Available online at www.sciencedirect.com
SCIENCE DIRECT®
Vision Research 44 (2004) 621–642
Vision Research
www.elsevier.com/locate/visres

A Neurodynamical cortical model of visual attention and invariant object recognition

Gustavo Deco [a], Edmund T. Rolls [b,*]

[a] Department of Technology, Computational Neuroscience, Institució Catalana de Recerca i Estudis Avançats (ICREA), Universitat Pompeu Fabra, Passeig de Circumval.lació, 08003 Barcelona, Spain

[b] Department of Experimental Psychology, Centre for Computational Neuroscience, University of Oxford, South Parks Road, Oxford OX1 3UD, UK

Received 10 February 2003

Abstract

We describe a model of invariant visual object recognition in the brain that incorporates feedback biasing effects of top–down attentional mechanisms on a hierarchically organized set of visual cortical areas with convergent forward connectivity, reciprocal feedback connections, and local intra-area competition. The model displays space-based and object-based covert visual search by using attentional top–down feedback from either the posterior parietal or the inferior temporal cortex (IT) modules, and interactions between the two processing streams occurring in $V1$ and $V2$. The model explains the gradually increasing magnitude of the attentional modulation that is found in fMRI experiments from earlier visual areas ($V1$, $V2$) to higher ventral stream visual areas ($V4$, IT); how the effective size of the receptive fields of IT neurons becomes smaller in natural cluttered scenes; and makes predictions about interactions between stimuli in their receptive fields.

Keywords: Attention; Inferior temporal visual cortex; Parietal cortex; Biassed competition; Attractor dynamics

1. Introduction

Vision is a sufficiently complex problem that it benefits from a computational neuroscience approach that is closely linked to empirical neurophysiological investigations. A number of different approaches to issues such as invariant object recognition, and visual attention, are described by Rolls and Deco (2002). In the research described here, we describe a neurophysiologically based model for invariant visual object recognition and attention in primates that combines a feature hierarchy approach to invariant object recognition (exemplified by VisNet) (Elliffe, Rolls, & Stringer, 2002; Rolls, 1992; Rolls & Milward, 2000; Stringer & Rolls, 2000, 2002; Wallis & Rolls, 1997; Wallis, Rolls, & Foldiak, 1993) with a model of spatial and object attention that incorporates backprojections for top–down attentional effects, and interactions between a dorsal 'where' and ventral 'what' visual stream (Deco, 2001; Deco & Lee, 2002; Deco & Rolls, 2002; Deco & Zihl, 2001; Rolls & Deco, 2002). This model is defined in a neurodynamical framework, i.e. the underlying dynamics is founded on biophysical mathematical models of single neurons, with the neurons interconnected to form networks which correspond to particular brain areas. In this paper we focus in particular on the locally implemented but gradually increasing global character of the competition that is produced in a hierarchical network with convergent forward connectivity from area to area; and on the interaction between space-based and object-based attentional top–down feedback processes.

VisNet is a four-layer feedforward network with convergence to each part of a layer from a small region of the preceding layer, with competition between the neurons within a layer, and with a trace learning rule to help it learn transform invariance. The trace rule is a modified Hebbian rule, which modifies synaptic weights according to both the current firing rates and the firing rates to recently seen stimuli (Rolls & Stringer, 2001). This enables neurons to learn to respond similarly to the gradually transforming inputs it receives, which over the

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

0042-6989/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2003.09.037

short term are statistically likely to be about the same object. This kind of hierarchical feature analysis system has the capability of representing the spatial relations between features, by incorporating fixed (non-dynamic) feature combination neurons which respond to a combination of a small number of features that are in the correct spatial relation to each other (Elliffe et al., 2002).

The attentional model of Deco (2001) consists of a set of modules with feedforward and also feedback connections between each module; a neurodynamical formulation expressed in terms of mean field theory that allows a 'biased competition' model of attention to operate; and a set of modules representing the 'what' pathway with another set of modules that can interact through $V1$ that represent the 'where' pathway. Attention then appears as an emergent effect [1] related to the dynamical evolution of the whole network to a state where the constraints given by the stimulus and the external attentional object or spatial bias are satisfied (Corchs & Deco, 2002; Deco & Lee, 2002; Deco & Zihl, 2001; Rolls & Deco, 2002).

The aim of the research described here, and the new model presented, is to combine the feedforward feature hierarchy approach used by VisNet, and the multimodular attentional architecture with both forward and 'top–down' backprojections, into a single unified model. [2] First, we show that the computational principles of both models are captured in the combined model. Second, the model accounts for the gradually increasing magnitude of the attentional modulation that is found in fMRI experiments from earlier visual areas ($V1$, $V2$) to higher ventral stream visual areas ($V4$, IT). Third, the model shows how the effective size of the receptive fields of IT neurons becomes smaller in natural cluttered scenes. Fourth, the model also makes new experimental predictions about two types of interaction between stimuli in the receptive fields of IT neurons, which are due to competition at early vs. late stages of processing in the ventral stream. This single integrated model will provide a basis for new aspects of the operation of the cortical visual system to be explored, because it incorporates several aspects of the cortical architecture of the visual systems found in the brain, including the hierarchies present in the ventral or 'what' visual system, and the backprojections in the ventral and dorsal visual systems which enable these streams to interact. This approach is very different from some earlier models of visual attention based on saliency maps which used only feedforward processing, directed attention only to the location of salient features in the environment, and did not address the issue of object identification (Itti & Koch, 2000). The present model in contrast shows how spatial and object representations in separate dorsal and ventral processing streams in the brain could interact using top–down processing to model both identification of a location given an object search target, and identification of an object given a spatial location as a search cue. Moreover, the model described here includes a model of invariant object recognition, and is a full dynamical model which enables the timing in different modules during global settling of the whole network taking into account the constraints given to be investigated.

2. The combined neurodynamical model of 'what' and 'where' visual stream processing

2.1. Neurophysiological background

2.1.1. The dorsal and ventral paths of the visual cortex

A widely accepted description of the many cortical areas (Felleman & Van Essen, 1991; Rolls & Deco, 2002; Ungerleider & Mishkin, 1982; Van Essen, Felleman, DeYoe, Olavarria, & Knierim, 1990) is into a ventral or 'what' stream that runs from $V1$, to $V2$, $V4$, and the inferior temporal cortical areas TEO and TE that computes properties of objects such as shape and colour; and a dorsal or 'where' stream that runs from $V1$, to $V2$, $V3$, MT and the medial superior temporal area MST, and on to the posterior parietal cortex (PP), including area 7a (Ungerleider & Mishkin, 1982). Neurons in the temporal cortical visual areas typically have large translation-invariant receptive fields, and have distributed encoding of shapes, objects or faces in which the spatial arrangement of the features can be important (Desimone, Albright, Gross, & Bruce, 1984; Logothetis, Pauls, & Poggio, 1995; Perrett, Rolls, & Caan, 1982; Rolls, 1984, 1992, 2000; Rolls & Deco, 2002; Tovee, Rolls, & Azzopardi, 1994). On the other hand, neurons in the parietal lobe are frequently sensitive to the location of the stimulus on the retina or with respect to the animal's head (Andersen, Snyder, & Bradley, 1997). Neurons in the posterior parietal cortex (PP) show an enhanced response to attended targets within their receptive fields, even when no eye movements are made (Bushnell, Goldberg, & Robinson, 1981), and there is correspondingly suppression of responses to unattended items (see Rolls & Deco, 2002). Consistent with this latter finding, Posner, Walker, Friedrich, and Rafal (1984) showed that damage to the

[1] Emergent effects are those effects that are not a scaling up or adaptation of anything its parts do. The dynamical evolution and the global attractors of the cortical networks are genuine emergent effects because they are only due to the connections between each part.

[2] Previous computational studies have already considered the role of feedforward bottom–up effects in visual attention (Itti & Koch, 2001). We stress here the role of biassed competition mechanisms for spatial and object-based attention involving interactions between the dorsal visual stream and the ventral visual stream through early cortical areas, and therefore the role of top–down vs. bottom–up attentional interactions.

parietal lobe in humans can block the ability to move the attentional focus away from the presently attended location to other objects in the visual field. Haxby et al. (1994) found consistent evidence for a segregation of processing streams in humans. They showed in a positron emission tomography (PET) study that when humans performed a face-matching task activation was observed in the inferior and temporal cortex and in the occipital lobe. On the other hand, when humans performed a spatial task (involving face rotation), activation was detected in the parietal and occipital cortex.

We include in the computational model we describe this what–where segregation by providing a set of 'ventral stream' modules that correspond approximately to visual areas $V1$, $V2$, $V4$, IT, and a set of 'dorsal stream' modules that correspond approximately to $V1$, $V2$ and PP.

2.1.2. The biased competition hypothesis of attention and visual search

The dichotomy between parallel and serial operations in visual search has been challenged by psychological models suggesting that all types of search task can be solved by a single parallel competitive mechanism. Duncan (1980) and Duncan and Humphreys (1989) have proposed a scheme that integrates both attentional modes (parallel and serial) as an instantiation of a common principle. They explain searches for conjunctions of features as well as for single features on the basis of the same operations involving grouping between items in the visual field, and matching of those items or groups to a memory template of the target. The matching process leads to support of items with features consistent with the template and inhibits those with different features. This process would operate for all stimulus features: colour, shape, location, etc. This process of feature selection suggests that subjects utilize top–down information (from the feature-based or object memory template) independently of stimulus location in space. The attentional theory of Duncan and Humphreys (1989) proposes that there is both parallel activation of a target template (from multiple items in the field), and competition between items (and between the template and non-matching items), so that, finally, only one object is selected. There is evidence suggesting that parallel competitive processes in the brain are responsible for human performance in visual selective attention tasks (Duncan, Humphreys, & Ward, 1997).

