Available online at www.sciencedirect.com
ScienceDirect
Journal of Physiology - Paris 100 (2006) 43–62
Journal of Physiology Paris
www.elsevier.com/locate/jphysparis

Invariant visual object recognition: A model, with lighting invariance

Edmund T. Rolls *, Simon M. Stringer

Oxford University, Centre for Computational Neuroscience, Department of Experimental Psychology, South Parks Road, Oxford OX1 3UD, England, United Kingdom

Abstract

How are invariant representations of objects formed in the visual cortex? We describe a neurophysiological and computational approach which focusses on a feature hierarchy model in which invariant representations can be built by self-organizing learning based on the statistics of the visual input. The model can use temporal continuity in an associative synaptic learning rule with a short term memory trace, and/or it can use spatial continuity in Continuous Transformation learning. The model of visual processing in the ventral cortical stream can build representations of objects that are invariant with respect to translation, view, size, and in this paper we show also lighting. The model has been extended to provide an account of invariant representations in the dorsal visual system of the global motion produced by objects such as looming, rotation, and object-based movement. The model has been extended to incorporate top-down feedback connections to model the control of attention by biased competition in for example spatial and object search tasks. The model has also been extended to account for how the visual system can select single objects in complex visual scenes, and how multiple objects can be represented in a scene.

Keywords: Ventral visual stream; View invariance; Global motion recognition; Trace learning; Dorsal visual stream; Inferior temporal visual cortex; Attention

1. Introduction

One of the major problems that is solved by the visual system in the cerebral cortex is the building of a representation of visual information which allows recognition to occur relatively independently of size, contrast, spatial frequency, position on the retina, angle of view, lighting, etc.

The neurophysiological findings reviewed elsewhere (Rolls, 1992, 2000, 2006, 2007) and wider considerations on the possible computational properties of the cerebral cortex (Rolls and Treves, 1998; Rolls and Deco, 2002), lead to the following outline working hypotheses on object recognition by visual cortical mechanisms (see Rolls, 1992; Rolls and Deco, 2002). A related approach to invariant object recognition is described by Riesenhuber and Poggio (1999b) (see also Riesenhuber and Poggio, 1999a, 2000) and is a feature hierarchy approach which uses alternate 'simple cell' and 'complex cell' layers in a way analogous to Fukushima (1980). Differences between these approaches are described in Section 3.

Cortical visual processing for object recognition is organized as a set of hierarchically connected cortical regions consisting at least of V1, V2, V4, posterior inferior temporal cortex (TEO), inferior temporal cortex (e.g. TE3, TEa and TEm), and anterior temporal cortical areas (e.g. TE2 and TE1). There is convergence from each small part of a region to the succeeding region (or layer in the hierarchy) in such a way that the receptive field sizes of neurons (e.g. 1 degree near the fovea in V1) become larger by a factor of approximately 2.5 with each succeeding stage (see Fig. 1). Such zones of convergence would overlap continuously with each other (see Fig. 1). This connectivity would be part of the architecture by which translation invariant representations are computed.

* Corresponding author. Tel.: +44 1865 271348; fax: +44 1865 310447.
E-mail address: Edmund.Rolls@psy.ox.ac.uk (E.T. Rolls).
URL: www.cns.ox.ac.uk (E.T. Rolls).

0928-4257/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jphysparis.2006.09.004

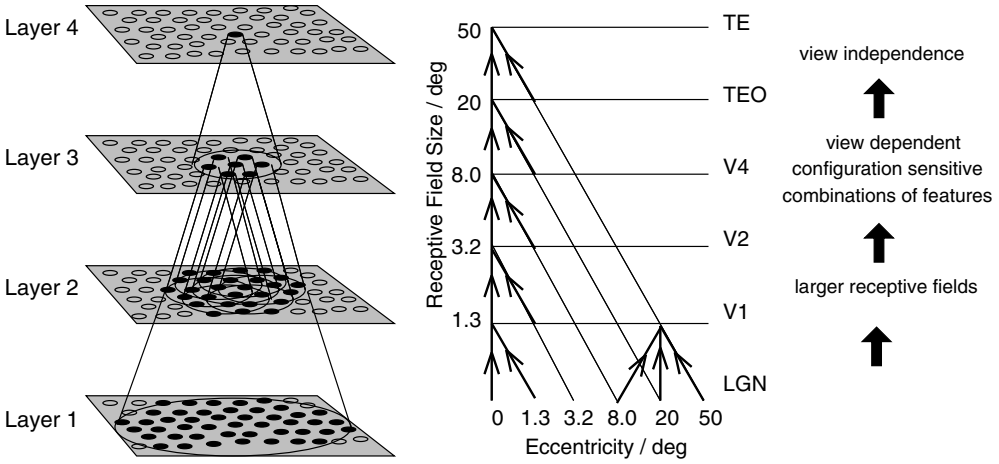

Fig. 1. Convergence in the visual system. Right – as it occurs in the brain. V1: visual cortex area V1; TEO: posterior inferior temporal cortex; TE: inferior temporal cortex (IT). Left – as implemented in VisNet. Convergence through the network is designed to provide fourth layer neurons with information from across the entire input retina.

Each layer is considered to act partly as a set of local self-organizing competitive neuronal networks with overlapping inputs. These competitive nets operate by a single set of forward inputs leading to (typically non-linear, e.g. sigmoid) activation of output neurons; by competition between the output neurons mediated by a set of feedback inhibitory interneurons which receive from many of the principal (in the cortex, pyramidal) cells in the net and project back (via inhibitory interneurons) to many of the principal cells and serve to decrease the firing rates of the less active neurons relative to the rates of the more active neurons; and then by synaptic modification by a modified Hebb rule, such that synapses to strongly activated output neurons from active input axons strengthen, and from inactive input axons weaken (see Hertz et al., 1991; Rolls and Deco, 2002). A biologically plausible form of this learning rule for the change of the synaptic weight δw_{ij} that operates well in such networks is

$$\delta w_{ij} = \alpha y_i (x_j - w_{ij}) \quad (1)$$

where α is a learning rate constant, x_j is the presynaptic firing rate, y_i is the postsynaptic firing rate, and w_{ij} is the synaptic weight or strength (see Rolls and Deco, 2002). Such competitive networks operate to detect correlations between the activity of the input neurons, and to allocate output neurons to respond to each cluster of such correlated inputs.

Translation invariance would be computed in such a system by utilizing competitive learning to detect regularities in inputs when real objects are translated in the physical world. The hypothesis is that because objects have continuous properties in space and time in the world, an object at one place on the retina might activate feature analyzers at the next stage of cortical processing, and when the object was translated to a nearby position, because this would occur in a short period (e.g. 0.5 s), the membrane of the postsynaptic neuron would still be in its 'Hebb-modifiable' state (caused for example by calcium entry as a result of the voltage dependent activation of NMDA receptors), and the presynaptic afferents activated with the object in its new position would thus become strengthened on the still-activated postsynaptic neuron. It is suggested that the short temporal window (e.g. 0.5 s) of Hebb-modifiability helps neurons to learn the statistics of objects moving in the physical world, and at the same time to form different representations of different feature combinations or objects, as these are physically discontinuous and present less regular correlations to the visual system. Földiák (1991) has proposed computing an average activation of the postsynaptic neuron to assist with the same problem. One 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 objects. Another suggestion is that a memory trace for what has been seen in the last 300 ms appears to be implemented by a mechanism as simple as continued firing of inferior temporal neurons after the stimulus has disappeared (Rolls and Tovee, 1994; Rolls et al., 1994). Rolls also suggested that other invariances, for example size, spatial frequency, and rotation invariance, could be learned by a comparable process. It is suggested that this process takes place at each stage of the multiple-layer cortical processing hierarchy, so that invariances are learned first over small regions of space, and then over successively larger regions. This limits the size of the connection space within which correlations must be sought.

Increasing complexity of representations could also be built in such a multiple layer hierarchy by similar mechanisms. At each stage or layer the self-organizing competitive nets would result in combinations of inputs becoming the effective stimuli for neurons. In order to avoid the combinatorial explosion, it is proposed that low-order combinations of inputs would be what is learned by each neuron. Evidence consistent with this suggestion that neurons are responding to combinations of a few variables represented at the preceding stage of cortical pro-

cessing is reviewed by Rolls and Deco (2002), and includes the finding that posterior inferior temporal cortex neurons respond to stimuli which may require two or more simple features to be present (Tanaka et al., 1990); and in the temporal cortical face processing areas the presence of several features in a face may be required by some neurons (such as eyes, hair, and mouth) in order to respond (Perrett et al., 1982; Yamane et al., 1988).

It is suggested that view-independent representations could be formed by the same type of computation, operating to combine a limited set of views of objects. The plausibility of providing view-independent recognition of objects by combining a set of different views of objects has been proposed by a number of investigators (Koenderink and Van Doorn, 1979; Poggio and Edelman, 1990; Logothetis et al., 1994; Ullman, 1996). Consistent with the suggestion that the view-independent representations are formed by combining view-dependent representations in the primate visual system, is the fact that in the temporal cortical areas, neurons with view-independent representations of faces are present in the same cortical areas as neurons with view-dependent representations (from which the view-independent neurons could receive inputs) (Hasselmo et al., 1989; Perrett et al., 1985; Booth and Rolls, 1998).

2. A feature hierarchy network model of invariant object recognition: VisNet

We now consider some of the computational issues that arise in feature hierarchy systems, with the help of a particular model, VisNet, which requires precise specification of the hypotheses, and at the same time enables them to be explored and tested numerically and quantitatively.

2.1. The architecture of VisNet

Fundamental elements of Rolls' (1992) theory for how cortical networks might implement invariant object recognition are described above and by Rolls and Deco (2002). They provide the basis for the design of VisNet, and can be summarized as:

- A series of competitive networks, organized in hierarchical layers, exhibiting mutual inhibition over a short range within each layer. These networks allow combinations of features or inputs occurring in a given spatial arrangement to be learned by neurons, ensuring that higher order spatial properties of the input stimuli are represented in the network.
- A convergent series of connections from a localized population of cells in preceding layers to each cell of the following layer, thus allowing the receptive field size of cells to increase through the visual processing areas or layers.
- A modified Hebb-like learning rule incorporating a temporal trace of each cell's previous activity, which, it is suggested, will enable the neurons to learn transform invariances.

2.1.1. The trace rule

The learning rule implemented in most VisNet simulations utilizes the spatio-temporal constraints placed upon the behaviour of 'real-world' objects to learn about natural object transformations. By presenting consistent sequences of transforming objects the cells in the network can learn to respond to the same object through all of its naturally transformed states, as described by Földiák (1991) and Rolls (1992). The learning rule incorporates a decaying trace of previous cell activity and is henceforth referred to simply as the 'trace' learning rule. The learning paradigm we describe here is intended in principle to enable learning of any of the transforms tolerated by inferior temporal cortex neurons (Rolls, 1992, 2000; Rolls and Deco, 2002).

To clarify the reasoning behind this point, consider the situation in which a single neuron is strongly activated by a stimulus forming part of a real-world object. The trace of this neuron's activation will then gradually decay over a time period in the order of 0.5 s. If, during this limited time window, the net is presented with a transformed version of the original stimulus then not only will the initially active afferent synapses modify onto the neuron, but so also will the synapses activated by the transformed version of this stimulus. In this way the cell will learn to respond to either appearance of the original stimulus. Making such associations works in practice because it is very likely that within short time periods different aspects of the same object will be being inspected. The cell will not, however, tend to make spurious links across stimuli that are part of different objects because of the unlikelihood in the real world of one object consistently following another.

