Invariant Object Recognition in the Visual System with Novel Views of 3D Objects

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

To form view-invariant representations of objects, neurons in the inferior temporal cortex may associate together different views of an object, which tend to occur close together in time under natural viewing conditions. This can be achieved in neuronal network models of this process by using an associative learning rule with a short-term temporal memory trace. It is postulated that within a view, neurons learn representations that enable them to generalize within variations of that view. When three-dimensional (3D) objects are rotated within small angles (up to, e.g., 30 degrees), their surface features undergo geometric distortion due to the change of perspective. In this article, we show how trace learning could solve the problem of in-depth rotation-invariant object recognition by developing representations of the transforms that features undergo when they are on the surfaces of 3D objects. Moreover, we show that having learned how features on 3D objects transform geometrically as the object is rotated in depth, the network can correctly recognize novel 3D variations within a generic view of an object composed of a new combination of previously learned features. These results are demonstrated in simulations of a hierarchical network model (VisNet) of the visual system that show that it can develop representations useful for the recognition of 3D objects by forming perspective-invariant representations to allow generalization within a generic view.

1 Introduction

There is now much evidence demonstrating that over successive stages, the visual system develops neurons that respond with view, size, and position (translation) invariance to objects or faces (Desimone, 1991; Rolls, 1992, 2000; Rolls & Tovee, 1995; Tanaka, Saito, Fukada, & Moriya, 1991; Tanaka, 1996; Logothetis, Pauls & Poggio, 1995; Booth & Rolls, 1998). Rolls (1992, 2000) has proposed a biologically plausible mechanism to explain this behavior based on the following: (1) a series of hierarchical competitive networks with local graded inhibition, (2) convergent connections to each neuron from a topologically corresponding region of the preceding layer, leading to an increase in the receptive field size of cells through the visual processing areas, and (3) synaptic plasticity based on a modified Hebb-like learning rule with a temporal trace of each cell's previous activity. The idea underlying the trace learning rule is to learn from the natural statistics of real-world visual input, where, for example, the successive transformed versions of the same image tend to occur close together in time (Földiák, 1991; Rolls, 1992; Rolls & Tovee, 1995; Rolls & Deco, 2002). Rolls's hypothesis about the functional architecture and operation of the ventral visual system (the "what" pathway, in which representations of objects are formed) was tested in a model, VisNet, of ventral stream cortical visual processing, where it was found that invariant neurons did indeed develop as long as the Hebbian learning rules incorporated a trace of recent cell activity, where the trace is a form of temporal average (Wallis & Rolls, 1997). The types of invariance demonstrated included translation, size, and the view of an object.

This hypothesis of how object recognition could be implemented in the brain postulates that trace rule learning helps invariant representations to form in two ways (Rolls, 1992, 2000; Rolls & Deco, 2002). The first process enables associations to be learned between different generic 3D views of an object where there are different qualitative shape descriptors. One example of different generic views is usually provided by the front and back of an object. Another example is when different surfaces come into view, and new surfaces define the viewed boundary, when most 3D objects are rotated in three dimensions. For example, a catastrophic rearrangement of the shape descriptors (Koenderink, 1990) occurs when a cup is tilted so that one can see inside it. The second process is that within a generic view, as the object is rotated in depth, there will be no catastrophic changes in the qualitative 3D boundary shape descriptors, but instead the quantitative (metric) values of the shape descriptors alter (see further Koenderink, 1990; Biederman, 1987). For example, while the cup is being rotated within a generic view seen from somewhat below, the curvature of the cusp forming the top boundary will alter, but the qualitative shape descriptor will remain a cusp. In addition to the changes of the metric values of the object boundary shape descriptors that occur within a generic view, the surface features on a 3D object also undergo geometric transforms as it rotates in depth, as described below and illustrated in Figure 3. Trace learning could potentially help with both the between- and within-generic view processes. That is, trace learning could help to associate together qualitatively different sets of shape descriptors that occur close together in time and describe the generically different views of a cup. Trace learning could also help with the second process and learn to associate together the different quantitative (or metric) values of shape descriptors that typically occur when objects are rotated within a generic view.