A number of neurophysiological experiments (Chelazzi, 1998; Chelazzi, Miller, Duncan, & Desimone, 1993; Miller, Gochin, & Gross, 1993; Moran & Desimone, 1985; Motter, 1993, 1994a, 1994b; Reynolds & Desimone, 1999; Rolls & Tovee, 1995; Spitzer, Desimone, & Moran, 1988) have been performed suggesting biased competition neural mechanisms which are consistent with the theory of Duncan and Humphreys (1989) (i.e., with a role for a top–down memory target template in visual search). The biased competition hypothesis proposes that multiple stimuli in the visual field activate populations of neurons that engage in competitive interactions. Attending to a stimulus at a particular location or with a particular feature biases this competition in favour of neurons that respond to the location of or the features in the attended stimulus. This attentional effect is produced by generating signals in areas outside the visual cortical areas which are then fed back to extrastriate areas, where they bias the competition in such a way that when multiple stimuli appear in the visual field, the cells representing the attended stimulus win, thereby suppressing cells representing distracting stimuli (Desimone & Duncan, 1995; Duncan, 1996; Duncan & Humphreys, 1989).

In addition, there is consistent evidence for similar mechanisms in human extrastriate cortex at the macroscopic level of functional magnetic resonance imaging (fMRI) (Kastner, De Weerd, Desimone, & Ungerleider, 1998; Kastner, Pinsk, De Weerd, Desimone, & Ungerleider, 1999). These studies have shown that multiple stimuli in the visual field interact in a mutually suppressive way when presented simultaneously but not when presented sequentially, and that spatially directed attention to one stimulus location reduces the mutually suppressive effect. They also revealed increased activity in extrastriate visual cortex in the absence of visual stimulation when subjects covertly directed attention to a peripheral location where the onset of visual stimuli was expected. This increased activity in extrastriate visual cortex was related to a top–down bias of neural signals in favour of the attended location, which was presumably derived from frontal and parietal cortical areas.

Our model implements biased competition at the microscopic level of neuronal pools and at the mesoscopic level of visual areas in a multi-modular architecture with 'what' and 'where' streams (Corchs & Deco, 2002; Deco & Zihl, 2001). At the neuronal pool level, dynamical competition is implemented by introducing mutual inhibition using pools of inhibitory neurons. Intermodular competition and mutual biasing result from the interaction between modules corresponding to different visual areas. In the model, feature attention biases intermodular competition between $V1$, $V2$, $V4$ and IT, whereas spatial attention biases intermodular competition between $V1$, $V2$, $V4$, and PP. The model allows simulation of single cell, fMRI and neuropsychological findings, and produces results which are consistent with the experimental observations of biased competition effects (Corchs & Deco, 2002; Deco & Zihl, 2001; Rolls & Deco, 2002).

2.2. A large-scale neurodynamical model of the visual cortex

The neurophysiological findings described above, wider considerations on the possible computational theory underlying hierarchical feedforward processing in the visual cortical areas with layers of competitive networks trained with a trace learning rule (Elliffe et al., 2002; Rolls, 1992; Rolls & Deco, 2002; Wallis & Rolls, 1997), and the analysis of the role of attentional feedback connections and interactions between an object and a spatial processing stream (Deco, 2001; Deco & Lee, 2002; Deco & Zihl, 2001; Rolls & Deco, 2002), lead to the neurodynamical model that we present here for invariant hierarchical object recognition and selective visual attention. The equations for the model are provided in Appendix A, and the following text explains the model. Within each module a competitive network is implemented by local lateral inhibitory connections, and the modules are connected hierarchically by convergent feedforward connections. A modified Hebb-like learning rule that incorporates a temporal trace of each cell's previous activity enables the neurons to learn transform invariant responses. The model implements biased competition by assuming mutually feedforward and feedback biasing between different modules corresponding to different brain areas. The different modules are organized in a hierarchical structure incorporating the overall architectural arrangement of the visual cortical areas with ventral and dorsal pathways, which can interact in $V1$ and $V2$. The model is fully autonomous and each component of its functional behaviour is explicitly described in a complete mathematical framework (provided in Appendix A), which at the microscopic level corresponds to the neurodynamical equations derived by Wilson and Cowan (1972) for a pool of spiking neurons.

Fig. 1 shows the overall systems-level diagram of the multi-area neurodynamical architecture used for modelling the primate visual cortical areas. The system is essentially composed of five modules or networks structured such that they resemble the two known main visual processing streams 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 modules $V2$, $V4$ to IT (the inferior temporal cortex), and is mainly concerned with object recognition, independently of position and scaling. The occipito-parietal stream leads dorsally into PP (the posterior parietal complex) and is responsible for maintaining a spatial map of an object's location and/or the spatial relationship of an object's parts as well as for moving the spatial allocation of attention.

The ventral stream consists of the four modules $V1$, $V2$, $V4$ and IT. This part of the architecture is similar to VisNet in architecture and training (Elliffe et al., 2002; Rolls & Deco, 2002; Rolls & Milward, 2000; Wallis & Rolls, 1997), except that backprojections are incorporated, and the numbers of neurons are reduced for simplicity. These different modules allow combinations of features or inputs that occur 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 (Elliffe et al., 2002). This is implemented via convergent connections to each part of a layer from a small region of the preceding layer, thus allowing the receptive field size of cells to increase through the ventral visual processing areas, as is observed in the primate ventral visual stream (see Fig. 2). An external top–down bias, coming it is postulated from a short-term memory for shape features or objects in the more ventral part of the prefrontal cortex area v46, generates an object-based attentional component that is fed back down through the recurrent connections from IT through $V4$ and $V2$ to $V1$. The $V1$ module contains hyper columns, each covering a pixel in a topologically organized model of the scene. Each hypercolumn contains orientation columns of orientation-tuned (complex) cells with Gabor filter tuning at octave intervals to different spatial frequencies. $V1$ sends visual inputs to both the ventral and dorsal streams, and in turn receives backprojections from each stream, providing a high-resolution representation for the two streams to interact. This interaction between the two streams made possible by the backprojections to $V1$ is important in the model for implementing attentional effects. In the brain, there may be contributions to this interaction from further cross-links between the processing streams, occurring for example in $V2$, but the principle of the interaction is captured in the model by the common $V1$ module. The $V2$, $V4$ and IT modules each receive inputs from a small region of the preceding module, allowing the receptive field sizes of the neurons to increase gradually through the pyramidal structure of the network (see Fig. 2). Each of these modules acts like a competitive network (see Rolls & Deco, 2002; Rolls & Treves, 1998; Wallis & Rolls, 1997) which enables neurons to learn to respond to spatially organized combinations of features detected at the preceding stage, thus helping to solve the binding problem (Elliffe et al., 2002), and also implementing a certain degree of localized competitive interaction between different targets. All the feedforward connections are trained by an associative (Hebb-like) learning rule with a short-term memory (the trace learning rule) in a learning phase in order to produce invariant neuronal responses (Rolls, 1992; Wallis & Rolls, 1997). The backprojections between modules, a feature of cortical connectivity (Rolls & Deco, 2002; Rolls & Treves, 1998) are symmetric and

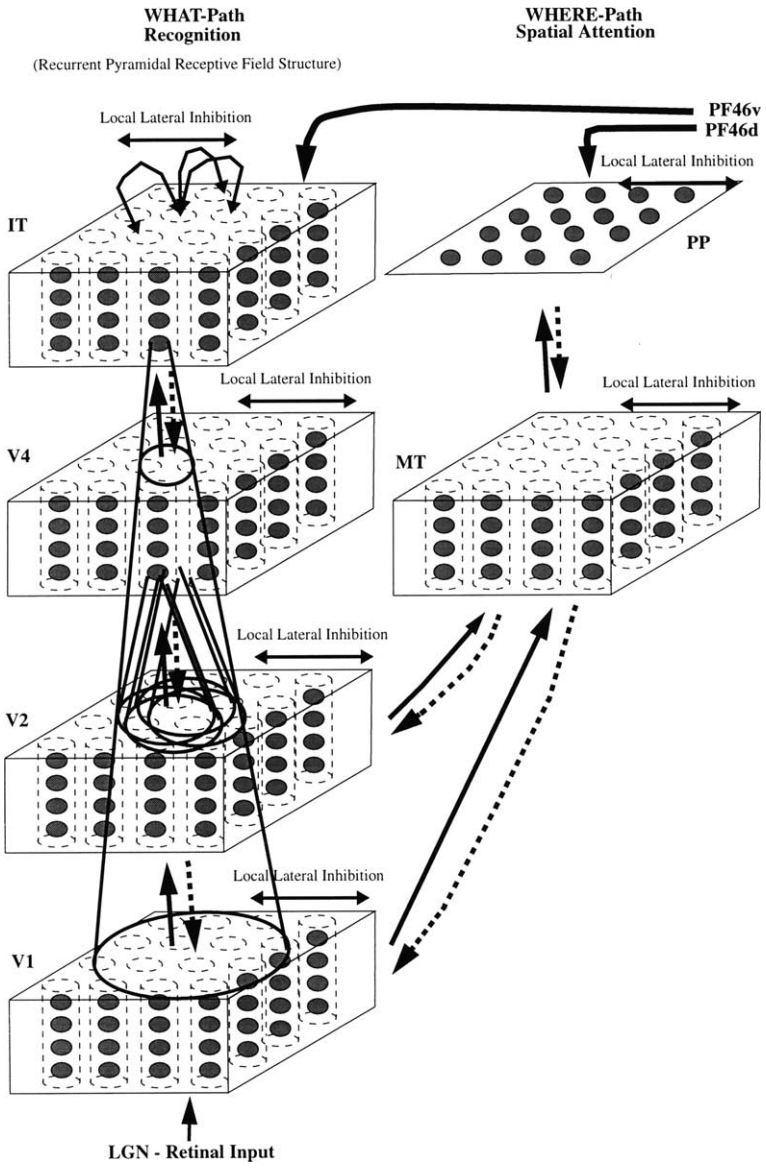

Fig. 1. Cortical architecture for hierarchical and attention-based visual perception. The system is essentially composed of five modules structured such that they resemble the two known main visual paths of the mammalian visual cortex. Information from the retino-geniculo-striate pathway enters the visual 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.

reciprocal in their connectivity with the forward connections. The average strength of the backprojections is set to be a specified fraction of the strength of the forward connections (by a single parameter in the model) so that the backprojections can influence but not dominate activity in the input layers of the hierarchy (Renart, Parga, & Rolls, 1999a, 1999b). Intramodular local competition is implemented in all modules by lateral local inhibitory connections between a neuron and its neighboring neurons via a Gaussian-like weighting factor as a function of distance (see Appendix A). The width of these Gaussian decays for $V1$, $V2$, $V4$ and IT are denoted as σ_{V1}, σ_{V2}, σ_{V4}, σ_{IT}.