Various biological bases for this temporal trace have been advanced:

- The persistent firing of neurons for as long as 100–400 ms observed after presentations of stimuli for 16 ms (Rolls and Tovee, 1994) could provide a time window within which to associate subsequent images. Maintained activity may potentially be implemented by recurrent connections between cortical areas (O'Reilly and Johnson, 1994; Rolls, 1994, 1995).
- The binding period of glutamate in the NMDA channels, which may last for 100 or more ms, may implement a trace rule by producing a narrow time window over which the *average* activity at each presynaptic site affects learning (Rolls, 1992; Rhodes, 1992; Földiák, 1992).
- Chemicals such as nitric oxide may be released during high neural activity and gradually decay in concentration over a short time window during which learning could be enhanced (Földiák, 1992; Montague et al., 1991).

The trace update rule used in the baseline simulations of VisNet (Wallis and Rolls, 1997) is equivalent to both

Földiák's used in the context of translation invariance and to the earlier rule of Sutton and Barto (1981) explored in the context of modelling the temporal properties of classical conditioning, and can be summarized as follows:

$$\delta w_j = \alpha \bar{y}^{\tau} x_j \tag{2}$$

where

$$\bar{y}^{\tau} = (1-\eta) y^{\tau} + \eta \bar{y}^{\tau-1} \tag{3}$$

and

x_j	jth input to the neuron
y	Output from the neuron
$\bar{y}^{\tau}$	trace value of the output of the neuron at time step τ
α	learning rate. Annealed between unity and zero
w_j	synaptic weight between jth input and the neuron
η	trace value. The optimal value varies with presentation sequence length

To bound the growth of each neuron's synaptic weight vector, $\mathbf{w}_i$ for the ith neuron, its length is explicitly normalized (a method similarly employed by Malsburg (1973) which is commonly used in competitive networks, see Rolls and Deco (2002)). An alternative, more biologically relevant implementation, using a local weight bounding operation which utilizes a form of heterosynaptic long-term depression (see Rolls and Deco, 2002), has in part been explored using a version of the Oja (1982) rule (see Wallis and Rolls, 1997).

2.1.2. The network implemented in VisNet

The network itself is designed as a series of hierarchical, convergent, competitive networks, in accordance with the hypothesis advanced above. The actual network consists of a series of four layers, constructed such that the convergence of information from the most disparate parts of the network's input layer can potentially influence firing in a single neuron in the final layer – see Fig. 1. This corresponds to the scheme described by many researchers (Van Essen et al., 1992; Rolls, 1992, for example) as present in the primate visual system – see Fig. 1. The forward connections to a cell in one layer are derived from a topologically related and confined region of the preceding layer. The choice of whether a connection between neurons in adjacent layers exists or not, is based upon a Gaussian distribution of connection probabilities which roll off radially from the focal point of connections for each neuron. (A minor extra constraint precludes the repeated connection of any pair of cells.) In particular, the forward connections to a cell in one layer come from a small region of the preceding layer (Wallis and Rolls, 1997; Rolls and Milward, 2000; Rolls and Deco, 2002). Fig. 1 shows the general convergent network architecture used. Localization and limitation of connectivity in the network is intended to mimic cortical connectivity, partially because of the clear retention of retinal topology through many regions of the ventral stream visual cortical areas.

2.1.3. Competition and lateral inhibition

In order to act as a competitive network some form of mutual inhibition is required within each layer, which should help to ensure that all stimuli presented are evenly represented by the neurons in each layer. This is implemented in VisNet by a form of lateral inhibition. The idea behind the lateral inhibition, apart from this being a property of cortical architecture in the brain, was to prevent too many neurons that received inputs from a similar part of the preceding layer responding to the same activity patterns. The purpose of the lateral inhibition was to ensure that different receiving neurons coded for different inputs. This is important in reducing redundancy (Rolls and Treves, 1998). The lateral inhibition is conceived as operating within a radius that was similar to that of the region within which a neuron received converging inputs from the preceding layer (because activity in one zone of topologically organized processing within a layer should not inhibit processing in another zone in the same layer, concerned perhaps with another part of the image).

The lateral inhibition and contrast enhancement just described are implemented in VisNet2 in two stages, as described in detail by Rolls and Milward (2000) and Rolls and Deco (2002).

2.1.4. The input to VisNet

VisNet is provided with a set of input filters which can be applied to an image to produce inputs to the network which correspond to those provided by simple cells in visual cortical area 1 (V1). The purpose of this is to enable within VisNet the more complicated response properties of cells between V1 and the inferior temporal cortex (IT) to be investigated, using as inputs natural stimuli such as those that could be applied to the retina of the real visual system. This is to facilitate comparisons between the activity of neurons in VisNet and those in the real visual system, to the same stimuli. In VisNet no attempt is made to train the response properties of simple cells, but instead we start with a defined series of filters to perform fixed feature extraction to a level equivalent to that of simple cells in V1, because we wish to simulate the more complicated response properties of cells between V1 and the inferior temporal cortex (IT). The elongated orientation-tuned input filters used accord with the general tuning profiles of simple cells in V1 (Hawken and Parker, 1987) and are oriented difference of Gaussians, or DOG filters. They are computed by weighting the difference of two Gaussians by a third orthogonal Gaussian (Wallis and Rolls, 1997; Rolls and Deco, 2002). Each individual filter is tuned to spatial frequency (0.0625–0.5 cycles pixels^{-1} over four octaves); orientation (0–135° in steps of 45°); and sign (± 1).

2.1.5. Measures for network performance

Measures of network performance based on information theory and similar to those used in the analysis of the firing of real neurons in the brain (Rolls et al., 1997a,b; Rolls and Deco, 2002) were introduced by Rolls and Milward (2000)

for VisNet2. A single cell information measure was introduced which is the maximum amount of information the cell has about any one stimulus/object independently of which transform (e.g. position on the retina) is shown. Because the competitive algorithm used in VisNet tends to produce local representations (in which single cells become tuned to one stimulus or object), this information measure can approach $\log_2 N_S$ bits, where N_S is the number of different stimuli. Rolls and Milward (2000) also introduced a multiple cell information measure, which has the advantage that it provides a measure of whether all stimuli are encoded by different neurons in the network. Again, a high value of this measure indicates good performance.

2.2. Initial experiments with VisNet

Having established a network model, Wallis and Rolls (1997) described four experiments in which the theory of how invariant representations could be formed was tested using a variety of stimuli undergoing a number of natural transformations. In each case the network produced neurons in the final layer whose responses were largely invariant across a transformation and highly discriminating between stimuli or sets of stimuli.

The first experiment was to learn translation invariant representations of 'T', 'L' and '+' stimuli, each of which consists of two bars, and which require the network to form representations in which the features are bound together in the correct relative spatial positions. The training consisted of sweeping a stimulus through a set of different training locations, doing the same with each of the other stimuli, for a number of training epochs. The network learned invariant representations, in that fourth layer neurons responded to one of the stimuli in all its different positions, and to none of the other stimuli. This developed gradually across the layers of the hierarchy, and only when the trace learning rule was used, not when a purely associative learning rule was used.

The second series of investigations described by Wallis and Rolls (1997) showed how the trace time constant η, which controls the exponential decay of the previous memory trace, can be set to optimize associations within a stimulus and to minimize those between stimuli (see further Wallis and Baddeley, 1997).

The third set of experiments described by Wallis and Rolls (1997) showed that the network can learn translation invariant representations of real biological stimuli, faces.

The fourth set of experiments described by Wallis and Rolls (1997) showed that the network can learn view invariant representations of real biological stimuli, faces.

2.3. Different forms of the trace learning rule, and their relation to error correction and temporal difference learning

The original trace learning rule used in the simulations of Wallis and Rolls (1997) and in our other investigations unless otherwise stated is shown in Eq. (2).

In the start of a series of investigations of different forms of the trace learning rule, Rolls and Milward (2000) demonstrated that VisNet's performance could be greatly enhanced with a modified Hebbian learning rule that incorporated a trace of activity from the preceding time steps, with no contribution from the activity being produced by the stimulus at the current time step. This rule took the form

$$\delta w_j = \alpha \bar{y}^{\tau-1} x_j^{\tau}. \tag{4}$$

The trace shown in Eq. (4) is in the postsynaptic term, and similar effects were found if the trace was in the presynaptic term, or in both the pre- and the postsynaptic term. The crucial difference from the earlier rule (see Eq. (2)) was that the trace should be calculated up to only the preceding time step, with no contribution to the trace from the firing on the current trial to the current stimulus. How might this be understood?

One way to understand this is to note that the trace rule is trying to set up the synaptic weight on trial τ based on whether the neuron, based on its previous history, is responding to that stimulus (in other positions). Use of the trace rule at $\tau - 1$ does this, that is it takes into account the firing of the neuron on previous trials, with no contribution from the firing being produced by the stimulus on the current trial. On the other hand, use of the trace at time τ in the update takes into account the current firing of the neuron to the stimulus in that particular position, which is not a good estimate of whether that neuron should be allocated to invariantly represent that stimulus. Effectively, using the trace at time τ introduces a Hebbian element into the update, which tends to build position encoded analyzers, rather than stimulus encoded analyzers.

Rule (4) corrects the weights using a postsynaptic trace obtained from the previous firing (produced by other transforms of the same stimulus), with no contribution to the trace from the current postsynaptic firing (produced by the current transform of the stimulus). Indeed, insofar as the current firing y^{τ} is not the same as $\bar{y}^{\tau-1}$, this difference can be thought of as an error. This leads to a conceptualization of using the difference between the current firing and the preceding trace as an error correction term. Rolls and Stringer (2001) developed a whole series of rules from this starting point, including rules that even performed a type of temporal difference learning. In terms of biological plausibility, Rolls and Stringer (2001) showed that all the rules are local learning rules, and in this sense are biologically plausible (see Rolls and Treves, 1998). (The rules are local in that the terms used to modify the synaptic weights are potentially available in the pre- and postsynaptic elements.) However, many of these rules involved subtraction of a preceding from a current state to produce an error term, and as this is perhaps more complicated for biological processes to implement, most of our other investigation have continued to use the very simple rule shown in Eq. (2), for biological plausibility.

2.4. The issue of feature binding, and a solution

In this section we address two key issues that arise in hierarchical layered network architectures, such as VisNet, other examples of which have been described and analyzed by Fukushima (1980), Ackley et al. (1985), and Rosenblatt (1961). One issue is whether the network can discriminate between stimuli that are composed of the same basic alphabet of features. The second issue is whether such network architectures can find solutions to the spatial binding problem. These issues are addressed in the next two paragraphs and by Elliffe et al. (2002).

2.4.1. Objects and parts of objects

The first issue investigated is whether a hierarchical layered network architecture of the type exemplified by VisNet can discriminate stimuli that are composed of a limited set of features and where the different stimuli include cases where the feature sets are subsets and supersets of those in the other stimuli. To address this issue Elliffe et al. (2002) trained VisNet with stimuli that are composed from a set of four features which are designed so that each feature is spatially separate from the other features, and no unique combination of firing caused for example by overlap of horizontal and vertical filter outputs in the input representation distinguishes any one stimulus from the others. They showed that VisNet can indeed learn correct invariant representations of stimuli which do consist of feature sets where individual features do not overlap spatially with each other and where the stimuli can be composed of sets of features which are supersets or subsets of those in other stimuli. VisNet solves this problem because as a competitive net, neurons at one layer can learn to allocate different neurons to individual features and to combinations of those features, and uses normalization of the synaptic weight vectors and the input stimulus vectors to achieve this (Rolls and Deco, 2002).

2.4.2. Spatial binding of features

The second issue is the spatial binding problem in architectures such as VisNet. This computational problem that needs to be addressed in hierarchical networks such as the primate visual system and VisNet is how representations of features can be (e.g. translation) invariant, yet can specify stimuli or objects in which the features must be specified in the correct spatial arrangement. This is the feature binding problem, discussed for example by Malsburg (1990), and arising in the context of hierarchical layered systems (Ackley et al., 1985; Fukushima, 1980; Rosenblatt, 1961). The issue is whether or not features are bound together with each feature in the correct spatial position relative to the other features, yet with the object or object part represented by the combination of features at the same time translation invariant.