The main aim of this article is to show that trace learning in an appropriate architecture can indeed learn the geometric shape transforms that are characteristic of the surface markings on objects as they rotate within a generic view and use this knowledge in invariant object recognition. We are able to address this particular issue here by using an object, a sphere, in which there are no catastrophic changes between different generic views as the object is rotated or metric changes in the defining boundary of the object as it is rotated. We also show here that once the network has learned the types of geometric transform characteristic of features on the surface of 3D objects, the network can generalize to novel views of the object when the object has been shown in only a limited number of views, and the surface markings are composed of new combinations of the previously learned features.

Examples of the types of perspective transforms in the surface markings on 3D objects that are typically encountered as the objects rotate within a generic view and that we investigated in this article are shown in Figure 3. The surface markings on the sphere that consist of combinations of three features in different spatial arrangements undergo characteristic transforms as the sphere is rotated from 0 degree toward −60 degrees and +60 degrees. Each object is identified by surface markings that consist of a different spatial arrangement of the same three features (a horizontal, vertical, and diagonal line, which becomes an arc on the surface of the object). Boundary shape changes and shading and stereo cues are excluded from the stimuli, so that the invariant learning must be about the surface marking transforms.

An interesting issue about the properties of feature hierarchy networks used for learning invariant representations is whether they can generalize to transforms of objects on which they have not been trained, that is, whether they assume that an initial exposure is required during learning to every transformation of the object to be recognized (Wallis & Rolls, 1997; Rolls & Deco, 2002; Riesenhuber & Poggio, 1998). We show here that this is not the case, for such feature hierarchy models can generalize to novel within-generic views of an object. VisNet can achieve this when it is given invariant training on the features from which the new object will be composed. After the new object has been shown in some views, VisNet generalizes correctly to other views of the object. This part of the research described here builds on an earlier result of Elliffe, Rolls, and Stringer (2002) that was tested with a purely isomorphic transform, translation.

2 Methods

2.1 The VisNet Model. In this section, we provide an overview of the VisNet model. Further details may be found in Wallis and Rolls (1997), Rolls and Milward (2000), and Rolls and Deco (2002). The simulations performed here use the latest version of the VisNet model (VisNet2), with the same model parameters that Rolls and Milward (2000) used. VisNet is a four-layer feedforward network of the primate ventral visual system, with the separate layers corresponding to V2, V4, the posterior inferior temporal cor-

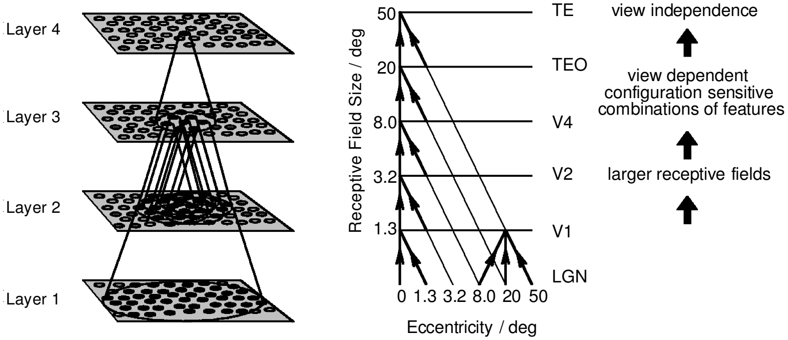

Figure 1: (Left) Stylized image of the VisNet four-layer network. Convergence through the network is designed to provide fourth-layer neurons with information from across the entire input retina. (Right) Convergence in the visual system (adapted from Rolls, 1992). V1: visual cortex area V1; TEO: posterior inferior temporal cortex; TE: inferior temporal cortex (IT).

tex, and the anterior inferior temporal cortex, as shown in Figure 1. For each layer, the connections to individual cells are derived from a topologically corresponding region of the preceding layer, with connection probabilities based on a gaussian distribution. 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 activation of neurons within a layer is convolved with a local spatial filter that operates over several pixels. Next, contrast enhancement is applied by means of a sigmoid activation function where the sigmoid threshold is adjusted to control the sparseness of the firing rates.