The inputs to module $V1$ of the network are provided by neurons with simple cell-like receptive fields. This input filtering enables real images to be presented to the network. Following Daugman (1988) the receptive fields of these input neurons are modelled by 2D-Gabor functions. The Gabor receptive fields have five degrees of freedom given essentially by the product of an elliptical Gaussian and a complex plane wave. The first two degrees of freedom are the 2D-locations of the receptive field's centre; the third is the size of the receptive field; the fourth is the orientation of the boundaries separating excitatory and inhibitory regions; and the fifth is the symmetry. This fifth degree of freedom is given in the

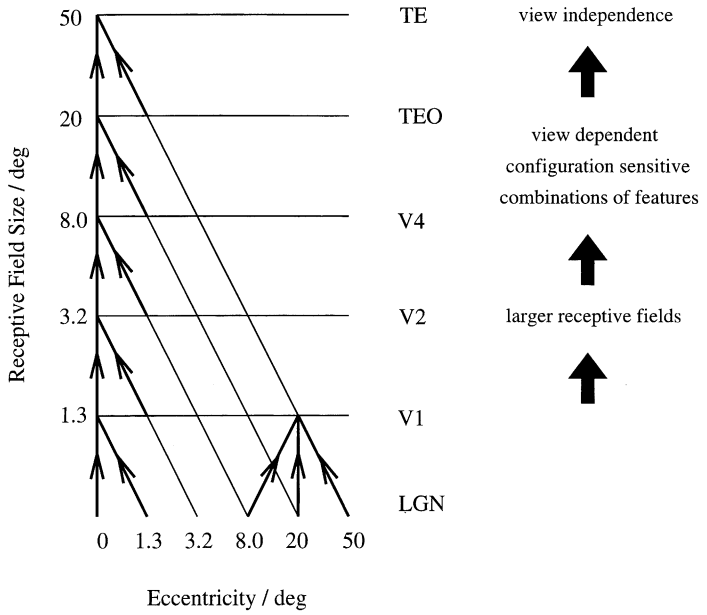

Fig. 2. Hierarchical convergent forward projection in the ventral 'what' path of the visual system achieved by a pyramidal multi-layer network, corresponding to the brain areas $V1$, $V2$, $V4$, TEO and TE (or IT), with convergence to each part of a layer from a small region of the preceding layer. The right part of the figure shows the different types of representation that may be built by implementing the biased competition hypothesis at each stage of the system. The attentive bias may correspond to recurrent attentional feedback connections, and the local competition between the neurons within a layer, may correspond to lateral local inhibitory connections. The local character of the competition within layers reveals itself effectively as a gradually increasing global competition between objects and/or parts of objects at the retina when deeper ventral layers are considered. Abbreviations: LGN, lateral geniculate nucleus; TEO posterior inferior temporal cortex; TE (or IT) inferior temporal cortex.

standard Gabor transform by the real and imaginary part, i.e. by the phase of the complex function representing it, whereas in a biological context this can be done by combining pairs of neurons with even and odd receptive fields. This design is supported by the experimental work of Pollen and Ronner (1981), who found simple cells in quadrature-phase pairs. Even more, Daugman (1988) proposed that an ensemble of simple cells is best modelled as a family of 2*D*-Gabor wavelets sampling the frequency domain in a log-polar manner as a function of eccentricity. Experimental neurophysiological evidence constrains the relation between the free parameters that define a 2*D*-Gabor receptive field (De Valois & De Valois, 1988). There are three constraints fixing the relation between the width, height, orientation, and spatial frequency (Lee, 1996). The first constraint posits that the aspect ratio of the elliptical Gaussian envelope is 2:1. The second constraint postulates that the plane wave tends to have its propagating direction along the short axis of the elliptical Gaussian. The third constraint assumes that the half-amplitude bandwidth of the frequency response is about 1–1.5 octaves along the optimal orientation. Further, we assume that the mean is zero in order to have an admissible wavelet basis (Lee, 1996). The neuronal pools in our $V1$ module complex cells are modelled here by the power modulus of a 2*D*-Gabor function sensitive to a particular location, orientation, symmetry, and spatial frequency according to the constraints described above. The $V1$ module contains $N_{V1} \times N_{V1}$ hypercolumns, covering a $N \times N$ pixel scene. Each hypercolumn contains L orientation columns of complex cells with K octave levels corresponding to different spatial frequencies. The cortical magnification factor is explicitly modelled by introducing more high spatial resolution neurons in a hypercolumn the nearer this hypercolumn is to the fovea. The density of the fine spatial resolution neurons across the visual field decreases in the model according to a Gaussian function centered on the fovea. In other words, in the periphery far from the fovea only coarse spatial resolution $V1$ pools are in the respective hypercolumn, whereas in regions near to the fovea, the $V1$ hypercolumns include also high spatial resolution input neurons.

The modules $V2$, $V4$ and IT consist also of C-dimensional columns of neuronal pools (i.e., each column contains C pools) distributed in a topographical lattice with $N_{V2} \times N_{V2}$, $N_{V4} \times N_{V4}$ and $N_{\mathrm{IT}} \times N_{\mathrm{IT}}$ neurons, respectively. The connectivity between modules $V1$–$V2$, $V2$–$V4$ and $V4$–IT is intended to mimic the convergent forward connectivity of the cerebral cortex. This connectivity helps to implement the gradually increasing receptive field size as one proceeds up the cortical hierarchy, and the formation of neurons that respond to combinations of inputs with features in a defined spatial configuration (Elliffe et al., 2002; Rolls, 1992; Wallis & Rolls, 1997). The connections to neuronal pools in a column in an upper module are limited to neuronal

Table 1
Network dimensions

	Dimensions	Radius
IT	1×1×2	1
$V4$	4×4×2	1
$V2$	16×16×2	2
$V1$	32×32×16	16
Retina	256×256	–

pools in a column in the immediately connected lower module that are within a certain radius around the focal point of connection (see Fig. 1). We denote these radii at each level by R_{V1}, R_{V2} and R_{V4}. This connectivity is reciprocated by the backprojections.

Table 1 shows the dimensions utilized in the present implementation. We used $N = 128$, $N_{V1} = 128$, $K = 8$, $L = 2$, $N_{V2} = 16$, $N_{V4} = 4$, $N_{\mathrm{IT}} = 1$, $R_{V1} = 16$, $R_{V2} = 2$, $R_{V4} = 1$, $R_{\mathrm{IT}} = 1$, $C = 2$.

The dorsal stream includes a PP module which receives connections from $V1$ and $V2$, and which has reciprocal backprojections (see Fig. 1). This causes the effective resolution of PP neurons to be coarser than the highest resolution $V1$ neurons. An external top–down bias to the PP module, coming from a spatial short-term memory and denoted as prefrontal cortex area d46 in the model, generates a spatial attentional component. The backprojections from PP influence the activity in the $V2$ and $V1$ modules, and thus can indirectly influence activity in the ventral stream modules. A lattice of $N_{\mathrm{PP}} \times N_{\mathrm{PP}}$ nodes provides topological organization in module PP. Each node on the lattice corresponds to the spatial position of each pixel in the input image (i.e., $N_{\mathrm{PP}} = N$). Each of these assemblies monitors the activities from columns in $V1$ and $V2$ via a Gaussian weighting function that relates topologically homologous locations. Local competition in PP is implemented via local lateral inhibitory connections between a neuron and its neighboring neurons weighted with a Gaussian-like factor. The width of the Gaussian decay is denoted σ_{PP}.

The system operates in two different modes, the learning mode and the recognition mode. During the learning mode the synaptic connections between $V1$–$V2$, $V2$–$V4$ and $V4$–IT are trained by means of an associative (Hebb-like) trace learning rule during a number of presentations of a given object as it is shifted to neighboring positions in the visual field (Wallis & Rolls, 1997). This learning rule 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 when it is presented in different locations, as described by Földiák (1991), Rolls (1992) and Wallis and Rolls (1997). The learning rule incorporates a decaying trace of previous cell activity and is henceforth referred to simply as the 'trace' learning rule (see Eq. (A.25)). This learning paradigm is intended in principle to enable learning of any of the transforms tolerated by inferior temporal cortex neurons (Rolls, 1992; Wallis & Rolls, 1997). 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 each 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 (Wallis & Rolls, 1997). Various biological bases for this temporal trace have been advanced. One is the continuing firing of neurons for as long as 100–400 ms observed after presentations of stimuli for 16 ms (Rolls & Tovee, 1994), which could provide a time window within which to associate subsequent images. Maintained activity may potentially be implemented by recurrent connections between cortical areas (Rolls & Deco, 2002). A second is the binding period of glutamate in the NMDA channels, which may last for 100 or more ms, and may implement a trace rule by producing a narrow time window over which the *average* activity at each presynaptic site affects learning (Földiák, 1992; Rolls, 1992).

During learning, the backprojections are disabled, partly for simplicity, partly to assist rapid learning, partly because before any training has occurred, backprojections would introduce only noise into the system, and partly in recognition of the fact that during any new learning, new information being fed into the system, rather than what is already stored, should dominate the activity of cortical areas (Rolls & Deco, 2002; Rolls & Treves, 1998). Consistent with these points, backprojections in the cerebral cortex make synapses on the apical parts of the dendrites of pyramidal cells, making it likely that when forward inputs which make synapses closer to the cell body are active, then the backprojection effects are shunted (Rolls & Deco, 2002; Rolls & Treves, 1998). After the forward connections have been trained using the trace rule, the values of the backprojection synapses are set to be the same values as the forward connections between any two neurons, scaled by a single scaling factor. Further details are provided in Appendix A.