2.4.2.1. Temporal synchronization. Von der Malsburg suggested that one potential solution is the addition of a temporal dimension to the neuronal response, so that features that should be bound together would be linked by temporal binding. There has been considerable neurophysiological investigation of this possibility (Singer, 1999; Singer et al., 1990; Abeles, 1991; Hummel and Biederman, 1992; Singer and Gray, 1995). We note that one problem with this approach is that temporal binding might enable features 1, 2 and 3, which might define one stimulus to be bound together and kept separate from for example another stimulus consisting of features 2, 3 and 4, but would require a further temporal binding (leading in the end potentially to a combinatorial explosion) to indicate the relative spatial positions of the 1, 2 and 3 in the 123 stimulus, so that it can be discriminated from e.g. 312. A second problem with this approach is that, when the stimulus-dependent temporal synchronization has been rigourously tested with information theoretic approaches, it has so far been found that most of the information available is in the number of spikes, with rather little, less than 5% of the total information, in stimulus-dependent synchronization (Aggelopoulos et al., 2005; Franco et al., 2004; Rolls et al., 2004). For example, Aggelopoulos et al. (2005) showed that when macaques used object-based attention to search for one of two objects to touch in a complex natural scene between 99% and 94% of the information was present in the firing rates of inferior temporal cortex neurons, and less that 5% in any stimulus-dependent synchrony that was present between the simultaneously recorded inferior temporal cortex neurons. The implication of these results is that any stimulus-dependent synchrony that is present is not quantitatively important as measured by information theoretic analyses under natural scene conditions when feature binding, segmentation of objects from the background, and attention are required. This has been found for the inferior temporal cortex, a brain region where features are put together to form representations of objects (Rolls and Deco, 2002), and where attention has strong effects, at least in scenes with blank backgrounds (Rolls et al., 2003). It would of course also be of interest to test the same hypothesis in earlier visual areas, such as V4, with quantitative, information theoretic, techniques. In connection with rate codes, it should be noted that a rate code implies using the number of spikes that arrive in a given time, and that this time can be very short, as little as 20–50 ms, for very useful amounts of information to be made available from a population of neurons (Tovee et al., 1993; Rolls and Tovee, 1994; Rolls et al., 1994, 1999, 2006; Tovee and Rolls, 1995; Rolls, 2003).

2.4.2.2. Sigma–Pi neurons. Another approach to a binding mechanism is to group spatial features based on local mechanisms that might operate for closely adjacent synapses on a dendrite (Finkel and Edelman, 1987; Mel et al., 1998). This might implement a Sigma–Pi computation in which a product is formed between a set of local synapses, and the neuron would sum different such products (Rolls and Deco, 2002). If each local product corre-

sponded to a feature combination at one location in the world, then the neuron might respond to the feature combination in any one of the different locations, this producing an invariant representation. A problem for such architectures is how to force one particular neuron to respond to the same feature combination invariantly with respect to all the ways in which that feature combination might occur in a scene.

2.4.2.3. Binding of features and their relative spatial position by feature combination neurons. The approach to the spatial binding problem that is proposed for VisNet is that individual neurons at an early stage of processing are set up (by learning) to respond to low-order combinations of input features occurring in a given relative spatial arrangement and position on the retina (Rolls, 1992, 1994, 1995; Wallis and Rolls, 1997; Rolls and Treves, 1998; Rolls and Deco, 2002) (cf. Feldman, 1985). (By low-order combinations of input features we mean combinations of a few input features. By forming neurons that respond to combinations of a few features in the correct spatial arrangement the advantages of the scheme for syntactic binding are obtained, yet without the combinatorial explosion that would result if the feature combination neurons responded to combinations of many input features so producing potentially very specifically tuned neurons which very rarely responded.) Then invariant representations are developed in the next layer from these feature combination neurons which already contain evidence on the local spatial arrangement of features. Finally, in later layers, only one stimulus would be specified by the particular set of low-order feature combination neurons present, even though each feature combination neuron would itself be somewhat invariant. Elliffe et al. (2002) showed that VisNet can solve this spatial binding problem in the way proposed. They trained the first two layers of VisNet with feature pair combinations, forming representations of feature pairs with some translation invariance in layer 2. Then they used feature triples as input stimuli, allowed no more learning in layers 1 and 2, and then investigated whether layers 3 and 4 could be trained to produce invariant representations of the triples where the triples could only be distinguished if the local spatial arrangement of the features within the triple had effectively to be encoded in order to distinguish the different triples.

Three conclusions follow from these results. First, a hierarchical network which seeks to produce invariant representations in the way used by VisNet can solve the feature binding problem. In particular, when feature pairs in layer 2 with some translation invariance are used as the input to later layers, these later layers can nevertheless build invariant representations of objects where all the individual features in the stimulus must occur in the correct spatial position relative to each other. This is possible because the feature combination neurons formed in the first layer (which could be trained just with a Hebb rule) do respond to combinations of input features in the correct spatial configuration, partly because of the limited size of their receptive fields. The second conclusion is that even though early layers can in this case only respond to small feature subsets, these provide, with no further training of layers 1 and 2, an adequate basis for learning to discriminate in layers 3 and 4 stimuli consisting of combinations of larger numbers of features. Third, because some invariance for low-order feature combinations had been built into the first two layers of VisNet, training with new objects composed of new combinations of those feature pairs could generalize after training in a few locations to other locations. (This occurs for example because if the new object was then shown in a new location, the same set of layer 3 neurons would be active because they respond with spatial invariance to feature combinations, and given that the layer 3–4 connections had already been set up by the new object, the correct layer 4 neurons would be activated by the new object in its new untrained location, and without any further training.) This is an important point, for it shows that after prior training on objects, feature hierarchy networks need not be trained on every possible transform of a new object, but only on a few transforms.

2.5. Operation in a cluttered environment

In natural environments, objects may not only appear against cluttered (natural) backgrounds, but also the object may be partially occluded. Stringer and Rolls (2000) considered factors that influence the performance of feature hierarchy networks in cluttered backgrounds, and showed that, for previously trained objects, performance was little affected by background clutter or by partial occlusion of the object. One of the reasons for this is that after training the network operates partly as an associative look-up system, and thus previously trained objects tend to dominate the competitive interactions in each layer, and also generalization allows recovery from partial occlusion.

Training a feature hierarchy network in a cluttered natural scene is likely to be more complicated, but may be facilitated by the following factors. First, the receptive fields of inferior temporal cortex neurons shrink from in the order of 70° in diameter when only one object is present in a blank scene to much smaller values of as little as 5–10° close to the fovea in complex natural scenes (Rolls et al., 2003). The proposed mechanism for this is that if there is an object at the fovea, this object, because of the high cortical magnification factor at the fovea, dominates the activity of neurons in the inferior temporal cortex by competitive interactions (Trappenberg et al., 2002; Deco and Rolls, 2004) (see Section 2.6). This allows primarily the object at the fovea to be represented in the inferior temporal cortex, and, it is proposed, for learning to be about this object, and not about the other objects in a whole scene. Second, top-down spatial attention (Deco and Rolls, 2004, 2005a) could bias the competition towards a region of visual space where the object to be learned is located.

2.6. Attention in natural scenes – a computational account

In this section, we consider how attention operates in complex natural scenes, and in particular describe how the inferior temporal visual cortex operates to enable the selection of an object in a complex natural scene (Rolls and Deco, 2006). The inferior temporal visual cortex contains distributed and invariant representations of objects and faces (Rolls, 2000, 2006; Rolls and Deco, 2002; Booth and Rolls, 1998; Rolls et al., 1997a; Rolls and Tovee, 1995; Tovee et al., 1994; Hasselmo et al., 1989; Rolls and Baylis, 1986; Franco et al., 2007; Rolls, 2007).

To investigate how attention operates in complex natural scenes, and how information is passed from the inferior temporal cortex (IT) to other brain regions to enable stimuli to be selected from natural scenes for action, Rolls et al. (2003) analyzed the responses of inferior temporal cortex neurons to stimuli presented in complex natural backgrounds. The monkey had to search for two objects on a screen, and a touch of one object was rewarded with juice, and of another object was punished with saline. Neuronal responses to the effective stimuli for the neurons were compared when the objects were presented in the natural scene or on a plain background. It was found that the overall response of the neuron to objects was hardly reduced when they were presented in natural scenes, and the selectivity of the neurons remained. However, the main finding was that the magnitudes of the responses of the neurons typically became much less in the real scene the further the monkey fixated in the scene away from the object, that is, the receptive fields decreased from approximately 70° with a plain background to as little as a few deg in a complex scene (Rolls et al., 2003).

It is proposed that this reduced translation invariance in natural scenes helps an unambiguous representation of an object which may be the target for action to be passed to the brain regions which receive from the primate inferior temporal visual cortex. It helps with the binding problem, by reducing in natural scenes the effective receptive field of at least some inferior temporal cortex neurons to approximately the size of an object in the scene.

It is also found that in natural scenes, the effect of object-based attention on the response properties of inferior temporal cortex neurons is relatively small compared to the blank background condition (Rolls et al., 2003).

Trappenberg et al. (2002) have suggested what underlying mechanisms could account for these findings, and simulated a model to test the ideas. The model utilizes an attractor network representing the inferior temporal visual cortex (implemented by the recurrent excitatory connections between inferior temporal cortex neurons), and a neural input layer with several retinotopically organized modules representing the visual scene in an earlier visual cortical area such as V4 (see Fig. 2). The attractor network aspect of the model produces the property that receptive fields of IT neurons can be large in blank scenes by enabling a weak input in the periphery of the visual field

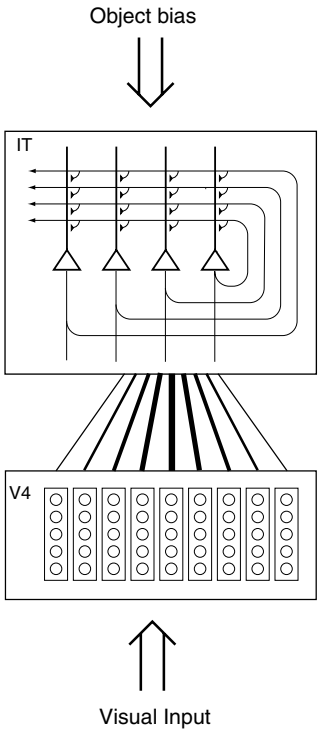

Fig. 2. The architecture of the inferior temporal cortex (IT) model of Trappenberg et al. (2002) operating as an attractor network with inputs from the fovea given preferential weighting by the greater magnification factor of the fovea. The model also has a top-down object-selective bias input. The model was used to analyze how object vision and recognition operate in complex natural scenes.

to act as a retrieval cue for the object attractor. On the other hand, when the object is shown in a complex background, the object closest to the fovea tends to act as the retrieval cue for the attractor, because the fovea is given increased weight in activating the IT module because the magnitude of the input activity from objects at the fovea is greatest due to the cortical higher magnification factor of the fovea incorporated into the model. (The cortical magnification factor can be expressed as the number of mm of cortex representing 1 degree of visual field. The cortical magnification factor decreases rapidly with increasing eccentricity from the fovea (Rolls and Cowey, 1970; Cowey and Rolls, 1975).) This results in smaller receptive fields of IT neurons in complex scenes, because the object tends to need to be close to the fovea to trigger the attractor into the state representing that object. (In other words, if the object is far from the fovea, then it will not trigger neurons in IT which represent it, because neurons in IT are preferentially being activated by another object at the fovea.) This may be described as an attractor model in which the competition for which attractor state is retrieved is weighted towards objects at the fovea.

