulation for position and size invariance. The simulations show that the network can develop global planar representations from noisy local motion inputs (experiment 1), invariant representations of rotating optic flow fields (experiment 2), invariant representations of looming optic flow fields (experiment 3), and invariant representations of asymmetrical objects rotating about one of their axes (experiment 4). These are fundamental problems in motion analysis, and they have all been studied neurophysiologically, including local versus planar motion (Movshon et al., 1985; Newsome et al., 1989); position-invariant representation of rotating flow fields and looming (Lagae, Maes, Raiguel, Xiao, & Orban, 1994); and object-based rotation (Hasselmo et al., 1989; Sakata et al., 1986). The model thus shows principles by which the different types of motion-related invariant neuronal responses in the dorsal cortical visual system could be produced.

The theory is unifying in the sense that the same theory, but with different inputs, can account for invariant representations of objects in the ventral visual system (Rolls, 1992; Wallis & Rolls, 1997; Elliffe, Rolls, & Stringer, 2002; Rolls & Deco, 2002). It is a strength of the unifying concept introduced in this article that the same hierarchical network that can perform computations of the type important in the ventral visual system can also perform computations of a type important in the dorsal visual system.

Our simulations support the hypothesis that the different response properties of MT and MST neurons from V1 neurons are determined in part by the sizes of their receptive fields, with a larger receptive field needed to analyze some global motion patterns. Similar conclusions were drawn from simulation experiments performed by Sereno (1989) and Sereno and Sereno (1991). This type of self-organization can occur with a Hebbian associative learning rule operating on the feedforward connections to a competitive network. However, experiment 1 showed that even for the computation of planar global motion in intermediate layers such as MT, a trace-based associative learning rule is better than a purely associative Hebbian rule with noisy (probabilistic) local motion inputs, because the trace rule allows temporal averaging to contribute to the learning. In experiments 2 and 3, the trace rule is crucial to the success of the learning, in that the stimuli when presented in different training locations did not overlap, so that the only process by which the different transforms can be linked is by the temporal trace learning rule implemented in the model (Rolls & Milward, 2000; Rolls & Stringer, 2001). (We note that in a new development, it has been shown that if different transforms of the training stimuli do overlap continuously in space, then this overlap can provide a useful learning principle for invariant representations to be formed and requires only associative synaptic modification; Stringer et al., 2006. It would be of interest to extend this concept, which has been applied to the ventral visual system, to the dorsal visual system.)

One type of perceptual analysis that can be understood with the theory and simulations described here is how neurons can self-organize to respond to the motion inputs produced by small objects when they are seen on different parts of the retina. This is achieved by using memory-trace-based synaptic modification in the type of architecture illustrated in Figure 4a. The crucial stage for this learning is the top layer in Figure 4a labeled Layer 2/3. The forward connections to the neurons in this layer can form the required representation if they use a trace or similar learning rule, and the object motion occurs with some temporospatial continuity. (Temporospatial continuity has been shown to be important in human face invariance learning [Wallis & Bulthoff, 2001], and spatial continuity over continuous transforms may be a useful learning principle [Stringer et al., 2006].) This aspect of the architecture is what is formally similar to the architecture of the ventral visual system, which can learn invariant representations of stationary objects. The only difference required of the networks is that the ventral visual stream network should receive inputs from neurons that respond to stationary features such as lines or edges and that the dorsal visual stream network should receive inputs from neurons that respond to local motion cues. It is this concept that allows us to propose that there is a unifying hypothesis that applies to some of the computations performed by both the ventral and the dorsal visual streams.

The way in which position-invariant representations in the model develop is illustrated in Figure 4a, where, in the top layer labeled layer 3, individual neurons receive information from different parts of layer 2, where different neurons can represent the same object motion but in different parts of visual space. In the model, layer 2 can thus be thought of as corresponding to some neurons in area MT, in which direction selectivity for elementary optic flow components such as rotation, deformation, and expansion and contraction is not position invariant (Lagae et al., 1994). Layer 3 in the model can in the same way be thought of as corresponding to area MST, in which direction selectivity for elementary optic flow components such as rotation, deformation, and expansion and contraction is position invariant for 40% of neurons (Lagae et al., 1994). A further correspondence between the model and the brain is that neurons that respond to global planar motion are found in the brain in area MT (Movshon et al., 1985; Newsome et al., 1989) and in the model in layer 1, whereas neurons in V1 and V2 do not respond to global motion (Movshon et al., 1985; Newsome et al., 1989), and correspond in the model to the input layer of Figure 4a.