During the recognition mode the forward and the backprojection pathways operate as a dynamical system that is implemented by a mean-field set of differential equations, as described in Appendix A. External top–down bias from the prefrontal cortex can be introduced in order to model the effects of object or spatial attention. The full mathematical description of the model is given in Appendix A. The model has the ability to simulate covert visual search, that is search without eye movements in which given an object attentional bias applied to the IT module the network settles in the PP module at the correct spatial location; and given a spatial attentional bias cue in the PP module the network identifies the correct object in the IT module (see Rolls & Deco, 2002). The model also can simulate invariant object recognition, in the way described previously (Elliffe et al., 2002; Rolls & Deco, 2002; Rolls & Stringer, 2001; Wallis & Rolls, 1997). However, in this paper we describe new issues that the combined model can address, as described next.

3. Operation of the model: simulations of fMRI and single-cell data

We describe in the following sections simulations performed with the model to test the model against experimental results and to provide an explanation of fMRI and single-cell findings. The experiments show how spatial and object-based attentional inputs applied at the top of the spatial or object processing stream produce attentional effects throughout both processing streams by virtue of the feedback pathways and the interactions that occur through $V1$ and $V2$, and how competition though implemented locally gradually has a more global character as one proceeds up the ventral stream hierarchy. First, we explain the gradually increasing magnitude of the attentional modulation from earlier visual areas ($V1$, $V2$) to higher ventral stream areas ($V4$, IT) as found in fMRI experiments. Second, we explain the variation of the effective size of receptive fields of IT neurons in natural cluttered scenes.

3.1. fMRI data: gradually increasing attentional and more global lateral inhibitory modulation along the ventral stream

Functional magnetic resonance imaging (fMRI) studies show that when multiple stimuli are present simultaneously in the visual field, their cortical representations within the object recognition pathway interact in a competitive, suppressive fashion (Kastner et al., 1998, 1999). Directing attention to one of the stimuli can counteract the suppressive influence of nearby stimuli. The model we describe here is able to simulate and account for these results. In the first experimental condition the authors (Kastner et al., 1998, 1999) showed the presence of suppressive interactions among stimuli presented simultaneously (SIM) within the visual field in the absence of directed attention (UNATT). The comparison condition was sequential (SEQ) presentation. (In the SEQ condition, each of the complex image stimuli was shown separately in one of four locations. In the SIM condition, the stimuli appeared together in all four locations. The presentation time was 250 ms, followed by a blank period of 750 ms, on average, in each location. A 15 s block design was used. An attentional modulation index (AMI) was defined as $\mathrm{AMI} = \frac{[\mathrm{ATT}-\mathrm{UNATT}]}{\mathrm{ATT}}$ where ATT = the fMRI response in the attended condition. The AMI was computed separately for the sequential and simultaneous conditions.) In a second experimental condition they showed that spatially directed attention increased the fMRI signal more strongly for simultaneously presented stimuli than for sequentially presented stimuli. Thus, the suppressive interactions were partially cancelled out by attention. This effect was indicated by a larger increase of the AMI_{SIM} in comparison to AMI_{SEQ} caused by attention. The results further showed that attention had a greater effect (the AMI was higher) for higher (IT, $V4$ and $V2$) than for earlier ($V1$) visual areas, as shown in Fig. 3a. In a third experimental condition the effects of attention were investigated in the absence of the visual stimuli.

The dynamical evolution of neural activity at the level of what occurs in different cortical areas as measured by fMRI signals can be simulated in the framework of the present model by integrating the activity of neuronal pools in a given simulated cortical area over space and time. The temporal integration was set so that it has the temporal resolution of fMRI experiments. In this section we describe simulations of the fMRI signals from $V1$, $V2$, $V4$, and IT under the experimental conditions defined by Kastner et al. (1999).

In the simulations, as in the experiments, images were presented in four nearby locations in the upper right quadrant. In all the simulations described in this paper, the background image was of a complex natural scene (similar to that in Fig. 5.18 of Rolls & Deco, 2002); and the stimuli consisted of letters of the alphabet made up of smaller letters (similar to those in Fig. 10.3 of Rolls & Deco, 2002). The neurodynamics (as defined by the differential equations that specify the interactions between neurons and between modules provided in Appendix A) were solved using the Euler method with a $\mathrm{d}t = 1$ ms (note that this means that 1000 iterations represents 1 s of time in the fMRI measurements). Two attentional conditions were simulated: an unattended condition, during which no external top–down bias from prefrontal areas was present (i.e., $I_{ij}^{\mathrm{PP},A}$ was set to zero everywhere), and an attended condition which started (in an expectation period) 10 s before the onset of visual presentations and continued during the subsequent 10 s

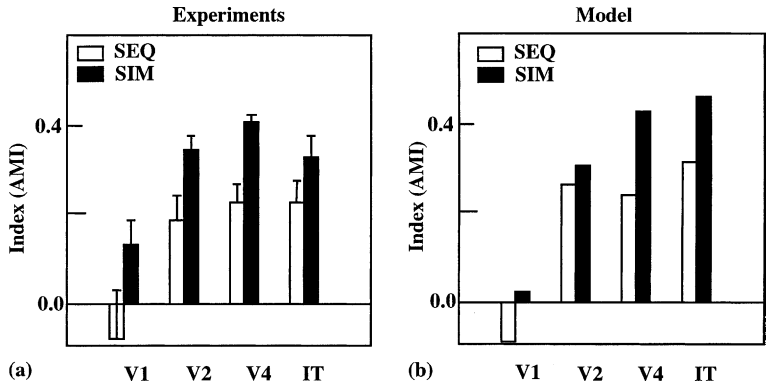

Fig. 3. Attentional modulation index (AMI) in visual cortex with sequentially and simultaneously presented stimuli. The AMI normalizes the attentional effects to the activity evoked in the corresponding attended condition. The effect of attention on a particular visual area is evident as the difference between the AMI observed by sequentially and simultaneously presented stimuli. Higher ventral stream visual cortical areas show more attentional modulation.

block. This attended condition corresponded to providing a spatial cue applied to the PP (parietal) module, and testing for identification of the correct object in the IT (inferior temporal cortex) module. The spatial attentional cue was implemented by setting $I_{ij}^{PP,A}$ equal to 0.07 for the locations associated with the lowest left stimulus and zero elsewhere. Fig. 3b shows the results of our simulations. The simulations show a gradually increasing magnitude of attentional modulation from earlier visual areas ($V1$, $V2$) to higher ventral stream visual areas ($V4$, IT), which is similar to that found in the experiments of Kastner et al. (1999). This attentional modulation is location-specific, and its effects are mediated by the PP attentional biassing input having an effect via the backprojections in $V2$ and $V1$, from which the effect is fed up the ventral stream in the forward direction to reach the IT module. The gradually increasing influence of attentional modulation from early visual cortical areas to higher ventral stream areas is a consequence of the gradually increasing global character of the competition between objects and/or parts of objects as one ascends through the ventral visual system, and the locally implemented lateral inhibition becomes effectively more global due to the convergence in the forward direction in the hierarchical pyramidal architecture of the ventral stream illustrated in Fig. 2. For clarification, the competition in the IT module is global, in that just two neurons reflect the two possible objects in the scene, but in a more biologically realistic implementation with distributed representations the global competition would occur because neurons in the distributed representation would be intermingled in a non-topologically based representation, and would thus interact.

The simulation data describe quite well the qualitative behaviour found in the experiments. The quantitative differences found between the simulated and empirical data are due to the numerical values of the parameters used in the model. Closer results could be obtained by adjusting these parameters. However, our intention is primarily to provide a qualitative analysis of the underlying processes that give rise to the experimental fMRI results.

3.2. The receptive field size of IT neurons to stimuli presented in complex natural backgrounds

Translation invariance is an important property of visual processing in object recognition. Inferior temporal visual cortex neurons that respond to specific objects or faces show considerable translation invariance, not only under anesthesia (Gross, Desimone, Albright, & Schwartz, 1985), but also in the awake behaving primate (Tovee et al., 1994). These neurons have large receptive fields when a single object is presented in a blank background. In most cases the responses of the neurons were little affected by which part of the face was fixated, and the neurons responded (with a greater than half-maximal response) even when the monkey fixated 2°–5° beyond the edge of a face that subtended 8°–17° at the retina. Moreover, the stimulus selectivity between faces was maintained this far eccentrically within the receptive field.

If more than one object is present on the retina, it was found that the mean firing rate across a sample of anterior inferior temporal cortex cells to a fixated effective face with a non-effective face in the parafovea (centred 8.5° from the fovea) was 34 spikes/s. On the other hand, the average response to a fixated non-effective face with an effective face in the periphery was 22 spikes/s (Rolls & Tovee, 1995). Thus these cells gave a reliable output about which stimulus is actually present at the fovea, in that their response was larger to a fixated effective face than to a fixated non-effective face, even when there were other parafoveal stimuli effective for the neuron. Thus the neurons provide information biassed

towards what is present at the fovea, and not equally about what is present anywhere in the visual field.