Attentional top-down object-based inputs can bias the competition implemented in this attractor model, but have relatively minor effects (in for example increasing receptive field size) when they are applied in a complex natural scene,

as then as usual the stronger forward inputs dominate the states reached. In this network, the recurrent collateral connections may be thought of as implementing constraints between the different inputs present, to help arrive at firing in the network which best meets the constraints. In this scenario, the preferential weighting of objects close to the fovea because of the increased magnification factor at the fovea is a useful principle in enabling the system to provide useful output. The attentional object biasing effect is much more marked in a blank scene, or a scene with only two objects present at similar distances from the fovea, which are conditions in which attentional effects have frequently been examined. The results of the investigation (Trappenberg et al., 2002) thus suggest that attention may be a much more limited phenomenon in complex, natural, scenes than in reduced displays with one or two objects present. The results also suggest that the alternative principle, of providing strong weight to whatever is close to the fovea, is an important principle governing the operation of the inferior temporal visual cortex, and in general of the output of the ventral visual system in natural environments. This principle of operation is very important in interfacing the visual system to action systems, because the effective stimulus in making inferior temporal cortex neurons fire is in natural scenes usually on or close to the fovea. This means that the spatial coordinates of where the object is in the scene do not have to be represented in the inferior temporal visual cortex, nor passed from it to the action selection system, as the latter can assume that the object making IT neurons fire is close to the fovea in natural scenes (see Rolls and Deco, 2002; Rolls et al., 2003).

There may of course be in addition a mechanism for object selection that takes into account the locus of covert attention when actions are made to locations not being looked at. However, the simulations described in this section suggest that in any case covert attention is likely to be a much less significant influence on visual processing in natural scenes than in reduced scenes with one or two objects present.

Given these points, one might question why inferior temporal cortex neurons can have such large receptive fields, which show translation invariance (Rolls, 2000; Rolls et al., 2003). At least part of the answer to this may be that inferior temporal cortex neurons must have the capability to be large if they are to deal with large objects (Rolls and Deco, 2002). A V1 neuron, with its small receptive field, simply could not receive input from all the features necessary to define an object. On the other hand, inferior temporal cortex neurons may be able to adjust their size to approximately the size of objects, using in part the interactive attentional effects of bottom-up and top-down effects described elsewhere in this paper.

The implementation of the simulations is described by Trappenberg et al. (2002), and some of the results obtained with the architecture shown Fig. 2, follow. In one simulation only one object was present in the visual scene in a plain background at different eccentricities from the fovea. As shown in Fig. 3a by the line labelled 'simple background', the receptive fields of the neurons were very large. The value of the object bias $k^{\mathrm{IT_{BIAS}}}$ was set to 0 in these simulations. Good object retrieval (indicated by large correlations) was found even when the object was far from the fovea, indicating large IT receptive fields with a blank background. The reason that any drop is seen in performance as a function of eccentricity is because some noise was present in the recall process. This demonstrates that the attractor dynamics can support translation invariant object recognition even though the translation invariant weight vectors between V4 and IT are explicitly modulated by the modulation factor $k^{\mathrm{IT-V4}}$ derived from the cortical magnification factor.

In a second simulation individual objects were placed at all possible locations in a natural and cluttered visual scene. The resulting correlations between the target pattern and the asymptotic IT state are shown in Fig. 3a with the

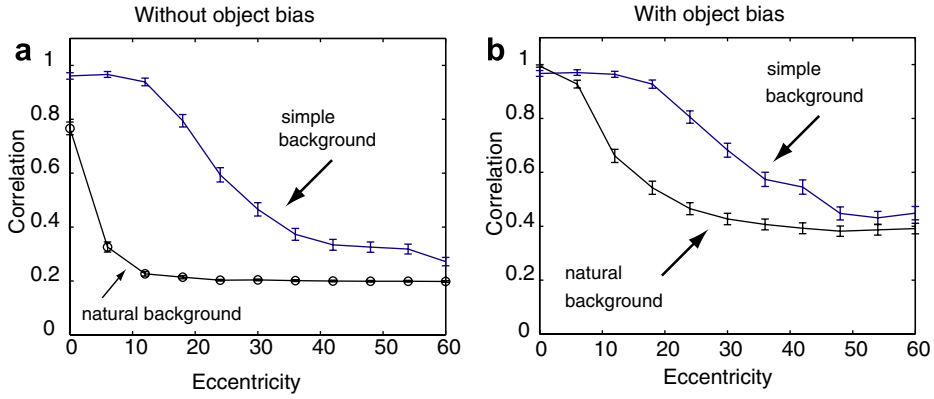

Fig. 3. Correlations as measured by the normalized dot product between the object vector used to train IT and the state of the IT network after settling into a stable state with a single object in the visual scene (blank background) or with other trained objects at all possible locations in the visual scene (natural background). There is no object bias included in the results shown in graph (a), whereas an object bias is included in the results shown in (b) with $k^{\mathrm{IT_{BIAS}}} = 0.7$ in the experiments with a natural background and $k^{\mathrm{IT_{BIAS}}} = 0.1$ in the experiments with a blank background.

line labelled 'natural background'. Many objects in the visual scene are now competing for recognition by the attractor network, and the objects around the foveal position are enhanced through the modulation factor derived from the cortical magnification factor. This results in a much smaller size of the receptive field of IT neurons when measured with objects in natural backgrounds.

In addition to this major effect of the background on the size of the receptive field, which parallels and we suggest may account for the physiological findings outlined above, there is also a dependence of the size of the receptive fields on the level of object bias provided to the IT network. Examples are shown in Fig. 3b where an object bias was used. The object bias biases the IT network towards the expected object with a strength determined by the value of $k^{\mathrm{IT_{BIAS}}}$, and has the effect of increasing the size of the receptive fields in both blank and natural backgrounds (see Fig. 3b compared to a). This models the effect found neurophysiologically (Rolls et al., 2003).

Some of the conclusions are as follows. When single objects are shown in a scene with a blank background, the attractor network helps neurons to respond to an object with large eccentricities of this object relative to the fovea. When the object is presented in a natural scene, other neurons in the inferior temporal cortex become activated by the other effective stimuli present in the visual field, and these forward inputs decrease the response of the network to the target stimulus by a competitive process. The results found fit well with the neurophysiological data, in that IT operates with almost complete translation invariance when there is only one object in the scene, and reduces the receptive field size of its neurons when the object is presented in a cluttered environment. The model described here provides an explanation of the responses of real IT neurons in natural scenes.

In natural scenes, the model is able to account for the neurophysiological data that the IT neuronal responses are larger when the object is close to the fovea, by virtue of fact that objects close to the fovea are weighted by the cortical magnification factor related modulation $k^{\mathrm{IT-V4}}$.

The model accounts for the larger receptive field sizes from the fovea of IT neurons in natural backgrounds if the target is the object being selected compared to when it is not selected (Rolls et al., 2003). The model accounts for this by an effect of top-down bias which simply biases the neurons towards particular objects compensating for their decreasing inputs produced by the decreasing magnification factor modulation with increasing distance from the fovea. Such object-based attention signals could originate in the prefrontal cortex and could provide the object bias for the inferotemporal cortex (Renart et al., 2000).

Important properties of the architecture for obtaining the results just described are the high magnification factor at the fovea and the competition between the effects of different inputs, implemented in the above simulation by the competition inherent in an attractor network.

We have also been able to obtain similar results in a hierarchical feedforward network where each layer operates as a competitive network (Deco and Rolls, 2004). This network thus captures many of the properties of our hierarchical model of invariant object recognition (Rolls, 1992; Wallis and Rolls, 1997; Rolls and Milward, 2000; Stringer and Rolls, 2000, 2002; Rolls and Stringer, 2001, 2006; Elliffe et al., 2002; Rolls and Deco, 2002; Stringer et al., 2006), but incorporates in addition a foveal magnification factor and top-down projections with a dorsal visual stream so that attentional effects can be studied, as shown in Fig. 4.

Deco and Rolls (2004) trained the network shown in Fig. 4 with two objects, and used the trace learning rule (Wallis and Rolls, 1997; Rolls and Milward, 2000) in order to achieve translation invariance. In a first experiment we placed only one object on the retina at different distances from the fovea (i.e. different eccentricities relative to the fovea). This corresponds to the blank background condition. In a second experiment, we also placed the object at different eccentricities relative to the fovea, but on a cluttered natural background. Larger receptive fields were found with the blank as compared to the cluttered natural background.

Deco and Rolls (2004) also studied the influence of object-based attentional top-down bias on the effective size of the receptive field of an inferior temporal cortex neuron for the case of an object in a blank or a cluttered background. To do this, they repeated the two simulations but now considered a non-zero top-down bias coming from prefrontal area 46v and impinging on the inferior temporal cortex neuron specific for the object tested. When no attentional object bias was introduced, a shrinkage of the receptive field size was observed in the complex vs the blank background. When attentional object bias was introduced, the shrinkage of the receptive field due to the complex background was somewhat reduced. This is consistent with the neurophysiological results (Rolls et al., 2003). In the framework of the model (Deco and Rolls, 2004), the reduction of the shrinkage of the receptive field is due to the biasing of the competition in the inferior temporal cortex layer in favour of the specific IT neuron tested, so that it shows more translation invariance (i.e. a slightly larger receptive field). The increase of the receptive field size of an IT neuron, although small, produced by the external top-down attentional bias offers a mechanism for facilitation of the search for specific objects in complex natural scenes.

2.7. The representation of multiple objects in a scene

When objects have distributed representations, there is a problem of how multiple objects (whether the same or different) can be represented in a scene, because the distributed representations overlap, and it may not be possible to determine whether one has an amalgam of several objects, or a new object (Mozer, 1991), or multiple