Another type of perceptual analysis that can be understood with the theory and simulations described here is the object-based view-independent representation of objects, exemplified by the ability to see that an "ended" object is rotating clockwise about one of its axes. It was shown in experiment 4 that these representations can be formed by combining information from both the dorsal visual stream (about global motion) and the ventral visual stream (about object shape and/or luminance features). For these representations to be learned, a trace associative or similar learning rule must be used while the object transforms from one view to another (e.g., from upright to inverted).

A hierarchical network with the general architecture shown in Figure 2 with separate analyses of form and motion that are combined at a final stage (as in experiment 4) is also useful for biological motion, such as representing a person walking (Giese & Poggio, 2003). However, the network described by Giese and Poggio is not very biologically plausible, in that it performs MAX functions to help with the computational issue of transform invariance and does not self-organize on the basis of the inputs so that it must be largely hand-wired. The issue here is that Giese and Poggio suppose that a MAX function is performed to select the maximally active afferent to a neuron, but there is no account of how afferents of just one type (e.g., a bar with a particular orientation and contrast) are being received by a given neuron. Not only is no principle suggested by which this could be achieved, but also no learning algorithm is given to achieve this. We suggest therefore that it would be of interest to investigate whether the more biologically plausible self-organizing type of network described in this article can learn on the basis of the inputs being received to respond to biological motion. To do this, some form of sequence sensitivity would be useful.

The theory described here is appropriate for the global motion analysis required to analyze the flow fields of objects as they translate, rotate, expand (loom), or contract, as shown in experiments 1 to 3. The theory thus provides a model of some of the computations that appear to occur along the pathway V1–V2–MT–MST, as neurons of these types are generated along this pathway (see section 1). The theory described here can also account for global motion in an object-based coordinate frame as shown in experiment 4. Neurons with these properties have been found in the cortex in the anterior part of the macaque superior temporal sulcus, in which neurons respond to a head when it is rotating clockwise about its own axis but not counterclockwise, regardless of whether it is upright or inverted (Hasselmo et al., 1989). The result of the inversion experiment shows that these neurons are not just responding to global flow across the visual field, but are taking into account information about the shape and features of the object. Area STPa (the cortex in the anterior part of the macaque superior temporal sulcus) contains neurons that respond to a rotating sphere (Anderson & Siegel, 2005), and as shown in experiment 4, the present theory could account for such neurons. Whether the present model could account for the structure from motion also observed for these neurons is not yet known. The theory could also account for neurons in area 7a of the parietal cortex that may also respond to motion of an object about one of its axes in an object-based way (Sakata et al., 1986). Neurons have also been found in the primary motor cortex (M1) that respond similarly to neurons in area 7a when a monkey is solving a visually presented maze (Crowe, Chafee, Averbeck, & Georgopoulos, 2004), but their visual properties are not sufficiently understood to know whether the present model might apply. Area LIP contains neurons that perform processing related to saccadic eye movements to visual targets (Andersen, 1997), and the present theory may not apply to this type of processing.

The model of processing utilized here in a series of hierarchically organized competitive networks with convergence at each stage (as illustrated in Figure 2) is intended to capture some of the main anatomical and physiological characteristics of the ventral visual stream of visual cortical areas, and is intended to provide a model for how processing in these areas could operate, as described in detail elsewhere (Rolls & Deco, 2002; Rolls & Treves, 1998). To enable learning along this pathway to result by self-organization in the correct representations being formed, associative learning using a short-term memory trace has been proposed (Rolls, 1992; Wallis & Rolls, 1997; Rolls & Milward, 2000; Rolls & Stringer, 2001; Rolls & Deco, 2002). Another approach used in continuous transformation learning utilizes associative learning without a temporal trace and relies on close exemplars of stimuli being provided during the training (Stringer et al., 2006). What we propose here is that similar connectivity and learning processes in the series of cortical pathways in the dorsal visual stream that includes V1–V2–MT–MST and onward connections to the cortex in the superior temporal sulcus and area 7a could account for the invariant representations of the flow fields produced by moving objects.