Recently, Rolls, Aggelopoulos, and Zheng (2004) (see also Rolls & Deco, 2002) investigated how information is passed from the inferior temporal cortex (IT) to other brain regions to enable stimuli presented in complex natural scenes to be selected for action. They analyzed the responses of single and simultaneously recorded IT neurons to stimuli presented in complex natural backgrounds. In one situation, a visual fixation task was performed in which the monkey fixated at different distances from the effective stimulus. In another situation 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 a drop of saline. In both situations 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 sometimes a little reduced when they were presented in natural scenes, though 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 field sizes of the neurons became smaller in natural scenes. This effect is shown in Fig. 4 (after Rolls & Deco, 2002). Rolls et al. (2004) showed that the receptive fields were large (78°) with a single stimulus in a blank background, and were greatly reduced in size (to 22°) when presented in a complex natural scene. They also showed that the receptive fields were smaller in complex scenes if the object was not the target of attention than when it was being searched for, although the effect of attention was much smaller in a complex natural scene than it was when tested as has been usual in studies of attention in the past with objects shown on a blank screen. In the most recent experiments it has been found with smaller objects that the receptive field can shrink to approximately the size of an object (Rolls et al., 2003).

Rolls et al. (2004) and Rolls and Deco (2002) 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.

In this section, we develop a computational hypothesis that can account for these effects in the theoretical framework of our neurodynamical model. We trained the network described above with two objects, and used the trace learning rule 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 (a forest scene from the 'still images collections' at www.visionscience.com).

Fig. 5 shows the average firing activity of the inferior temporal cortex neuron specific for the test object as a function of the position of the object on the retina rel-

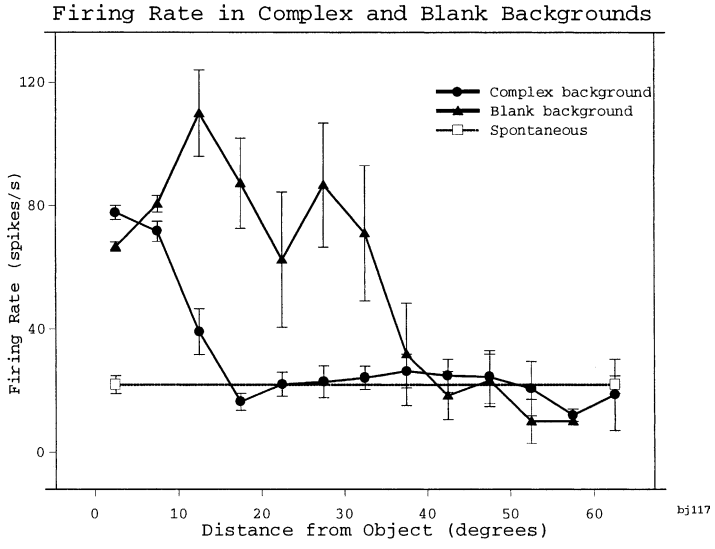

Fig. 4. Firing of a temporal cortex cell to an effective stimulus presented either in a blank background or in a natural scene, as a function of the angle in degrees at which the monkey was fixating away from the effective stimulus. The task was to search for and touch the stimulus (after Rolls et al., 2003).

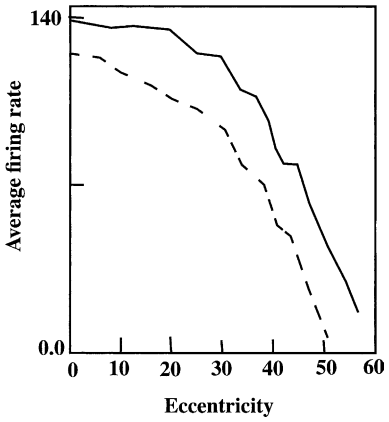

Fig. 5. Average firing activity of an inferior temporal cortex neuron as a function of eccentricity from the fovea, in the simulation. When the object was in a blank background (solid line), large receptive fields are observed because of the translation invariance of inferior temporal neurons. The decay is mainly due to the magnification factor implemented in $V1$. When the object was presented in a complex cluttered natural background (dashed line), the effective size of the receptive field of the same inferior temporal neuron was reduced because of competitive effect between the object features and the background features within each layer of the ventral stream.

ative to the fovea (eccentricity). In both cases relatively large receptive fields are observed, because of the translation invariance obtained with the trace learning rule and the competition mechanisms implemented within each layer of the ventral stream. (The receptive field size is defined as the width of the receptive field at the point where there is a half-maximal response.) However, when the object was in a blank background, larger receptive fields were observed. The decrease in neuronal response as a function of distance from the fovea is mainly due to the effect of the magnification factor implemented in $V1$. On the other hand, when the object was in a complex cluttered background, the effective size of the receptive field of the same inferior temporal cortex neuron shrinks because of competitive effects between the object features and the background features in each layer of the ventral stream. In particular, the global character of the competition expressed in the inferior temporal cortex module (due to the large receptive fields and the local character of the inhibition, in our simulations, between the two object specific pools) is the main cause of the reduction of the receptive fields in the complex scene.

We also studied the influence of object-based attentional top–down bias on the effective size of an inferior temporal cortex neuron for the case of an object in a blank or a cluttered background. To do this, we 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. Fig. 6 shows the results. We plot the average firing activity normalized to the maximum value to compare the neuronal activity as a function of the eccentricity. When no attentional object bias is introduced (a), a shrinkage of the receptive field size is observed. When attentional object bias is introduced (b), the shrinkage of the receptive field due to the complex background is slightly reduced. Rolls et al. (2004) also found that in natural scenes, the effect of object-based attention on the response properties of inferior temporal cortex neurons was relatively small. They found only a small difference in the receptive field size or firing rate in the complex background when the effective stimulus was selected for action, vs. when it was not. In the framework of our model, the reduction of the shrinkage of the receptive field is due to the biasing of the competition in

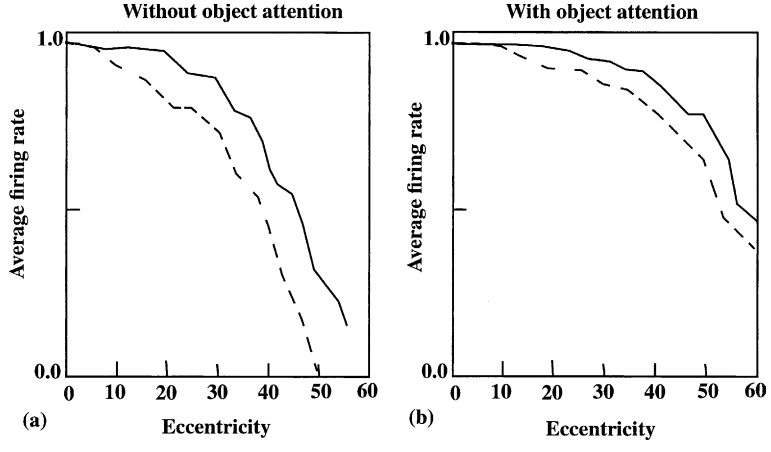

Fig. 6. Influence of object-based attentional top–down bias from prefrontal area 46v on the effective size of an inferior temporal cortex neuron for the case of an object in a blank (solid line) or a cluttered (dashed line) background. The average firing activity was normalized to the maximum value in order to compare the neuronal activity as a function of the eccentricity. When no attentional object bias was introduced (a), a reduction of the receptive field was observed. When attentional object bias was introduced (b), the reduction of the receptive field size due to the complex background was slightly reduced.

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 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.

3.3. The effective receptive field size of IT neurons in scenes with a blank background and a second distracting object

In this section, we analyze a set of experiments where we placed two objects, a target and a distractor, in a blank background, in order to study the influence of a single distractor object on the receptive field of an IT neuron specific for the target object. The target object, to which the monitored IT neuron specifically responds, is placed on one side of the fovea at different eccentricities in order to follow the decay of the average firing rate of the corresponding IT neuron as a function of the distance from the fovea. The distractor object is placed on the other side of the fovea at a fixed location $D°$ from the fovea. Fig. 7 shows the results of this simulation. The average firing rate of an IT neuron specific for the target object is plotted as a function of the position of the target relative to the fovea. The different curves correspond to different locations $D = 15$, 25, 30, 40 and 45 of the distractor object on the other side of the fovea. The corresponding D-value is shown for each curve. The single target object case (i.e., without a distractor) is also plotted (upper curve 'one object'). The size of the receptive field for the target object (shown by the decrease in the respective plot) decreases with decreasing distance of the distractor object from the fovea. This can be interpreted mainly as an effect of the global character of the local implemented competition at the inferior temporal cortex layer due to its large receptive fields and intermingling of neurons. When a second distracting object is near the fovea, due to the large magnification factor, one will have much more activation in the upper layers, and particularly in the inferior temporal cortex layer, in neurons associated with features of the distractor object. Consequently, much more competition in all layers, and particularly in the inferior temporal cortex layer producing a global character to the competition, causes a stronger suppression of the firing activity of the target specific IT neuron. This is indicated by a rapid decay of the activation curve, i.e. in the size of the corresponding receptive field. Increasing the distance from the fovea of the location of the distractor object, again due to the Gaussian decay of the magnification factor, will again produce lower activity at all levels, and particularly at the inferior temporal cortex layer, which results in weak competition that will cause an increase in the translation invariance at the IT level (i.e., there will be an increase of the IT receptive field size corresponding to the target object as a function of the distance of the distractor object from the fovea).

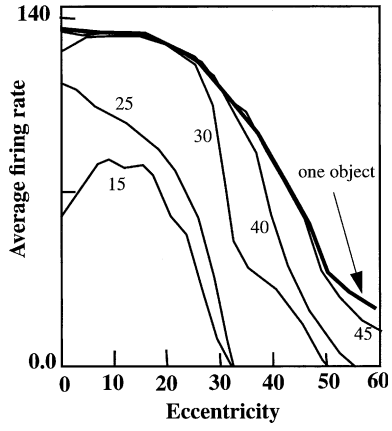

Fig. 7. The average firing rate of an IT neuron specific for the target object is plotted as a function of the position of the target relative to the fovea. The different curves correspond to different locations $D = 15$, 25, 30, 40 and 45 of the distractor object on the other side of the fovea. The corresponding D-value is attached at each curve. The single target object case (i.e., without a distractor) is also plotted (upper curve 'one object'). The size of the receptive field for the target object (curve decay) decreases with decreasing distance of the distractor object from the fovea.