The trace learning rule (Földiák, 1991; Rolls, 1992; Wallis & Rolls, 1997) encourages neurons to develop invariant responses to input patterns that tended to occur close together in time, because these are likely to be from the same object. The rule used was

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

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: jth input to the neuron

$\overline{y}^{\tau}$: Trace value of the output of the neuron at time step τ

w_j: Synaptic weight between jth input and the neuron

y: Output from the neuron

α: Learning rate, annealed between unity and zero

η: Trace value; the optimal value varies with presentation sequence length

The parameter η may be set anywhere in the interval [0, 1], and for the simulations described here, η was set to 0.8. (A discussion of the good performance of this rule and its relation to other versions of trace learning rules is provided by Rolls & Milward, 2000, and Rolls & Stringer, 2001.)

2.2 Training and Test Procedure. The stimuli were designed to allow higher-order representations (of combinations of three features) to be built from lower-order feature representations (of pairs of features) (see Elliffe et al., 2002). The stimuli take the form of images of surface features on 3D rotating spheres, with each image presented to VisNet's retina being a 2D projection of the surface features of one of the spheres. (For the actual simulations described here, the surface features and their deformations were what VisNet was trained and tested with, and the remaining blank surface of each sphere was set to the same gray scale as the background.) Each stimulus is uniquely identified by two or three surface features, where the surface features are (1) vertical, (2) diagonal, and (3) horizontal arcs and where each feature may be centered at three different spatial positions, designated A, B, and C, as shown in Figure 2. The stimuli are thus defined in terms of what features are present and their precise spatial arrangement with respect to each other. We refer to the two- and three-feature stimuli as pairs and triples, respectively. Individual stimuli are denoted by three numbers that refer to the individual features present in positions A, B, and C, respectively. For example, a stimulus with positions A and C containing a vertical and diagonal bar, respectively, would be referred to as stimulus 102, where the 0 denotes no feature present in position B. In total, there are 18 pairs (120, 130, 210, 230, 310, 320, 012, 013, 021, 023, 031, 032, 102, 103, 201, 203, 301, 302) and 6 triples (123, 132, 213, 231, 312, 321).

Further image construction details are as follows, with an illustration of the images used shown in Figure 3. In the experiments presented later, the end points of the individual surface features subtend an angle of approximately 10 degrees with respect to the center of the spheres. In addition, each pair of spatial positions (A, B, C) subtends an angle of approximately 15 degrees with respect to the centers of the spheres. The 2D projections of the stimuli are scaled to be 128 × 128 pixels, and these are placed on a blank background and preprocessed by a set of input filters before the final images are presented to VisNet's 128 × 128 pixel input retina. The input filters accord with the general tuning profiles of simple cells in V1 (see Rolls & Milward, 2000, for more details).

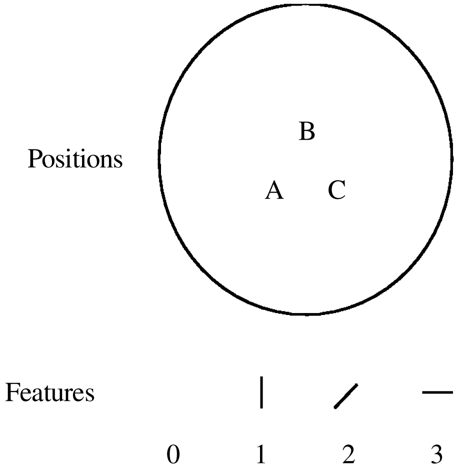

Figure 2: Details of the 3-dimensional visual stimuli used in the rotation invariance experiments of section 3. Each stimulus is a sphere that is identified by its unique combination of two or three surface features. There are three possible types of features that take the form of (1) a vertical arc, (2) a diagonal arc, and (3) a horizontal arc and lie on the surfaces of the spheres (a blank space is denoted by 0). These three features can be placed in any of three relative positions A, B, and C, which are themselves separated by 15 degrees of arc.