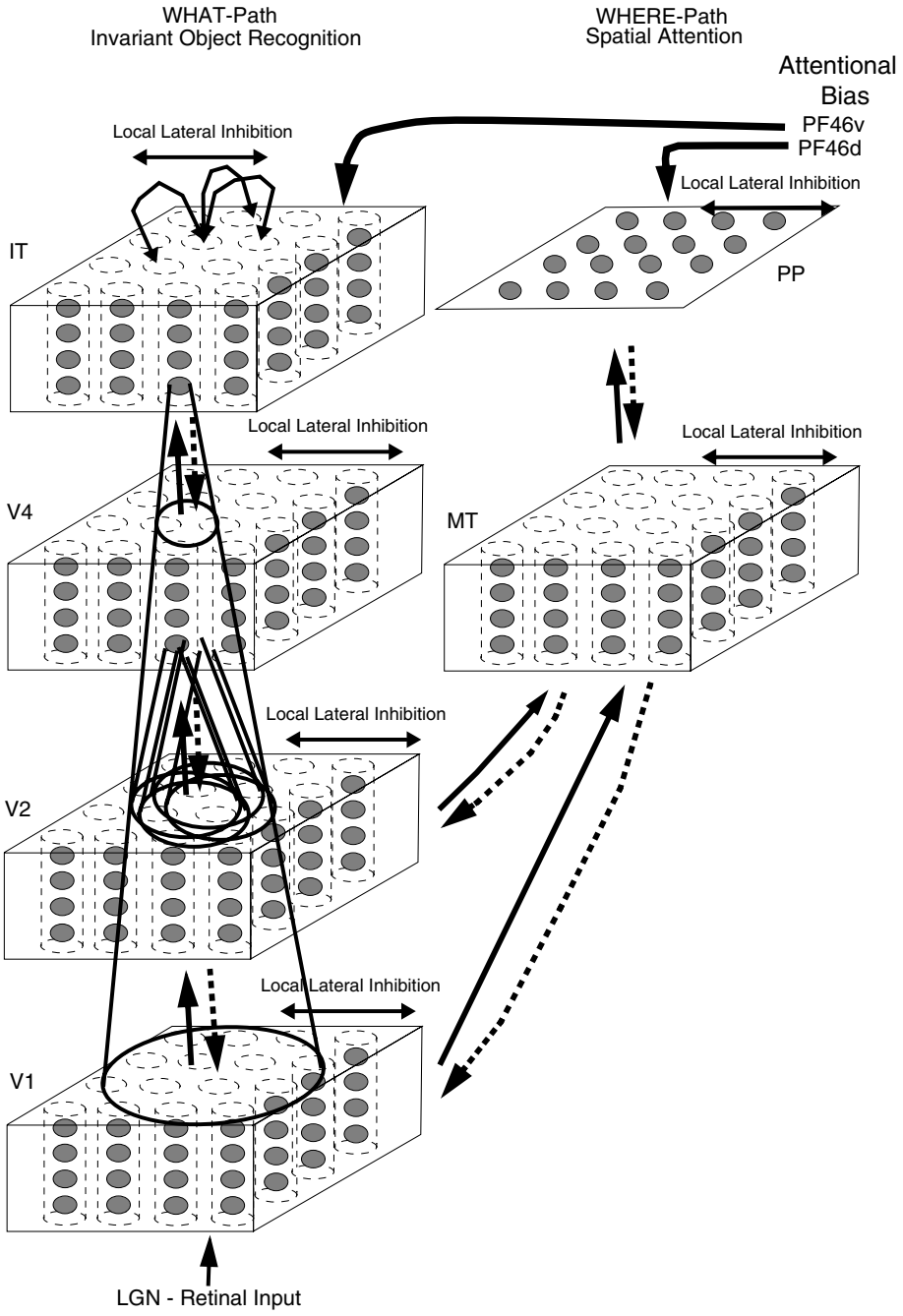

Fig. 4. Cortical architecture for hierarchical and attention-based visual perception (Deco and Rolls, 2004). The system is essentially composed of five modules structured such that they resemble the two known main visual paths of the mammalian visual cortex. Information from the retino-geniculo-striate pathway enters the visual cortex through area V1 in the occipital lobe and proceeds into two processing streams. The occipital-temporal stream leads ventrally through V2–V4 and IT (inferior temporal visual cortex), and is mainly concerned with object recognition. The occipito-parietal stream leads dorsally into PP (posterior parietal complex), and is responsible for maintaining a spatial map of an object's location. The solid lines with arrows between levels show the forward connections, and the dashed lines the top-down backprojections. Short term memory systems in the prefrontal cortex (PF46) apply top-down attentional bias to the object or spatial processing streams.

instances of the same object, let alone the relative spatial positions of the objects in a scene. Yet humans can determine the relative spatial locations of objects in a scene even in short presentation times without eye movements (Biederman, 1972) (and this has been held to involve some spotlight of attention). Aggelopoulos and Rolls (2005) analyzed this issue by recording from single inferior temporal cortex neurons with five objects simultaneously present in the receptive field. They found that although all the neurons responded to their effective stimulus when it was at the fovea, some could also respond to their effective stimulus when it was in some but not other parafoveal positions 10° from the fovea. An example of such a neuron is shown in Fig. 5. The asymmetry is much more evident in a scene

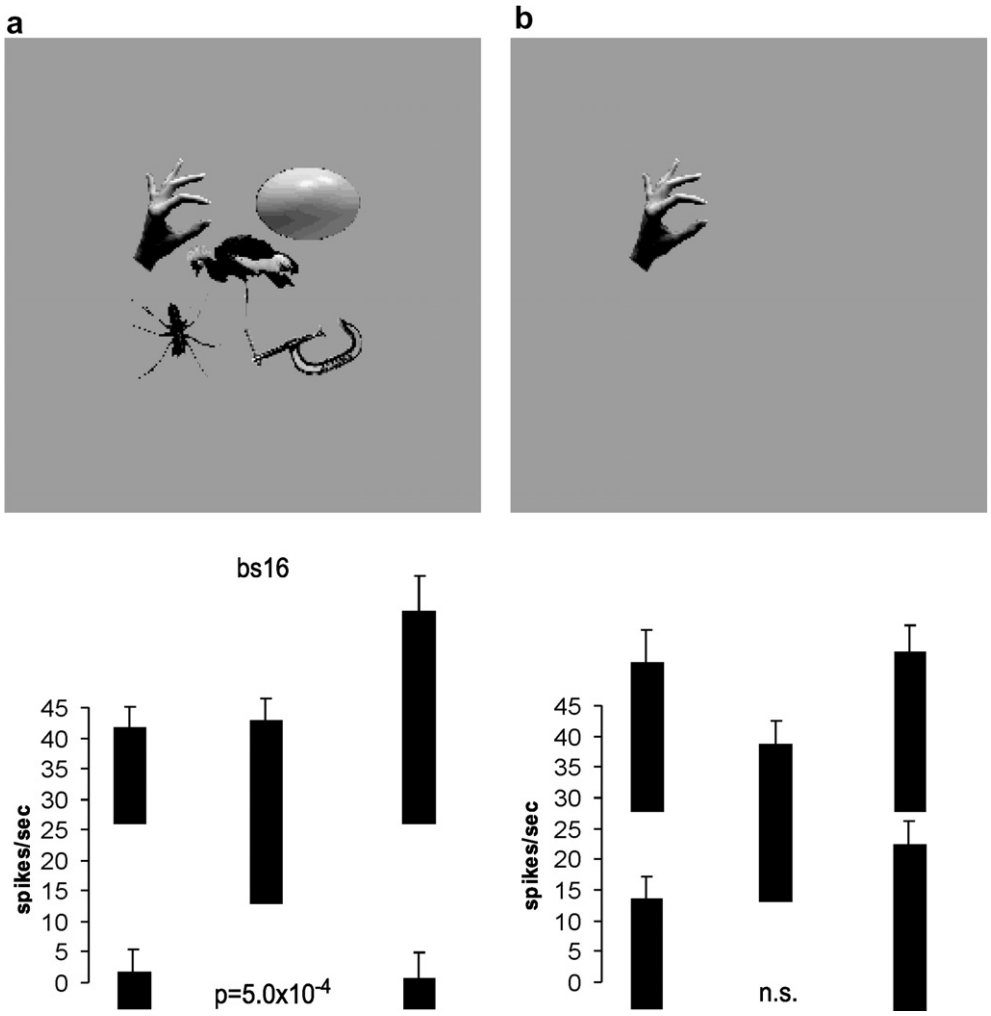

Fig. 5. (a) The responses (firing rate with the spontaneous rate subtracted, means ± sem) of an inferior temporal cortex neuron when tested with 5 stimuli simultaneously present in the close (10°) configuration with the parafoveal stimuli located 10° from the fovea. (b) The responses of the same neuron when only the effective stimulus was presented in each position. The firing rate for each position is that when the effective stimulus for the neuron was in that position. The p value is that from the ANOVA calculated over the four parafoveal positions.

with 5 images present (Fig. 5a) than when only one image is shown on an otherwise blank screen (Fig. 5b). Competition between different stimuli in the receptive field thus reveals the asymmetry in the receptive field of inferior temporal visual cortex neurons.

The asymmetry provides a way of encoding the position of multiple objects in a scene. Depending on which asymmetric neurons are firing, the population of neurons provides information to the next processing stage not only about which image is present at or close to the fovea, but where it is with respect to the fovea. This information is provided by neurons that have firing rates that reflect the relevant information, and stimulus-dependent synchrony is not necessary. Top-down attentional biasing input could thus, by biasing the appropriate neurons, facilitate bottom-up information about objects without any need to alter the time relations between the firing of different neurons. The exact position of the object with respect to the fovea, and effectively thus its spatial position relative to other objects in the scene, would then be made evident by the subset of asymmetric neurons firing.

This is thus the solution that these experiments indicate is used for the representation of multiple objects in a scene, an issue that has previously been difficult to account for in neural systems with distributed representations (Mozer, 1991) and for which 'attention' has been a proposed solution.

2.8. Learning 3D transforms

There is evidence that some neurons in the inferior temporal cortex may show two different types of 3D invariance. First, Booth and Rolls (1998) showed that some inferior temporal cortex neurons can respond to different generic views of familiar 3D objects. Second, some neurons do generalize across quantitative changes in the values of 3D shape descriptors while faces (Hasselmo et al., 1989) and objects (Tanaka, 1996; Logothetis et al., 1995) are rotated within-

generic views. Indeed, Logothetis et al. (1995) showed that a few inferior temporal cortex neurons can generalize to novel (untrained) values of the quantitative shape descriptors typical of within-generic-view object rotation. In addition to the qualitative shape descriptor changes that occur catastrophically between different generic views of an object, and the quantitative changes of 3D shape descriptors that occur within a generic view, there is a third type of transform that must be learned for correct invariant recognition of 3D objects as they rotate in depth. This third type of transform is that which occurs to the surface features on a 3D object as it transforms in depth.

Stringer and Rolls (2002) showed that trace learning can in the VisNet architecture solve the problem of in-depth rotation invariant object recognition by developing representations of the transforms which features undergo when they are on the surfaces of 3D objects. Moreover, they showed that having learned how features on 3D objects transform as the object is rotated in depth, the network can correctly recognize novel 3D variations within a generic view of an object which is composed of previously learned feature combinations.

The process investigated by Stringer and Rolls (2002) will only allow invariant object recognition over moderate 3D object rotations, since rotating an object through a large angle may lead to a catastrophic change in the appearance of the object that requires the new qualitative 3D shape descriptors to be associated with those of the former view. In that case, invariant object recognition must rely on the first process referred to at the start of this Section 2.8 in order to associate together the different generic views of an object to produce view invariant object identification. For that process, association of a few cardinal or generic views is likely to be sufficient (Koenderink, 1990). The process described in this section of learning how surface features transform is likely to make a major contribution to the within-generic view transform invariance of object identification and recognition.

2.9. A biased competition model of object and spatial attention

Visual attention exerts top-down influences on the processing of sensory information in the visual cortex, and therefore is intrinsically associated with interactions between cortical areas. Thus, elucidating the neural basis of visual attention is an excellent paradigm for understanding the basic mechanisms of intercortical neurodynamics. Recent cognitive neuroscience developments allow a more direct study of the neural mechanisms underlying attention in humans and primates. In particular, the work of Chelazzi et al. (1993) has led to a promising account of attention termed the '*biased competition hypothesis*' (see also Duncan, 1996; Moran and Desimone, 1985; Reynolds and Desimone, 1999). According to this hypothesis, attentional selection operates in parallel by biasing an underlying competitive interaction between multiple stimuli in the visual field toward one stimulus or another, so that behaviorally relevant stimuli are processed in the cortex while irrelevant stimuli are filtered out. Thus, attending to a stimulus at a particular location or with a particular feature biases the underlying neural competition in a certain brain area in favour of neurons that respond to the location, or the features, of the attended stimulus.

Neurodynamical models for biased competition have been proposed and successfully applied in the context of attention and working memory. In the context of attention, Usher and Niebur (1996) introduced an early model of biased competition. Deco and Zihl (2001) extended Usher and Niebur's model to simulate the psychophysics of visual attention by visual search experiments in humans. Their neurodynamical formulation is a large-scale hierarchical model of the visual cortex whose global dynamics is based on biased competition mechanisms at the neural level. Attention then appears as an emergent effect related to the dynamical evolution of the whole network. This large-scale formulation has been able to simulate and explain in a unifying framework (Deco and Rolls, 2005a) visual attention in a variety of tasks including object and spatial search with a simplified form of the architecture shown in Fig. 4, and at different cognitive neuroscience experimental measurement levels, namely: single-cells (Rolls and Deco, 2002; Deco and Lee, 2002; Deco and Rolls, 2005a), fMRI (Corchs and Deco, 2002, 2004), psychophysics (Deco et al., 2002; Deco and Rolls, 2005a), and neuropsychology (Deco and Rolls, 2002).

For example Deco and Rolls (2005b) extended previous concepts of the role of biassed competition in attention (Duncan, 1996; Desimone and Duncan, 1995; Usher and Niebur, 1996) by providing the first analysis at the integrate-and-fire neuronal level, which allows the neuronal non-linearities in the system to be explicitly modelled, in order to investigate realistically the processes that underlie the apparent gain modulation effect of top-down attentional control. In the integrate-and-fire model, the competition is implemented realistically by the effects of the excitatory neurons on the inhibitory neurons, and their return inhibitory synaptic connections. That was also the first integrate-and-fire analysis of top-down attentional influences in vision that explicitly models the interaction of several different brain areas. Part of the originality of the model is that in the form in which it can account for attentional effects in V2 and V4 in the paradigms of Reynolds et al. (1999) in the context of biased competition, the model with the same parameters effectively makes predictions which show that the 'contrast gain' effects in MT of Martinez-Trujillo and Treue (2002) can be accounted for by the same model. These detailed and quantitative analyses of neuronal dynamical systems are an important step towards understanding the operation of complex processes such as top-down attention, which necessarily involve the interaction of several brain areas. They are being extended to provide neurally plausible models of decision-making (Deco and Rolls, 2003, 2005c, 2006; Rolls, 2005).