3.4. Experimental predictions

We present in this section specific new simulation-based predictions indicating two different type of modulation of IT neuron receptive fields, namely, one due to local early layer competition, and the other associated with more global competition at higher layers in the ventral stream.

3.4.1. Asymmetric effective receptive field size of IT neurons in scenes with a blank background and a second distracting object

One specific prediction of the model can be tested by repeating the experiment with two objects, but now placing the second distracting object on the same side of the fovea where the target object is also located. With this variation, we tested the influence of competition effects at earlier layers, where the character of the competition is much more local, due to the much smaller receptive fields of neurons in the earlier layers. Fixing the distractor object at one location corresponding to the eccentricity D, and than testing the firing rate of an IT neuron associated with the target object when it is placed at different locations on the same side of the fovea and on the same line from the fovea where the distractor object is located (see Fig. 8a), we expect to observe maximal local competition

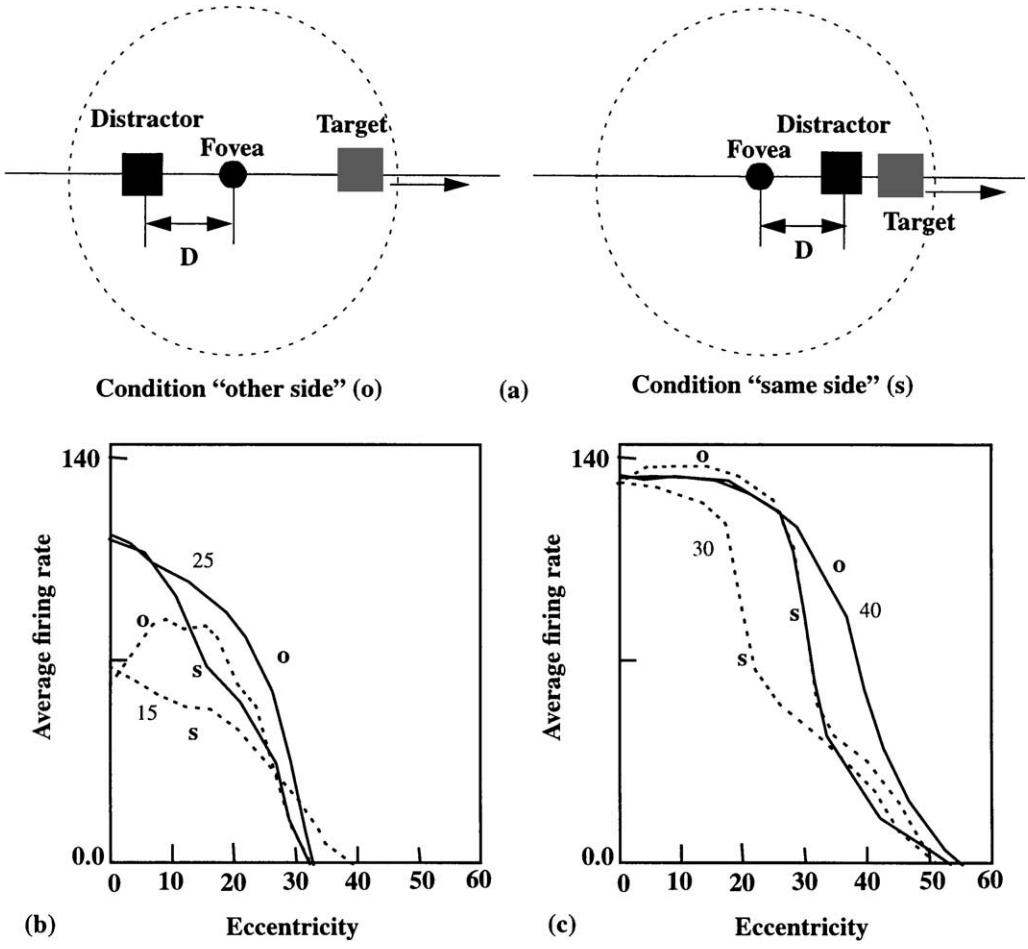

Fig. 8. (a) Diagrammatic description of condition 'same side' (s) and 'other side' (o) (see text). (b) and (c) For each D, two curves are plotted, one corresponding to the location of the distractor on the other side of the fovea (condition 'other side'), and the other corresponding to the location of the distractor on the same side of the fovea (condition 'same side'). The local activity reduction of the firing rate in the 'same side' condition with respect to the 'other side' condition was shifted with the shift of D, so that the maximum effect was always observed for a position of the target around D.

effects in the cases where both objects are near each other. Fig. 8b and c show the results of our simulations for different D. For each D (specified as a number near the curves), two curves are plotted, one corresponding to the location of the distractor on the other side of the fovea (condition 'other side' (o), as in Fig. 7, previous section), and the other corresponding to the location of the distractor on the same side of the fovea (condition 'same side' (s)). First, we remark that local competition effects at earlier layers (mainly $V1$) are present, as expected. The effect is again an increase of the competitive interaction, now occurring in earlier layers and only this particularly expressed when the two object are close to each other, so that the firing activity of the target neurons (in all layers and particularly in IT) are reduced. Second, this effect, local activity reduction of the firing rate in the 'same side' condition with respect to the 'other side' condition, is shifted with the shift of D, so that the maximum effect is always observed for a position of the target around D.

3.4.2. Local inhibitory effects on the effective receptive field size of IT neurons in a scene with a natural background and a surrounding grey circle around a small object

To examine the effects of local neuronal competition expressed in early layers of the visual processing, we simulated a condition in which the target object is presented in a natural complex cluttered background but is surrounded locally by a blank ring. In this way, we suppress the local effect of competition at earlier layers (mainly $V1$), because the ring removes the effect of the competing local activity of the $V1$ neurons in the neighborhood of the target object. Fig. 9 shows the results of this simulation. The simulation shows an increase of the receptive field size of an IT neuron responding to the target relative to that measured without the local blank ring in the complex full background. The simulation also shows that the receptive field size predicted in the 'blank ring' complex background condition is smaller that that corresponding to

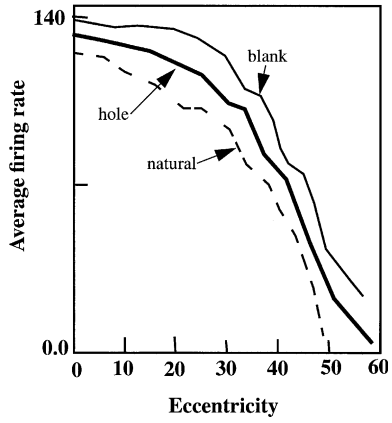

Fig. 9. Experimental prediction of receptive field size modulation when a single target object is surrounded locally by a blank ring. We predict an increase of the receptive field size of the target IT neuron relative to that measured with a complex background. On the other hand the receptive field predicted in the 'blank ring' (hole) is smaller than that corresponding to the 'blank background' condition.

the 'blank background' condition. This effect is due to the global competition effects that are effectively implemented in the higher ventral stream layers ($V4$, IT), where neurons are activated by the natural background features that are non-local to the target stimulus. (We note that in a neurophysiological test of this prediction, in order to prevent the grey circle being interpreted as an object in its own right, it might be appropriate to use instead of a grey circle, a region with low contrast.)

3.5. The interaction between the ventral and the dorsal system: visual search and object recognition

There are two possibilities for running the system. First, in *visual spatial search mode*, the spatial location of an object can be found in a scene by biasing the system with an external top–down (back projection) attentional component (from e.g. prefrontal area v46) to the TE (object) module. This drives the competition in TE in favour of the pool associated with the specific object to be searched for. Then, the intermodular backprojection attentional modulation TE–$V4$–$V2$–$V1$ will enhance the activity of the pools in $V4$, $V2$ and $V1$ associated with the component features of the specific object to be searched for. This modulation will add to the visual input being received by $V2$ and $V1$, resulting in greater local activity where the features in the topologically organized visual input feature representations match the feature representations being facilitated by the top–down attentional backprojections. Finally, the enhanced firing in a particular part of $V2$ and $V1$ will lead to increased activity in the forward pathway from $V1$ and $V2$ to PP, resulting in increased firing in the PP module in the location that corresponds to where the object being searched for is located. In this way, the architecture automatically finds the location of the object being searched for, and the location found is made explicit by which neurons in the spatially organized PP module are firing. Second, in *visual object identification mode*, the PP module receives a top–down (backprojection) input (from e.g. prefrontal area d46) which specifies the location at which to identify an object. The spatially biased PP module then drives by its backprojections the competition in the $V2$–$V1$ modules in favour of the pool associated with the specified location. This biasing effect in $V1$ and $V2$ will bias these modules to have a greater response for the specified location in space. The shape feature representations which happen to be present due to the visual input from the retina at that location in the $V1$ and $V2$ modules will therefore be enhanced, and the enhanced firing of these shape features will via the feedforward pathway $V1$–$V2$–$V4$–TE favour the TE object pool that contains the facilitated features, leading to recognition in TE of the object at the attentional location being specified in the PP module.

The operation of these two attentional modes is shown schematically in a simulation using real scenes in Fig. 10. We use as the target the picture of a monkey face placed in a natural background (the same background as that described and used above). For monitoring the performance and the dynamics of the network we plot the population *maximum* activity of the neuronal pools associated with the target and with the distracting rest of the world (background) at each point in time. Fig. 10 shows the results.