In relation to the number of stimuli that could be learned by the system, we note that the network simulated is relatively small and was designed to illustrate the new computational hypotheses introduced here rather than to analyze the capacity of such feature hierarchical systems. We note in particular that the network simulated has 1024 neurons in each layer and 100 inputs to each neuron in layers 2 to 4. In contrast, it has been estimated that perhaps half of the macaque brain is involved in visual processing, and typically each neuron has on the order of $10^4$ inputs. It will be of interest using much larger simulations in the future to address capacity issues of this class of network. However, we note that because the network can generalize to rotational flow fields generated by untrained stimuli, as shown in experiment 7, separate representations for the flow fields generated by every object may not be required, and this helps to reduce the number of separate representations that the network may be required to learn.

In contrast to some other theories, the theory developed here utilizes a single unified approach to self-organization in the dorsal and ventral visual systems. Predictions of the theory described here include the following. First, use of a trace rule in the dorsal as well as ventral visual system is predicted. (Thus, differences in, for example, the time constants of NMDA receptors, or persistent poststimulus firing, either of which could implement a temporal trace, would not be expected.) Second, a feature hierarchy is a useful way for understanding details of the operation of the ventral visual system, but can now be used as a clarifying concept for how the details of representations in the dorsal visual system may be built. Third, the theory predicts that neurons specialized for motion detection by using differences in the arrival times of sensory inputs from different retinal locations need occur at only one stage of the system (e.g., in V1) and need not occur elsewhere in the dorsal visual system. These are labeled as local motion neurons in Figure 4a.

## Acknowledgments

This research was supported by the Wellcome Trust and by the Medical Research Council.

## References

Andersen, R. A. (1997). Multimodal integration for the representation of space in the posterior parietal cortex. *Philosophical Transactions of the Royal Society of London B, 352*, 1421–1428.

Anderson, K. C., & Siegel, R. M. (2005). Three-dimensional structure-from-motion selectivity in the anterior superior temporal polysensory area, STPa, of the behaving monkey. *Cerebral Cortex, 15*, 1299–1307.

Bair, W., & Movshon, J. A. (2004). Adaptive temporal integration of motion in direction-selective neurons in macaque visual cortex. *Journal of Neuroscience, 24*, 7305–7323.

Bartlett, M. S., & Sejnowski, T. J. (1998). Learning viewpoint-invariant face representations from visual experience in an attractor network. *Network: Computation in Neural Systems, 9*, 399–417.

Crowe, D. A., Chafee, M. V., Averbeck, B. B., & Georgopoulos, A. P. (2004). Participation of primary motor cortical neurons in a distributed network during maze solution: Representation of spatial parameters and time-course comparison with parietal area 7a. *Experimental Brain Research, 158*, 28–34.

Deco, G., & Rolls, E. T. (2005). Neurodynamics of biased competition and cooperation for attention: A model with spiking neurons. *Journal of Neurophysiology, 94*, 295–313.

Desimone, R. (1991). Face-selective cells in the temporal cortex of monkeys. *Journal of Cognitive Neuroscience, 3*, 1–8.

Duffy, C. J. (2004). The cortical analysis of optic flow. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences* (Vol. 2, pp. 1260–1283). Cambridge, MA: MIT Press.

Duffy, C. J., & Wurtz, R. H. (1996). Optic flow, posture, and the dorsal visual pathway. In T. Ono, B. L. McNaughton, S. Molotchnikoff, E. T. Rolls, & H. Nishijo (Eds.), *Perception, memory and emotion: frontiers in neuroscience* (pp. 63–77). Cambridge: Cambridge University Press.

Elliffe, M. C. M., Rolls, E. T., & Stringer, S. M. (2002). Invariant recognition of feature combinations in the visual system. *Biological Cybernetics, 86*, 59–71.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation, 3*, 194–200.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36*, 193–202.

Geesaman, B. J., & Andersen, R. A. (1996). The analysis of complex motion patterns by form/cue invariant MSTd neurons. *Journal of Neuroscience, 16*, 4716–4732.

Giese, M. A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience, 4*, 179–192.

Graziano, M. S. A., Andersen, R. A., & Snowden, R. J. (1994). Tuning of MST neurons to spiral motions, *Journal of Neuroscience, 14*, 54–67.

Hasselmo, M. E., Rolls, E. T., Baylis, G. C., & Nalwa, V. (1989). Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research, 75*, 417–429.

Horn, B. K. P., & Schunk, B. G. (1981). Determining optic flow. *Artificial Intelligence, 17*, 185–203.