To train the network, each stimulus is presented to VisNet in a randomized sequence of five orientations with respect to VisNet's input retina, where the different orientations are obtained from successive in-depth rotations of the stimulus through 30 degrees. That is, each stimulus is presented to VisNet's retina from the following rotational views: (i) −60 degrees, (ii) −30 degrees, (iii) 0 degree (central position with surface features facing directly toward VisNet's retina), (iv) 30 degrees, and (v) 60 degrees. Figure 3 shows representations of the six visual stimuli with three surface features (triples) presented to VisNet during the simulations. Each row shows one of the stimuli rotated through the five different rotational views in which the stimulus is presented to VisNet. At each 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 training orientations, a new stimulus is chosen at random, and the process is repeated. The presentation of all the stimuli through all five 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 through 4 were 50, 100, 100, and 75, respectively.

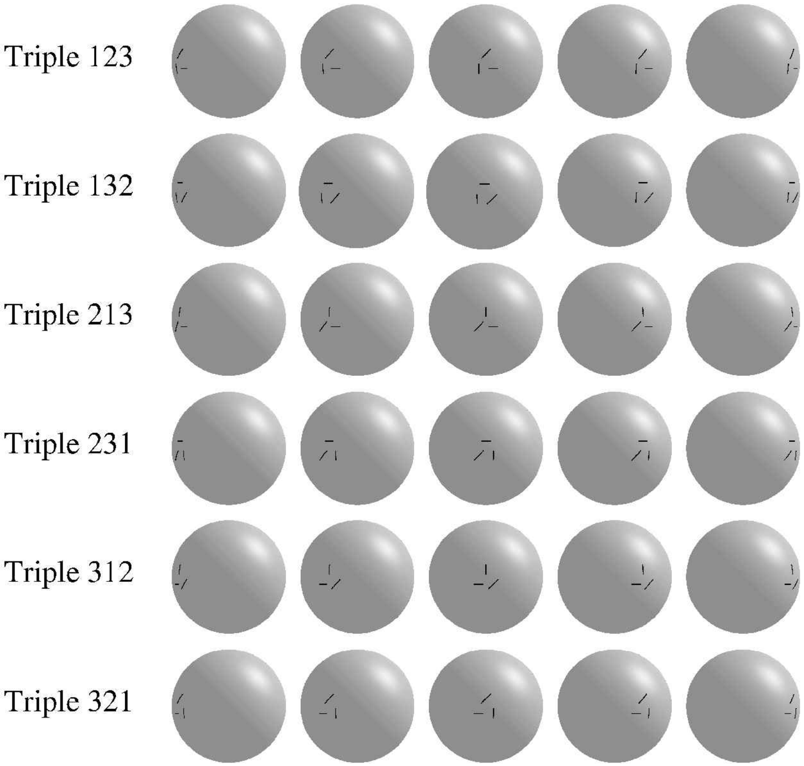

Figure 3: Representations of the six visual stimuli with three surface features (triples) presented to VisNet during the simulations. Each stimulus is a sphere that is identified by a unique combination of three surface features (a vertical, diagonal, and horizontal arc) that occur in three relative positions A, B, and C. Each row shows one of the stimuli rotated through the five different rotational views in which the stimulus is presented to VisNet. From left to right, the rotational views shown are −60 degrees, −30 degrees, 0 degree (central position), 30 degrees, and 60 degrees.

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 independently of view. 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 in Rolls, Treves, and Tovee (1997) and Rolls and Milward (2000). In the experiments presented later, the multiple cell information was calculated from only a small subset of the output cells. There were five cells selected for each stimulus, and these were the five cells that gave the highest single cell information values for that stimulus.

3 Results

The aim of the first experiment was to test whether the network could learn invariant representations of the surface markings on objects seen from different within-generic views and whether the learning would generalize to new views of objects (triples) after pretraining on feature subsets (pairs). To realize this aim, the VisNet network was trained in two stages. In the first stage, the 18 feature pairs were used as input stimuli, with each stimulus being presented to VisNet's retina in sequences of five orientations as described in section 2.2. During this stage, learning was allowed to take place only in layers 1 and 2. This led to the formation of neurons that responded to the feature pairs with some rotation invariance in layer 2. In the second stage, we used the six feature triples as stimuli, with learning allowed only in layers 3 and 4. During this second training stage, the triples were presented to VisNet's input retina only in the first four orientations, i through iv. After the two stages of training were completed, we examined whether the output layer of VisNet had formed top-layer neurons that responded invariantly to the six triples when presented in all five orientations, not just the four in which the triples had been presented during training. To provide baseline results for comparison, the results from experiment 1 were compared with results from experiment 2, which involved no training in layers 1,2 and 3,4, with the synaptic weights left unchanged from their initial random values.