In the context of working memory, further developments (Deco et al., 2004; Szabo et al., 2004) managed to model in a unifying form attentional and memory effects in the prefrontal cortex, integrating single-cell and fMRI data, and different paradigms in the framework of biased competition (Deco and Rolls, 2005a).

2.10. Invariant global motion in the dorsal visual system

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 Fig. 6a rotating clockwise independently of its position on the retina. This occurs even though the local motion for the wheels in the different positions may be opposite. 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 (in the range 1–2° at the fovea), and can therefore not detect global motion, and this is part of the aperture problem (Wurtz and Kandel, 2000). Neurons in MT, which receives inputs from V1 and V2, have larger receptive fields (e.g. 5° 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° to the overall motion (Movshon et al., 1985; Newsome et al., 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 et al., 1994; Geesaman and Andersen, 1996).

In a unifying hypothesis with the design of the ventral cortical visual system, Rolls and Stringer (2006) proposed 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 showed 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 vs 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.

2.11. Learning invariant representations using spatial continuity: continuous transformation learning

The temporal continuity typical of objects has been used in an associative learning rule with a short term memory trace to help build invariant object representations in the networks described previously in this paper. Stringer et al. (2006) showed that spatial continuity can also provide a basis for helping a system to self-organize invariant representations. They introduced a new learning paradigm 'continuous transformation (CT) learning' which operates by mapping spatially similar input patterns to the same postsynaptic neurons in a competitive learning system. As the inputs move through the space of possible continuous transforms (e.g. translation, rotation, etc.), the active synapses are modified onto the set of postsynaptic neurons. Because other transforms of the same stimulus overlap with previously learned exemplars, a common set of postsynaptic neurons is activated by the new transforms, and learning of the new active inputs onto the same postsynaptic neurons is facilitated.

The concept is illustrated in Fig. 7. During the presentation of a visual image at one position on the retina that activates neurons in layer 1, a small winning set of neurons in layer 2 will modify (through associative learning) their afferent connections from layer 1 to respond well to that image in that location. When the same image appears later at nearby locations, so that there is spatial continuity, the same neurons in layer 2 will be activated because some of the active afferents are the same as when the image was in the first position. The key point is that if these afferent connections have been strengthened sufficiently while the image is in the first location, then these connections will be able to continue to activate the same neurons in layer 2 when the image appears in overlapping nearby locations. Thus the same neurons in the output layer have learned to respond to inputs that have similar vector elements in common.

As can be seen in Fig. 7, the process can be continued for subsequent shifts, provided that a sufficient proportion of input cells stay active between individual shifts. This whole process is repeated throughout the network, both horizontally as the image moves on the retina, and hierarchically up through the network. Over a series of stages, transform invariant (e.g. location invariant) representations of images are successfully learned, allowing the network to perform invariant object recognition. A similar CT learning process may operate for other kinds of transformation, such as change in view or size.

Stringer et al. (2006) demonstrated that VisNet can be trained with continuous transform learning to form view invariant representations. They showed that CT learning requires the training transforms to be relatively close together spatially so that spatial continuity is present in the training set; and that the order of stimulus presentation is not crucial, with even interleaving with other objects possible during training, because it is spatial continuity rather the temporal continuity that drives the self-organizing learning with the purely associative synaptic modification rule.

Perry et al. (in press) extended these simulations with VisNet of view invariant learning using CT to more

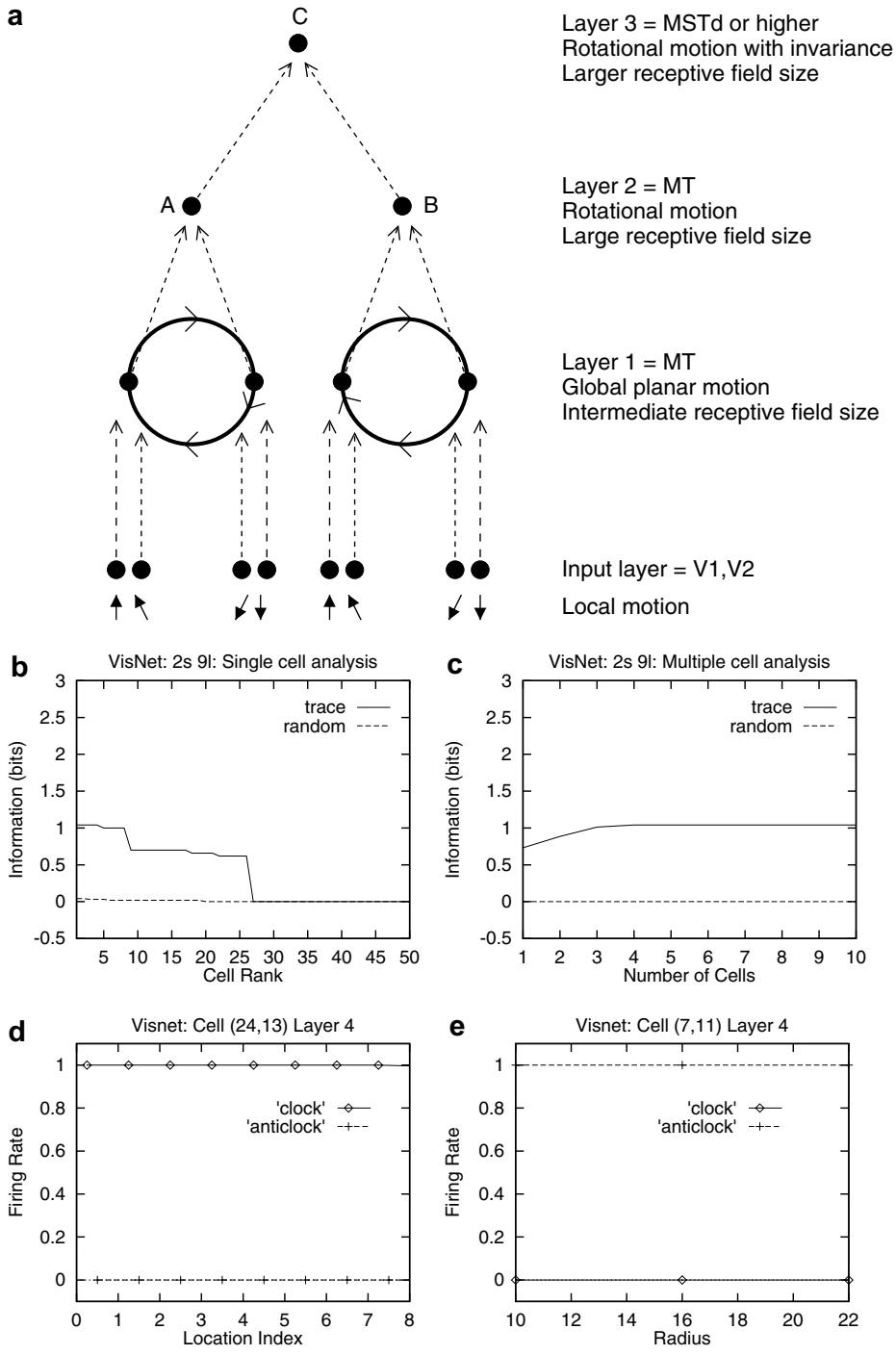

Fig. 6. (a) Two rotating wheels at different locations rotating in opposite directions. The local flow field is ambiguous. Clockwise or counterclockwise rotation can only be diagnosed by a global flow computation, and it is shown how the network is expected to solve the problem to produce position invariant 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, independently 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 vs anticlockwise) after training with the trace rule, but not with random initial synaptic weights in the untrained control condition. (c) The multiple cell information measure shows 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 9 positions. (e) Size invariance illustrated for a single cell from layer 4, which after training three different radii of rotating wheel, responded only to anticlockwise rotation, independently of the size of the rotating wheels. The training grid spacing was 32 pixels, and the radii of the wheels was 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 9 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 9 locations. For the size invariant simulations, the network was trained and tested with the set of clockwise vs anticlockwise rotating wheels presented in three sizes.

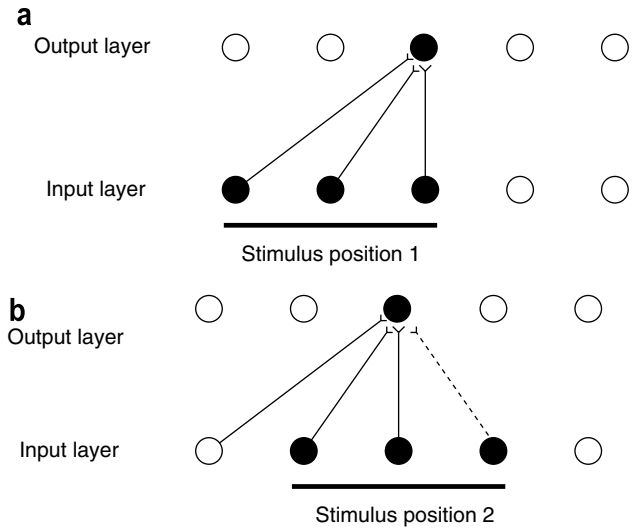

Fig. 7. An illustration of how CT learning would function in a network with a single layer of forward synaptic connections between an input layer of neurons and an output layer. Initially the forward synaptic weights are set to random values. The top part (a) shows the initial presentation of a stimulus to the network in position 1. Activation from the (shaded) active input cells is transmitted through the initially random forward connections to stimulate the cells in the output layer. The shaded cell in the output layer wins the competition in that layer. The weights from the active input cells to the active output neuron are then strengthened using an associative learning rule. The bottom part (b) shows what happens after the stimulus is shifted by a small amount to a new partially overlapping position 2. As some of the active input cells are the same as those that were active when the stimulus was presented in position 1, the same output cell is driven by these previously strengthened afferents to win the competition again. The rightmost shaded input cell activated by the stimulus in position 2, which was inactive when the stimulus was in position 1, now has its connection to the active output cell strengthened (denoted by the dashed line). Thus the same neuron in the output layer has learned to respond to the two input patterns that have similar vector elements in common. As can be seen, the process can be continued for subsequent shifts, provided that a sufficient proportion of input cells stay active between individual shifts.

complex 3D objects, and using the same training images in human psychophysical investigations, showed that view invariant object learning can occur when spatial but not temporal continuity applies in a training condition in which the images of different objects were interleaved. However, they also found that the human view invariance learning was better if sequential presentation of the images of an object was used, indicating that temporal continuity is an important factor in human invariance learning.

Perry et al. (2006) extended the use of continuous transformation learning to translation invariance. They showed that translation invariant representations can be learned by continuous transformation learning; that the transforms must be close for this to occur; that the temporal order of presentation of each transformed image during training is not crucial for learning to occur; that relatively large numbers of transforms can be learned; and that such continuous transformation learning can be usefully combined with temporal trace training.

2.12. Lighting invariance

Object recognition should occur correctly even despite variations of lighting. In an investigation of this, Stringer and Rolls as described here trained VisNet on a set of 3D objects generated with OpenGL in which the viewing angle and lighting source could be independently varied (see Fig. 8). After training with the trace rule on all the 180 views (separated by 1°, and rotated about the vertical axis in Fig. 8) of each of the four objects under the left lighting condition, we tested whether the network would recognize the objects correctly when they were shown again, but with the source of the lighting moved to the right so that the objects appeared different (see Fig. 8). Fig. 9 shows the single and multiple cell information measures for the set of objects tested with the light source in the same position as during training (Left-light), and that the measures were almost as good with testing with the light source moved to the right position (Right-light). Thus lighting invariant object recognition was demonstrated.