Lagae, L., Maes, H., Raiguel, S., Xiao, D.-K., & Orban, G. A. (1994). Responses of macaque STS neurons to optic flow components: A comparison of areas MT and MST. *Journal of Neurophysiology, 71*, 1597–1626.

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience, 19*, 577–621.

Movshon, J. A., Adelson, E. H., Gizzi, M. S., & Newsome, W. T. (1985). The analysis of moving visual patterns. In C. Chagas, R. Gattass, & C. Gross (Eds.), *Pattern recognition mechanisms* (pp. 117–151). New York: Springer-Verlag.

Newsome, W. T., Britten, K. H., & Movshon, J. A. (1989). Neuronal correlates of a perceptual decision. *Nature, 341*, 52–54.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*, 1019–1025.

Rolls, E. T. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philosophical Transactions of the Royal Society, 335*, 11–21.

Rolls, E. T. (2000). Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron, 27*, 205–218.

Rolls, E. T. (2006). The representation of information about faces in the temporal and frontal lobes of primates including humans. *Neuropsychologia*, PMID = 16797609.

Rolls, E. T., & Deco, G. (2002). *Computational neuroscience of vision*. New York: Oxford University Press.

Rolls, E. T., & Milward, T. (2000). A model of invariant object recognition in the visual system: Learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Computation, 12*, 2547–2572.

Rolls, E. T., & Stringer, S. M. (2001). Invariant object recognition in the visual system with error correction and temporal difference learning. *Network: Computation in Neural Systems, 12*, 111–129.

Rolls, E. T., & Tovee, M. J. (1994). Processing speed in the cerebral cortex and the neurophysiology of visual masking. *Proceedings of the Royal Society, B, 257*, 9–15.

Rolls, E. T., Tovee, M. J., Purcell, D. G., Stewart, A. L., & Azzopardi, P. (1994). The responses of neurons in the temporal cortex of primates, and face identification and detection. *Experimental Brain Research, 101*, 474–484.

Rolls, E. T., & Treves, A. (1998). *Neural networks and brain function*. New York: Oxford University Press.

Rolls, E. T., Treves, A., & Tovee, M. J. (1997). The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. *Experimental Brain Research, 114*, 149–162.

Rolls, E. T., Treves, A., Tovee, M., & Panzeri, S. (1997). Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. *Journal of Computational Neuroscience, 4*, 309–333.

Sakata, H., Shibutani, H., Ito, Y., & Tsurugai, K. (1986). Parietal cortical neurons responding to rotary movement of visual space stimulus in space. *Experimental Brain Research, 61*, 658–663.

Seltzer, B., & Pandya, D. N. (1978). Afferent cortical connections and architectonics of the superior temporal sulcus and surrounding cortex in the rhesus monkey. *Brain Research, 149*, 1–24.

Sereno, M. I. (1989). Learning the solution to the aperture problem for pattern motion with a Hebb rule. In D. Touretzky (Ed.), *Advances in neural information processing systems, 1* (pp. 468–476). San Mateo, CA: Morgan Kaufmann.

Sereno, M. I., & Sereno, M. E. (1991). Learning to see rotation and dilation with a Hebb rule. In D. Touretzky & R. Lippmann (Eds.), *Advances in neural information processing systems 3* (pp. 320–326). San Mateo, CA: Morgan Kaufmann.

Stringer, S. M., Perry, G., Rolls, E. T., & Proske, J. H. (2006). Learning invariant object recognition in the visual system with continuous transformations. *Biological Cybernetics, 94*, 128–142.

Stringer, S. M., & Rolls, E. T. (2000). Position invariant recognition in the visual system with cluttered environments. *Neural Networks, 13*, 305–315.

Stringer, S. M., & Rolls, E. T. (2002). Invariant object recognition in the visual system with novel views of 3D objects. *Neural Computation, 14*, 2585–2596.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience, 19*, 109–139.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of visual behavior* (pp. 549–586). Cambridge, MA: MIT Press.

Wallis, G., & Bulthoff, H. H. (2001). Effects of temporal assocation on recognition memory. *Proceedings of the National Academy of Sciences, 98*, 4800–4804.

Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology, 51*, 167–194.

Wurtz, R. H., & Kandel, E. R. (2000). Perception of motion depth and form. In E. R. Kandel, J. H. Schwartz, & T. M. Jessell (Eds.), *Principles of neural science*, 4th ed. (pp. 548–571). New York: McGraw-Hill.

Received January 11, 2006; accepted May 15, 2006.