In Figure 4, we present numerical results for the two experiments described. On the left are the single cell information measures for all top (fourth) layer neurons ranked in order of their invariance to the triples, and on the right side are multiple cell information measures. To help interpret these results, we can compute the maximum single cell information measure according to

$$\text{Maximum single cell information} = \log_2 \text{(number of triples)}, \qquad (3.1)$$

where the number of triples is six. This gives a maximum single cell information measure of 2.6 bits for these test cases. The information results from experiment 1 shown in Figure 4 demonstrate that even with the triples presented to the network in only four of the five orientations during training, layer 4 is indeed capable of developing rotation-invariant neurons that can discriminate effectively among the six different feature triples in all five ori-

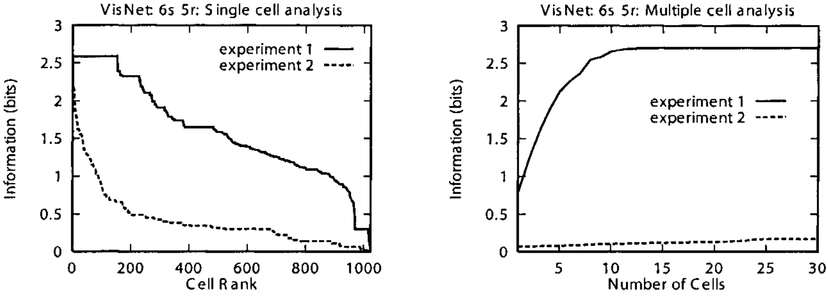

Figure 4: Numerical results for experiments 1 and 2: (Left) Single cell information measures. (Right) Multiple cell information measures. There were six stimuli each shown in five rotational views (6s 5r). In experiment 1, the first two layers were trained on feature pairs in all transforms, and layers 3 and 4 were trained on the six triple stimuli in four of the five transforms. For comparison, experiment 2 shows the performance with no training.

entations, that is, with correct recognition from all five perspectives. Indeed, from the single cell information measures, it can be seen that a number of cells have reached the maximum level of performance in experiment 1. In addition, the multiple cell information for experiment 1 reaches the maximal level of 2.6 bits, indicating that the network as a whole is capable of perfect discrimination between the six triples in any of the five orientations. The finding that some single neurons showed perfect performance on all five instances of every one of the six stimuli (as shown by the single cell information analysis of Figure 4) provides evidence that the network can indeed generalize to novel deformations of stimuli when the first two layers have been trained on the component feature combinations, but the top two layers have been trained on only some of the transforms of the complete stimuli.

Further results from experiment 1 are presented in Figure 5, where we show the response profiles of a top-layer neuron to the six triple-feature stimuli. This neuron has achieved excellent invariant responses to the six triple-feature stimuli. The performance was perfect in that the response profiles are independent of the orientation of the sphere but differentiate between triples in that the responses are maximal for triple 132 and minimal for all other triples. In particular, the cell responses are maximal for triple 132 presented in all five of the orientation transforms. The perfect performance of the neuron occurred even though the network was being tested with five transforms of the stimuli, with only four transforms having been trained in layers 3 and 4.

We performed a control experiment to show that the network really had learned invariant representations specific to the kinds of 3D deformations

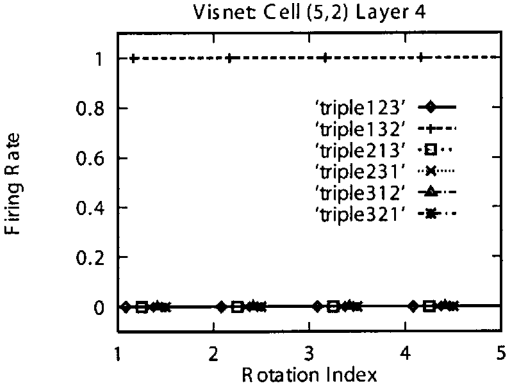

Figure 5: Numerical results for experiment 1: Response profiles of a top-layer neuron to the 6 triples in all 5 orientations.