Some insight into the good performance with a change of lighting is that some neurons in the inferior temporal visual cortex respond to the outlines of 3D objects (Vogels and Biederman, 2002), and these outlines will be relatively

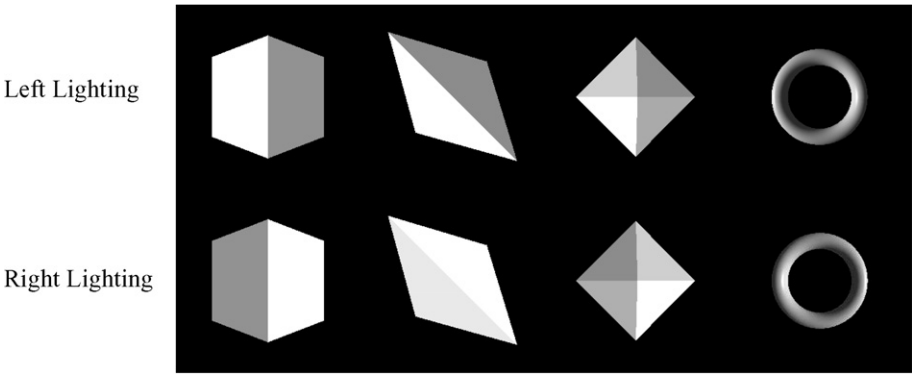

Fig. 8. Lighting invariance. VisNet was trained on a set of 3D objects (cube, tetrahedron, octahedron and torus) generated with OpenGL in which for training the objects had left lighting, and for testing the objects had right lighting. Just one view of each object is shown in the figure, but for training and testing 180 views of each object separated by 1° were used.

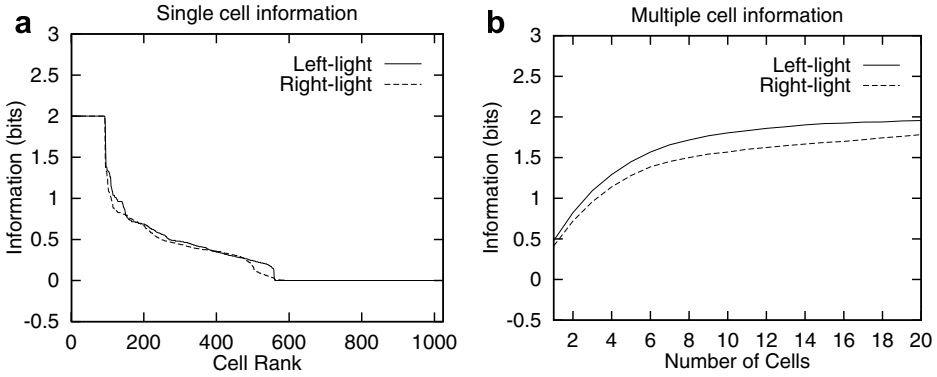

Fig. 9. Lighting invariance. The performance of the network after training with 180 views of each object lit from the left, when tested with the lighting again from the left (Left-light), and when tested with the lighting from the right (Right-light). The single cell information measure shows that many single neurons in layer 4 had the maximum amount of information about the objects, 2 bits, which indicates that they responded to all 180 views of one of the objects, and none of the 180 views of the other objects. The multiple cell information shows that the cells were sufficiently different in the objects to which they responded invariantly that all of the objects were perfectly represented when tested with the training images, and very well represented (with nearly 2 bits of information) when tested in the untrained lighting condition.

consistent across lighting variations. Although the features about the object represented in VisNet will include more than the representations of the outlines, the network may because it uses distributed representations of each object generalize correctly provided that some of the features are similar to those present during training. Under very difficult lighting conditions, it is likely that the performance of the network could be improved by including variations in the lighting during training, so that the trace rule could help to build representations that are explicitly invariant with respect to lighting.

3. Further approaches to invariant object recognition

A related approach to invariant object recognition is described by Riesenhuber and Poggio (1999b), and builds on the hypothesis that not just shift invariance (as implemented in the Neocognitron), but also other invariances such as scale, rotation and even view, could be built into a feature hierarchy system, as suggested by Rolls (1992) (see also Perrett and Oram, 1993). The approach of Riesenhuber and Poggio (1999b) (see also Riesenhuber and Poggio, 1999a, 2000) is a feature hierarchy approach which uses alternate 'simple cell' and 'complex cell' layers in a way analogous to Fukushima (1980). The function of each S cell layer is to build more complicated features from the inputs, and works by template matching. The function of each 'C' cell layer is to provide some translation invariance over the features discovered in the preceding simple cell layer (as in Fukushima (1980)), and operates by performing a MAX function on the inputs. The non-linear MAX function makes a complex cell respond only to whatever is the highest activity input being received, and is part of the process by which invariance is achieved according to this proposal. This C layer process involves 'implicitly scanning over afferents of the same type differing in the parameter of the transformation to which responses should be invariant (for instance, feature size for scale invariance), and then selecting the best-matching afferent' (Riesenhuber and Poggio, 1999b). Brain mechanisms by which this computation could be set up are not part of the scheme, and the model is effectively hand-wired, so does not yet provide a biologically plausible model of invariant object recognition. However, the fact that the model sets out to achieve some of the processes specified by Rolls (1992) and implemented in VisNet (see Section 2) does represent useful convergent thinking towards how invariant object recognition might be implemented in the brain.

Another approach to the implementation of invariant representations in the brain is the use of neurons with Sigma–Pi synapses. Sigma–Pi synapses, described by Rolls and Deco (2002), effectively allow one input to a synapse to be multiplied or gated by a second input to the synapse. The multiplying input might gate the appropriate set of the other inputs to a synapse to produce the shift or scale change required. For example, one set of inputs at the synapses could be a signal that varies with the shift required to compute translation invariance, effectively mapping the appropriate set of x_j inputs through to the output neurons depending on the shift required (Mel et al., 1998; Mel and Fiser, 2000; Olshausen et al., 1993; Olshausen et al., 1995). Local operations on a dendrite could be involved in such a process (Mel et al., 1998). The explicit neural implementation of the gating mechanism seems implausible, given the need to multiply and thus remap large parts of the retinal input depending on shift and scale modifying connections to a particular set of output neurons. Moreover, the explicit control signal to set the multiplication required in V1 has not been identified. Moreover, if this was the solution used by the brain, the whole problem of shift and scale invariance could in principle be solved in one layer of the system, rather than with the multiple hierarchically organized set of layers actually used in the brain, as shown schematically in Fig. 1. The multiple layers actually used in the brain are much more consistent with the type of scheme incorporated in VisNet. Moreover, if a multiplying system

of the type hypothesized by Mel et al. (1998), Olshausen et al. (1993) and Olshausen et al. (1995) was implemented in a multilayer hierarchy with the shift and scale change emerging gradually, then the multiplying control signal would need to be supplied to every stage of the hierarchy. A further problem with such approaches is how the system is trained in the first place.

In conclusion, the neurophysiological and computational approach taken here focusses on a feature hierarchy model in which invariant representations can be built by self-organizing learning based on the statistics of the visual input. The model can use temporal continuity in an associative synaptic learning rule with a short term memory trace, and/or it can use spatial continuity in Continuous Transformation learning. The model of visual processing in the ventral cortical stream can build representations of objects that are invariant with respect to translation, view, size, and in this paper we show also lighting. The model has been extended to provide an account of invariant representations in the dorsal visual system of the global motion produced by objects such as looming, rotation, and object-based movement. The model has been extended to incorporate top-down feedback connections to model the control of attention by biased competition in for example spatial and object search tasks. The model has also been extended to account for how the visual system can select single objects in complex visual scenes, and how multiple objects can be represented in a scene.

References

Abeles, M., 1991. Corticonics: Neural Circuits of the Cerebral Cortex. Cambridge University Press, Cambridge.

Ackley, D.H., Hinton, G.E., Sejnowski, T.J., 1985. A learning algorithm for Boltzmann machines. Cognitive Science 9, 147–169.

Aggelopoulos, N.C., Rolls, E.T., 2005. Natural scene perception: inferior temporal cortex neurons encode the positions of different objects in the scene. European Journal of Neuroscience 22, 2903–2916.

Aggelopoulos, N.C., Franco, L., Rolls, E.T., 2005. Object perception in natural scenes: encoding by inferior temporal cortex simultaneously recorded neurons. Journal of Neurophysiology 93, 1342–1357.

Biederman, I., 1972. Perceiving real-world scenes. Science 177, 77–80.

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.

Chelazzi, L., Miller, E., Duncan, J., Desimone, R., 1993. A neural basis for visual search in inferior temporal cortex. Nature (London) 363, 345–347.

Corchs, S., Deco, G., 2002. Large-scale neural model for visual attention: integration of experimental single cell and fMRI data. Cerebral Cortex 12, 339–348.

Corchs, S., Deco, G., 2004. Feature-based attention in human visual cortex: simulation of fMRI data. Neuroimage 21, 36–45.

Cowey, A., Rolls, E.T., 1975. Human cortical magnification factor and its relation to visual acuity. Experimental Brain Research 21, 447–454.

Deco, G., Lee, T.S., 2002. A unified model of spatial and object attention based on inter-cortical biased competition. Neurocomputing 44–46, 775–781.

Deco, G., Rolls, E.T., 2002. Object-based visual neglect: a computational hypothesis. European Journal of Neuroscience 16, 1994–2000.

Deco, G., Rolls, E.T., 2003. Attention and working memory: a dynamical model of neuronal activity in the prefrontal cortex. European Journal of Neuroscience 18, 2374–2390.

Deco, G., Rolls, E.T., 2004. A neurodynamical cortical model of visual attention and invariant object recognition. Vision Research 44, 621–644.

Deco, G., Rolls, E.T., 2005a. Attention, short term memory, and action selection: a unifying theory. Progress in Neurobiology 76, 236–256.

Deco, G., Rolls, E.T., 2005b. Neurodynamics of biased competition and cooperation for attention: a model with spiking neurons. Journal of Neurophysiology 94, 295–313.

Deco, G., Rolls, E.T., 2005c. Synaptic and spiking dynamics underlying reward reversal in the orbitofrontal cortex. Cerebral Cortex 15, 15–30.

Deco, G., Rolls, E.T., 2006. A neurophysiological model of decision-making and Weber's law. European Journal of Neuroscience 24, 901–916.

Deco, G., Zihl, J., 2001. Top-down selective visual attention: a neurodynamical approach. Visual Cognition 8, 119–140.

Deco, G., Pollatos, O., Zihl, J., 2002. The time course of selective visual attention: theory and experiments. Vision Research 42, 2925–2945.

Deco, G., Rolls, E.T., Horwitz, B., 2004. 'What' and 'where' in visual working memory: a computational neurodynamical perspective for integrating fMRI and single-neuron data. Journal of Cognitive Neuroscience 16, 683–701.

Desimone, R., Duncan, J., 1995. Neural mechanisms of selective visual attention. Annual Review of Neuroscience 18, 193–222.

Duncan, J., 1996. Cooperating brain systems in selective perception and action. In: Inui, T., McClelland, J.L. (Eds.), Attention and Performance XVI. MIT Press, Cambridge, Mass, pp. 549–578.

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.

Feldman, J.A., 1985. Four frames suffice: a provisional model of vision and space. Behavioural Brain Sciences 8, 265–289.

Finkel, L.H., Edelman, G.M., 1987. Population rules for synapses in networks. In: Edelman, G.M., Gall, W.E., Cowan, W.M. (Eds.), Synaptic Function. John Wiley & Sons, New York, pp. 711–757.

Földiák, P., 1991. Learning invariance from transformation sequences. Neural Computation 3, 193–199.

Földiák, P., 1992. Models of sensory coding, Technical Report CUED/F–INFENG/TR 91, University of Cambridge, Department of Engineering.