In the visual search task, i.e. when the system was looking for a particular object in a visual scene, the system functioned in an *object attention mode* as shown in Fig. 10a. Object attention was created by introducing a top–down bias to a particular cell pool in the TE module corresponding to the target. This ventral TE module pool backprojected the expected shape activity patterns over all spatial positions in the early V visual module through the top–down feedback connections, through $V4$ and $V2$. When the image containing the target object was presented, the early $V1$ and $V2$ visual modules whose activities were closest to the top–down 'template' became more excited because of the interactive activation or resonance between the forward visual inputs and the backprojected activity from the TE–$V4$ modules. Over time, these $V1$ hypercolumns with neuronal activities best matching the features in the encoded object dominated over all the other hypercolumns, resulting in a spatially localized response peak in the early $V1$ visual module. Meanwhile, the dorsal PP module was not idle but actively participated by having all its pools engaged in the competitive process to narrow down the location of the target. The simultaneous competition in the spatial domain and in the object domain in the two extrastriate modules as mediated by their reciprocal connections with the early $V1$ module

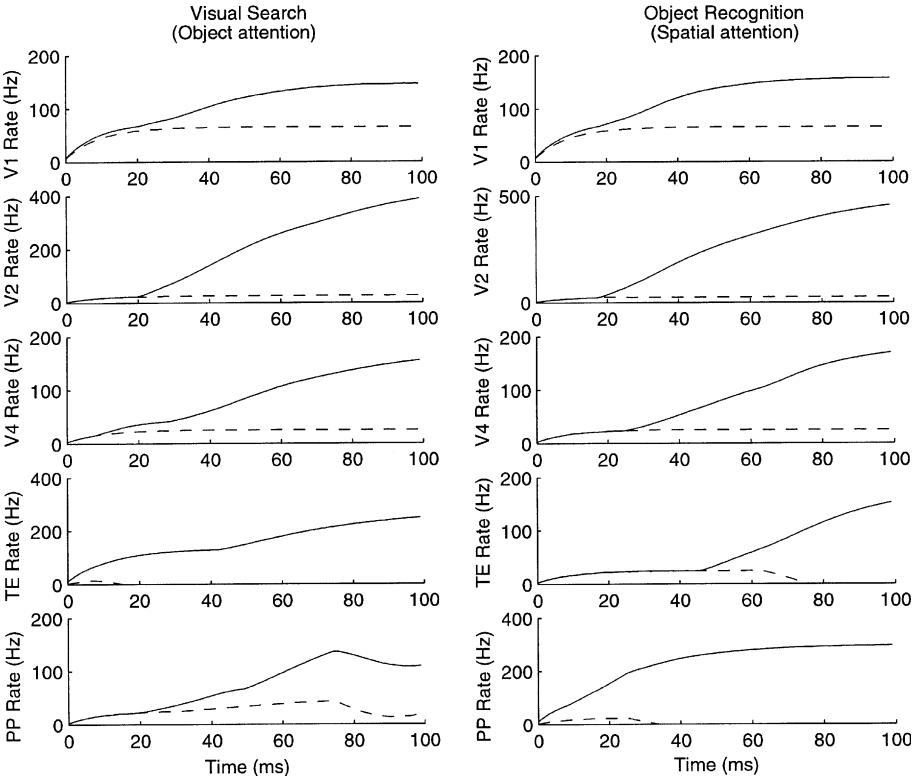

Fig. 10. Neuronal activity in all modules during: (a) visual search in object attention mode when the aim is to find the location of the object receiving attentional bias and (b) in spatial attention mode when the aim is to perform object recognition for the object at the spatial location receiving attentional bias. The visual stimuli consisted of a real scene with a target monkey face on a natural background. The maximum population activity of the neuronal pools corresponding to the identity or location of the target in the scene in all modules was compared against the maximum activity in pools coding any other locations or background objects.

finally resulted in a localized peak of activation in the spatially mapped dorsal stream PP module, with a corresponding peak of activity for the object mapped in the ventral stream IT module, and corresponding activity in the early $V1$ module. This corresponds to finding the object's location in the image in a visual search task, or linking 'where with what', or computing 'where' from 'what'. The temporal evolution of the population maximum activity shows the polarization of responses that started in the ventral stream TE module, and then backpropagated to the other ventral modules and finally to the PP module, where the object is localized.

In the object recognition task, the system functioned in a spatial attention mode as shown in Fig. 10b. Spatial attention was initiated by introducing a bias to a cell pool coding for a particular location. When the image was presented, the spatial bias interacted with the visual image inputs provided by the $V1$–$V2$ modules. If the hypercolumns in the $V1$ module 'designated' for spatial attention contained sufficiently strong neural activities, the activities in these hypercolumns would interact synergistically with the biased PP cell pool, so that over time that pool would dominate over all the other pools in the PP module, and the activities in the $V1$–$V2$ modules would be enhanced by the top–down bias from the PP module pool. This enhancement of neural activity highlighted the information in the attended location, effectively gating information in that area of the $V1$–$V2$ modules to the $V4$–TE modules for recognition (and in this way performing a type of shift invariance by using an attentional spatial modulation of early visual cortical processing). When the highlighted image patch contained one of the trained object classes, the activity of the TE cell pools started to polarize, resulting in only one cell pool surviving the competition. The winner indicated the object class being recognized, identifying 'what was where', or 'what from where'. The actual simulation results just outlined are shown in Fig. 10b, which shows the neuronal activities in all the modules during object recognition in the spatial attention mode. The temporal evolution of the population maximum activity shows the polarization of responses that started in the dorsal stream PP module, and then backpropagated to the other early ventral modules $V1$–$V2$ and finally to the $V4$ and TE modules, where the object is recognized.

Fig. 10 shows clearly the different temporal evolution of the activity in the different cortical modules under both conditions. We emphasize that the delays are not due to delays in the transmission between brain areas

(which we set to 0 in this paper for simplicity), but reflect instead the time it takes the distributed attractors in each module to emerge and to influence each other under both conditions.

In the case of visual search, the information processing is parallel, but the dynamics show a clear latency increase from TE through $V4$–$V2$–$V1$ to PP. On the other hand, during object recognition, the dynamics show a clear latency increase from PP through $V1$–$V2$–$V4$ to TE.

4. Discussion

In this paper, we extend and combine our previous computational neuroscience-based models for invariant visual object recognition and attention in order to consider the feedback biasing effects of top–down attentional mechanisms on a hierarchically organized set of visual areas with a pyramidal architecture, with convergent connectivity, and with local intra-area competition. In particular, the analysis of the interaction between space-based and object-based attentional top–down feedback, and the local and gradually increasing global character of laterally competing neurons in a pyramidal network for hierarchical feature integration, is the main goal of this paper. In order to implement this we performed here the fusion of two complementary models, namely the feedforward feature hierarchy network VisNet (Elliffe et al., 2002; Rolls & Deco, 2002; Rolls & Milward, 2000; Wallis & Rolls, 1997) and the multi-modular recurrent attentional model of Deco (2001), Deco and Lee (2002), Deco and Zihl (2001) and Rolls and Deco (2002).

We have shown that the model we describe does combine the ability to perform invariant image recognition, which is a property of hierarchical networks such as VisNet (Elliffe et al., 2002; Rolls & Deco, 2002; Rolls & Milward, 2000; Stringer & Rolls, 2000, 2002; Wallis & Rolls, 1997), with attentional phenomena such as object-cued and space-cued search which can be implemented in a system with backprojections and a dorsal as well as a ventral stream of cortical processing (Corchs & Deco, 2002; Deco & Lee, 2002; Deco & Zihl, 2001; Rolls & Deco, 2002). We emphasize that our modelling of object-based and space-based attention is an emergent effect of the dynamical interaction between different brain areas. It is based on the coupling between feedforward connections (e.g., from $V1$–$V2$, $V4$ to IT in the ventral stream and from $V1$–$V2$ to PP in the dorsal stream), and the re-entrant feedback connections (e.g., from IT–$V4$–$V2$ to $V1$ in the ventral stream and from PP to $V2$ and $V1$ in the dorsal stream). Even more, this re-entrant coupling results in the main modulatory effect of spatial or object attention being observed in $V1$ after a long latency, of around 120–200 ms (see detailed simulations in Deco & Lee, 2002). These simulation results are consistent with the experimental observations of Martinez et al. (2001) which have shown that attentional effects in $V1$ have a much longer latency (160–260 ms post stimulus onset) than those in extrastriate cortex (70–130 ms), suggesting that $V1$ activity may be modulated by delayed reentrant feedback from higher visual areas, as is implemented in our model. We note that other models of invariant object recognition (see e.g. review by Riesenhuber & Poggio, 2000) are feedforward models, and thus cannot address the issue of top–down processes and attention. We also note that the issue of feature binding can be dealt with in models of the type we describe by forming neurons at an early stage of processing that respond to combinations of feature in the correct spatial positions, and that if low order feature combinations are represented, and the natural statistics of images are taken into account, then the potential combinatorial explosion can be kept under control (Elliffe et al., 2002; Rolls & Deco, 2002).

The model accounts for the computational role of the locally implemented but gradually increasing global character of lateral inhibition and thus competition between neurons in the context of actual fMRI and single cell experiments. The model shows the same effects as found in fMRI experiments, namely a gradual increase in the magnitude of the attentional modulation from earlier visual areas ($V1$, $V2$) to high level ventral areas ($V4$, IT). The model allows us to understand how this occurs. The effect arises because the attentional modulation that is applied to the IT module can bias the system strongly here. The strong bias in IT arises because the attentional bias has to be applied throughout the IT module, so that the bias influences neurons that code for a particular object wherever they are in IT. The result is that the competition between the neurons representing different objects is effectively global in the IT module. On the other hand, in $V1$ the two stimuli are represented in different parts of a topologically organized map of visual space, and thus the local competition implemented by interneurons (neurons in the local inhibitory pools) will not operate to reduce the activation produced by one stimulus when the attended object is elsewhere in the visual field.

The same explanation can account for the relatively greater magnitude of the attentional modulation effects observed at the single cell level in IT than in $V1$, which is found in the model, and neurophysiologically in IT, at least when operating with a plain background. That is, the large receptive fields for an attended object vs. an unattended object in IT can be accounted for by strong effectively global competition effects in IT implemented by the widespread distribution of the attentional bias in IT. On the other hand, smaller neuronal competition effects are found in $V1$ because the different objects are represented in different parts of the topological map,

and so do not interact within $V1$ itself because of the locality of the lateral inhibition. The model is further able to make the predictions shown in Figs. 8 and 9 that because of the local lateral inhibition that operates within $V1$, stimuli close together in visual space will produce more mutual inhibition, and because of this effect that is contributed especially by the early visual cortical areas, neurons even in IT will show greater mutual interaction if they are close together in visual space.