undergone by the surface features as the objects rotated in depth. The design of the control experiment was to train the network on "spheres" with nondeformed surface features and then to test whether the network failed to operate correctly when it was tested with objects with the features present in the transformed way that they appear on the surface of a real 3D object. The training set of 2D feature triples was generated from the view of each feature triple present at 0 degree, which was 2D translated without deformation to the position on a sphere at which it would appear if the sphere had been rotated. It was found that when VisNet was first trained on undeformed triple stimuli and then tested on the true 3D triple images, that performance was poor, as shown by the single and multiple cell information measures, and by the fact that most layer 4 neurons did not respond to all five instances of each rotated triple stimulus. The results of this control experiment thus showed that the network had learned invariant representations specific to the kinds of 3D deformations undergone by the surface features as the objects rotated in depth in the first experiment.

4 Discussion

In this article, we were able to show first how trace learning can form neurons that can respond invariantly to the perspective transforms that surface markings on 3D objects show when the object is rotated within a generic view. The invariant learning was specifically about surface markings, in that boundary curvature changes were excluded by the use of spherical objects. Thus, VisNet can learn how the surface features on 3D objects transform as the object is rotated in depth and can use knowledge of the characteristics of these transforms to perform 3D object recognition.

Second, the results show that this could occur for a novel view of an object that was not an interpolation from previously seen views. This was possible given that the low-order feature combination sets from which an object was composed had been learned in early layers of VisNet previously. That is, it was shown that invariant learning on every possible transform of the triple-feature stimuli is not necessary. Once the invariance has been trained for the feature pairs, then just a few presentations of the triples (needed to train the weights from the intermediate to the higher layers) are needed, and the invariance is present not because of any invariant learning about the triples but because of invariant learning about the feature pairs.

The approach inherent in a feature hierarchy system such as that exemplified by VisNet is that the representation of an object at an upper level is produced by a combination of the firing of feature-sensitive neurons at lower levels that themselves have some invariant properties (Rolls & Deco, 2002). The concept is thus different from that suggested by Edelman (1999), who used existing invariant representations of different whole objects as a basis set for new whole objects. Effectively, a new whole object neuron at the whole object representation level in the network was trained to respond to linear combinations of the activity of the other pretrained object neurons. Provided that the new object was similar to some of the pretrained objects, the interpolation worked. The system we propose effectively learns invariant representations of low-order feature combinations at an early stage of the network. Then at an upper level, the system is trained on a new object constituted by an entirely new combination of lower-level neurons firing. The system then generalizes invariantly to different transforms of the new object. The system we describe is more closely related to the hierarchical nature of the cortical visual system and is also more powerful in that the new object need not be very similar at the object level to any previously learned object. It must just be composed of similar features.

Acknowledgments

This research was supported by the Wellcome Trust and the MRC Interdisciplinary Research Centre for Cognitive Neuroscience.

References

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*(2), 115–147.

Booth, M. C. A., & Rolls, E. T. (1998). View invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex, 8*, 510–523.

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

Edelman, S. (1999). *Representation and recognition in vision.* Cambridge, MA: MIT 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.

Koenderink, J. J. (1990). *Solid shape.* Cambridge, MA: MIT Press.

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology, 5,* 552–563.

Riesenhuber, M., & Poggio, T. (1998). Just one view: Invariances in inferotemporal cell tuning. In M. I. Jordan, M. J. Kearns & S. A. Solla (Eds.), *Advances in neural information processing systems, 10* (pp. 215–221). Cambridge, MA: MIT Press.

Rolls, E. T. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical areas. *Philosophical Transactions of the Royal Society, London [B], 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., & 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, 12,* 111–129.

Rolls, E. T., & Tovee, M. J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of Neurophysiology, 73,* 713–726.

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,* 177–185.

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

Tanaka, K., Saito, H., Fukada, Y., & Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology, 66,* 170–189.

Wallis, G., & Rolls, E. T. (1997). A model of invariant object recognition in the visual system. *Progress in Neurobiology, 51,* 167–194.

Received December 6, 2000; accepted April 26, 2002.