Franco, L., Rolls, E.T., Aggelopoulos, N.C., Treves, A., 2004. The use of decoding to analyze the contribution to the information of the correlations between the firing of simultaneously recorded neurons. Experimental Brain Research 155, 370–384.

Franco, L., Rolls, E.T., Aggelopoulos, N.C., Jerez, J.M., 2007. Neuronal selectivity, population sparseness, and ergodicity in the inferior temporal visual cortex, Submitted for publication.

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.

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., 1989. The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. Behavioural Brain Research 32, 203–218.

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.

Hawken, M.J., Parker, A.J., 1987. Spatial properties of the monkey striate cortex. Proceedings of the Royal Society, London [B] 231, 251–288.

Hertz, J.A., Krogh, A., Palmer, R.G., 1991. Introduction to the Theory of Neural Computation. Addison-Wesley, Wokingham, UK.

Hummel, J.E., Biederman, I., 1992. Dynamic binding in a neural network for shape recognition. Psychological Review 99, 480–517.
Koenderink, J.J., 1990. Solid Shape. MIT Press, Cambridge, MA.
Koenderink, J.J., Van Doorn, A.J., 1979. The internal representation of solid shape with respect to vision. Biological Cybernetics 32, 211–217.
Logothetis, N.K., Pauls, J., Bulthoff, H.H., Poggio, T., 1994. View-dependent object recognition by monkeys. Current Biology 4, 401–414.
Logothetis, N.K., Pauls, J., Poggio, T., 1995. Shape representation in the inferior temporal cortex of monkeys. Current Biology 5, 552–563.
Malsburg, C.v.d., 1973. Self-organization of orientation-sensitive columns in the striate cortex. Kybernetik 14, 85–100.
Malsburg, C.v.d., 1990. A neural architecture for the representation of scenes. In: McGaugh, J.L., Weinberger, N.M., Lynch, G. (Eds.), Brain Organization and Memory: Cells, Systems and Circuits. Oxford University Press, New York, pp. 356–372 (Chapter 19).
Martinez-Trujillo, J., Treue, S., 2002. Attentional modulation strength in cortical area MT depends on stimulus contrast. Neuron 35, 365–370.
Mel, B.W., Fiser, J., 2000. Minimizing binding errors using learned conjunctive features. Neural Computation 12, 731–762.
Mel, B.W., Ruderman, D.L., Archie, K.A., 1998. Translation-invariant orientation tuning in visual "complex" cells could derive from intradendritic computations. Journal of Neuroscience 18 (11), 4325–4334.
Montague, R., Gally, J., Edelman, G., 1991. Spatial signalling in the development and function of neural connections. Cerebral Cortex 1, 199–220.
Moran, J., Desimone, R., 1985. Selective attention gates visual processing in the extrastriate cortex. Science 229, 782–784.
Movshon, J.A., Adelson, E.H., Gizzi, M.S., Newsome, W.T., 1985. The analysis of moving visual patterns. In: Chagas, C., Gattass, R., Gross, C. (Eds.), Pattern Recognition Mechanisms. Springer-Verlag, New York, pp. 117–151.
Mozer, M.C., 1991. The Perception of Multiple Objects: A Connectionist Approach. MIT Press, Cambridge, Massachusetts.
Newsome, W.T., Britten, K.H., Movshon, J.A., 1989. Neuronal correlates of a perceptual decision. Nature 341, 52–54.
Oja, E., 1982. A simplified neuron model as a principal component analyzer. Journal of Mathematical Biology 15, 267–273.
Olshausen, B.A., Anderson, C.H., Van Essen, D.C., 1993. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. Journal of Neuroscience 13, 4700–4719.
Olshausen, B.A., Anderson, C.H., Van Essen, D.C., 1995. A multiscale dynamic routing circuit for forming size- and position-invariant object representations. Journal of Computational Neuroscience 2, 45–62.
O'Reilly, R., Johnson, M., 1994. Object recognition and sensitive periods: a computational analysis of visual imprinting. Neural Computation 6, 357–389.
Perrett, D., Oram, M., 1993. Neurophysiology of shape processing. Image and Vision Computing 11 (6), 317–333.
Perrett, D.I., Rolls, E.T., Caan, W., 1982. Visual neurons responsive to faces in the monkey temporal cortex. Experimental Brain Research 47, 329–342.
Perrett, D.I., Smith, P.A.J., Potter, D.D., Mistlin, A.J., Head, A.S., Milner, D., Jeeves, M.A., 1985. Visual cells in temporal cortex sensitive to face view and gaze direction. Proceedings of the Royal Society of London, Series B 223, 293–317.
Perry, G., Rolls, E.T., Stringer, S.M., 2006. Continuous transformation learning of translation invariant representations, submitted for publication.
Perry, G., Rolls, E.T., Stringer, S.M., in press. Spatial vs temporal continuity in view invariant visual object recognition learning. Vision Research.
Poggio, T., Edelman, S., 1990. A network that learns to recognize three-dimensional objects. Nature 343, 263–266.
Renart, A., Parga, N., Rolls, E.T., 2000. A recurrent model of the interaction between the prefrontal cortex and inferior temporal cortex in delay memory tasks. In: Solla, S., Leen, T., Mueller, K.-R. (Eds.), Advances in Neural Information Processing Systems, vol. 12. MIT Press, Cambridge, Mass, pp. 171–177.
Reynolds, J., Desimone, R., 1999. The role of neural mechanisms of attention in solving the binding problem. Neuron 24, 19–29.
Reynolds, J.H., Chelazzi, L., Desimone, R., 1999. Competitive mechanisms subserve attention in macaque areas V2 and V4. Journal of Neuroscience 19, 1736–1753.
Rhodes, P., 1992. The open time of the NMDA channel facilitates the self-organisation of invariant object responses in cortex. Society for Neuroscience Abstracts 18, 740.
Riesenhuber, M., Poggio, T., 1999a. Are cortical models really bound by the "binding problem"? Neuron 24, 87–93.
Riesenhuber, M., Poggio, T., 1999b. Hierarchical models of object recognition in cortex. Nature Neuroscience 2, 1019–1025.
Riesenhuber, M., Poggio, T., 2000. Models of object recognition. Nature Neuroscience Supplement 3, 1199–1204.
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., 1994. Brain mechanisms for invariant visual recognition and learning. Behavioural Processes 33, 113–138.
Rolls, E.T., 1995. Learning mechanisms in the temporal lobe visual cortex. Behavioural Brain Research 66, 177–185.
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., 2003. Consciousness absent and present: a neurophysiological exploration. Progress in Brain Research 144, 95–106.
Rolls, E.T., 2005. Emotion Explained. Oxford University Press, Oxford.
Rolls, E.T., 2006. The representation of information about faces in the temporal and frontal lobes of primates including humans. Neuropsychologia, in press.
Rolls, E.T., 2007. Invariant representations of objects in natural scenes in the temporal cortex visual areas. In: Funahashi, S. (Ed.), Representation and the Brain. Springer-Verlag, Berlin.
Rolls, E.T., Baylis, G.C., 1986. Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. Experimental Brain Research 65, 38–48.
Rolls, E.T., Cowey, A., 1970. Topography of the retina and striate cortex and its relationship to visual acuity in rhesus monkeys and squirrel monkeys. Experimental Brain Research 10, 298–310.
Rolls, E.T., Deco, G., 2002. Computational Neuroscience of Vision. Oxford University Press, Oxford.
Rolls, E.T., Deco, G., 2006. Attention in natural scenes: neurophysiological and computational bases. Neural Networks, in 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., Stringer, S.M., 2006. Invariant global motion recognition in the dorsal visual system: a unifying theory. Neural Computation, in press.
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., 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., 1998. Neural Networks and Brain Function. Oxford University Press, Oxford.
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., Tovee, M.J., 1997a. 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., 1997b. Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. Journal of Computational Neuroscience 4, 309–333.

Rolls, E.T., Tovee, M.J., Panzeri, S., 1999. The neurophysiology of backward visual masking: information analysis. Journal of Cognitive Neuroscience 11, 335–346.

Rolls, E.T., Aggelopoulos, N.C., Zheng, F., 2003. The receptive fields of inferior temporal cortex neurons in natural scenes. Journal of Neuroscience 23, 339–348.

Rolls, E.T., Aggelopoulos, N.C., Franco, L., Treves, A., 2004. Information encoding in the inferior temporal visual cortex: contributions of the firing rates and the correlations between the firing of neurons. Biological Cybernetics 90, 19–32.

Rolls, E.T., Franco, L., Aggelopoulos, N.C., Jerez, J.M., 2006. Information in the first spike, the order of spikes, and the number of spikes provided by neurons in the inferior temporal visual cortex. Vision Research, in press.

Rosenblatt, F., 1961. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan, Washington, DC.

Singer, W., 1999. Neuronal synchrony: a versatile code for the definition of relations? Neuron 24, 49–65.

Singer, W., Gray, C.M., 1995. Visual feature integration and the temporal correlation hypothesis. Annual Review of Neuroscience 18, 555–586.

Singer, W., Gray, C., Engel, A., Konig, P., Artola, A., Brocher, S., 1990. Formation of cortical cell assemblies. Cold Spring Harbor Symposium on Quantitative Biology 55, 939–952.

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.

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.

Sutton, R.S., Barto, A.G., 1981. Towards a modern theory of adaptive networks: expectation and prediction. Psychological Review 88, 135–170.

Szabo, M., Almeida, R., Deco, G., Stetter, M., 2004. Cooperation and biased competition model can explain attentional filtering in the prefrontal cortex. European Journal of Neuroscience 19, 1969–1977.

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

Tanaka, K., Saito, C., Fukada, Y., Moriya, M., 1990. Integration of form, texture, and color information in the inferotemporal cortex of the macaque. In: Iwai, E., Mishkin, M. (Eds.), Vision, Memory and the Temporal Lobe. Elsevier, New York, pp. 101–109 (Chapter 10).

Tovee, M.J., Rolls, E.T., 1995. Information encoding in short firing rate epochs by single neurons in the primate temporal visual cortex. Visual Cognition 2, 35–58.

Tovee, M.J., Rolls, E.T., Treves, A., Bellis, R.P., 1993. Information encoding and the responses of single neurons in the primate temporal visual cortex. Journal of Neurophysiology 70, 640–654.

Tovee, M.J., Rolls, E.T., Azzopardi, P., 1994. Translation invariance and the responses of neurons in the temporal visual cortical areas of primates. Journal of Neurophysiology 72, 1049–1060.

Trappenberg, T.P., Rolls, E.T., Stringer, S.M., 2002. Effective size of receptive fields of inferior temporal visual cortex neurons in natural scenes. In: Dietterich, T.G., Becker, S., Gharamani, Z. (Eds.), Advances in Neural Information Processing Systems, vol. 14. MIT Press, Cambridge, MA, pp. 293–300.

Ullman, S., 1996. High-level visionObject Recognition and Visual Cognition. Bradford/MIT Press, Cambridge, Mass.

Usher, M., Niebur, E., 1996. Modelling the temporal dynamics of IT neurons in visual search: a mechanism for top-down selective attention. Journal of Cognitive Neuroscience 8, 311–327.

Van Essen, D., Anderson, C.H., Felleman, D.J., 1992. Information processing in the primate visual system: an integrated systems perspective. Science 255, 419–423.

Vogels, R., Biederman, I., 2002. Effects of illumination intensity and direction on object coding in macaque inferior temporal cortex. Cerebral Cortex 12, 756–766.

Wallis, G., Baddeley, R., 1997. Optimal unsupervised learning in invariant object recognition. Neural Computation 9, 883–894.

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: Kandel, E.R., Schwartz, J.H., Jessell, T.M. (Eds.), Principles of Neural Science, fourth ed. McGraw-Hill, New York, pp. 548–571.

Yamane, S., Kaji, S., Kawano, K., 1988. What facial features activate face neurons in the inferotemporal cortex of the monkey? Experimental Brain Research 73, 209–214.