The model accounts for the reduction of the receptive field size of neurons in a complex scene by both global and local inhibition. In particular, the effectively global inhibition between the object representations in IT reduces the firing rates for most stimuli, but because of the large magnification factor at the fovea, the object at the fovea produces the strongest representation in IT. In addition, the local inhibition effectively implemented in $V1$ produces further competitive effects from the background on the stimulus, as shown by the experiment in which a small plain annulus round a stimulus led to less suppression of the neuronal responses (in IT) to that stimulus.

A different computational model of the reduction in the receptive field size of inferior temporal cortex neurons when objects are presented in natural scenes was described by Trappenberg, Rolls, and Stringer (2002). In that model, the input to the inferior temporal visual cortex modelled as an attractor network using the cortico-cortical recurrent connections was weighted by the cortical magnification factor of the projection of the visual field onto the cortex. The model could be activated in a blank scene by an object distant from the fovea because the weak peripheral inputs could capture the attractor, but if a complex background was present, this produced strong activation from foveal inputs, and peripheral inputs from a test object could not capture the attractor. The model described in this paper also utilized the greater magnification factor of the fovea than the periphery, but instead the competition was implemented by local lateral competition, which became effectively global in the inferior temporal visual cortex because there were global interactions within IT (implemented in the model because there were few neurons in IT, but more realistically by intermingled neurons in a distributed non-topological organization in IT).

Although the timing of the interactions between the ventral visual stream, early modules, and the dorsal visual stream is not the subject of this paper as it has been treated elsewhere (Deco & Lee, 2002; Rolls & Deco, 2002), we do note that the timing of interactions between the modules in the model shows similar effects to those found neurophysiologically (Martinez et al., 2001). For example, with object-based attentional bias applied to the IT module at the top of the ventral visual stream, attentional effects are found first in this IT module, then in $V4$, then in the early modules $V1$ and $V2$, then in the MT module, and finally in the parietal module at the top of the dorsal visual stream where the activation represents the spatial position of the object cued in the IT module (Deco & Lee, 2002; Rolls & Deco, 2002). Conversely, if a spatial attentional cue is applied to the PP module, then attentional modulation occurs in the following order: MT, $V2$ and $V1$, $V4$, and finally IT (see Fig. 1) where the activation represents the object that was at the spatial location cued in PP. We note that in this model $V1$ tends to show the top–down attentional effects investigated in this paper and elsewhere (Deco & Lee, 2002; Rolls & Deco, 2002) after $V2$ when the timing is with respect to the onset of the attentional bias, as this is a property of the architecture shown in Fig. 1. Of course, if the attentional bias is already present before a visual stimulus is presented, then attentional effects will be evident in $V1$ before $V2$, as the latencies of the neuronal response to the image will tend to be shorter in $V1$, which again is a straightforward property understandable from Fig. 1. Another property of the dynamical system implemented in the model is that it can account for multiplicative (as contrasted with additive) effects of attention on visual processing (McAdams & Maunsell, 1999), due to interactions between the non-linearity of the activation function and the mutual inhibition between the neurons, as explained by Rolls and Deco (2002). In addition, comparisons of the two classes of model combined in the architecture described here with other models of invariant recognition *or* attention (Olshausen, Anderson, & Van Essen, 1993; Riesenhuber & Poggio, 2000; Salinas & Abbott, 1997; Usher & Niebur, 1996) are provided by Rolls and Deco (2002).

In conclusion, we have shown in this model that a feature hierarchical network can be combined with a dynamical model using backprojections to account for the increasingly global character of attention-based top–down modulation evident in the inferior temporal visual cortex compared to the earlier modules. The local mutual inhibition within a layer also enables predictions to be made from the model about the interactions between stimuli in different locations in a visual scene. Overall, the model offers a way for studying the dynamical interactions between a dorsal visual 'where' stream and a ventral visual 'what' stream in a context in which invariant visual object representations are being formed in a hierarchically organized ventral visual stream.

Acknowledgements

This research was supported by MRC Programme Grant PG9826105, by the MRC Interdisciplinary Research Centre for Cognitive Neuroscience, and by the Oxford McDonnell Centre for Cognitive Neuroscience.

GD acknowledges support through ICREA, the German Ministery for Research, BMBF Grant 01IBC01A, and through the European Union, grant IST-2001-38099.

Appendix A. Mathematical formulation of the model

A.1. Neurodynamical equations

A model of brain functions requires the choice of an appropriate theoretical framework, which permits the investigation and simulation of large-scale biologically realistic neural networks. Starting from individual spiking neurons one can derive a differential equation that describes the dynamical evolution of the averaged activity of a pool of extensively many equivalent neurons. Several areas of the brain contain groups of neurons that are organized in populations of units with similar properties. These groups for mean-field modelling purposes are usually called pools of neurons and are constituted by a large and similar population of identical spiking neurons that receive similar external inputs and are mutually coupled by synapses of similar strength. Assemblies of motor neurons (Kandel, Schwartz, & Jessel, 2000) and the columnar organization in the visual and somatosensory cortex (Hubel & Wiesel, 1962) are examples of these pools. Each single cell in a pool can be described by a spiking model. Due to the fact that for large-scale cortical modelling, neuronal pools form a relevant computational unit, we adopt a population code. We take the activity level of each pool of neurons as the relevant dependent variable rather than the spiking activity of individual neurons. It is possible to derive dynamical equations for neuronal pool activity levels by utilizing the mean-field approximation (Abbott, 1991; Amit & Tsodyks, 1991; Wilson & Cowan, 1972). According to this approximation, we categorize each cell assembly by means of its activity $A(t)$. A pool of excitatory neurons without external input can be described by the dynamics of the pool activity given by

$$\tau \frac{\partial A(t)}{\partial t} = -A(t) + qF(A(t)), \tag{A.1}$$

where the first term on the right-hand side is a decay term and the second term (scaled by the constant q) takes into account the excitatory stimulation between the neurons in the pool. In the previous equation, the non-linearity

$$F(x) = \frac{1}{T_r - \tau \log\left(1 - \frac{1}{\tau x}\right)}, \tag{A.2}$$

is the response function (transforming current into discharge rate) for a spiking neuron with deterministic input, membrane time constant τ, and absolute refractory time T_r. Eq. (A.1) was derived by Gerstner (2000) assuming adiabatic conditions (i.e., the activity changes only slowly compared with the typical interval length) (see further Rolls & Deco, 2002, pp. 218–224).

We now present a formal description of the model. We consider a pixelized grey-scale image given by a $N \times N$ matrix $\Gamma_{ij}^{\mathrm{Orig}}$. The subindices ij denote the spatial position of the pixel. Each pixel value is given a grey level brightness value coded in a scale between 0 (black) and 255 (white). The first step in the preprocessing consists of removing the DC component of the image (i.e., the mean value of the grey-scale intensity of the pixels). (The equivalent in the brain is the low-pass filtering performed by the retinal ganglion cells and lateral geniculate cells. The visual representation in the LGN is essentially a contrast invariant pixel representation of the image, i.e. each neuron encodes the relative brightness value at one location in visual space referred to the mean value of the image brightness.) We denote this contrast-invariant LGN representation by the $N \times N$ matrix Γ_{ij} defined by the equation

$$\Gamma_{ij} = \Gamma_{ij}^{\mathrm{Orig}} - \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \Gamma_{ij}^{\mathrm{Orig}}. \tag{A.3}$$

Feedforward connections to a layer of $V1$ neurons perform the extraction of simple features like bars at different locations, orientations and sizes. Realistic receptive fields for $V1$ neurons that extract these simple features can be represented by $2D$-Gabor wavelets. Lee (1996) derived a family of discretized $2D$-Gabor wavelets that satisfy the wavelet theory and the neurophysiological constraints for simple cells mentioned above. They are given by an expression of the form

$$G_{pqkl}(x,y) = a^{-k} \Psi_{\Theta_l}(a^{-k}(x - 2p), a^{-k}(y - 2q)), \tag{A.4}$$

where

$$\Psi_{\Theta_l} = \Psi(x\cos(l\Theta_0) + y\sin(l\Theta_0), -x\sin(l\Theta_0) + y\cos(l\Theta_0)) \tag{A.5}$$

and the mother wavelet is given by

$$\Psi(x,y) = \frac{1}{\sqrt{2\pi}} e^{-(1/8)(4x^2+y^2)} [e^{i\kappa x} - e^{-\kappa^2/2}]. \tag{A.6}$$

In the above equations $\Theta_0 = \pi/L$ denotes the step size of each angular rotation; l the index of rotation corresponding to the preferred orientation $\Theta_l = l\pi/L$; k denotes the octave; and the indices pq the position of the receptive field centre at $c_x = p(N/N_{V1})$ and $c_y = q(N/N_{V1})$. In this form, the receptive fields at all levels cover the spatial domain in the same way, i.e. by always overlapping the receptive fields in the same fashion. In the model we use $a = 2$, $b = 1$ and $\kappa = \pi$ corresponding to a spatial frequency bandwidth of one octave.

The neurons in the pools in $V1$ have receptive fields performing a Gabor wavelet transform. Let us denote by I^{V1}_{pqkl} the sensory input activity to a pool A^{V1}_{pqkl} in $V1$ which is sensitive to a spatial frequency at octave k, to a preferred orientation defined by the rotation index l, and to stimuli at the centre location specified by the indices pq. The sensory input activity to a pool in $V1$ is therefore defined by the modulus of the complex valued convolution between the corresponding receptive fields and the image, i.e.

$$I^{VI}_{pqkl} = \|\langle G_{pqkl}, \Gamma\rangle\| = \left\| \sum_{i=1}^{N}\sum_{j=1}^{N} G_{pqkl}(i,j)\Gamma_{ij} \right\| \tag{A.7}$$

and is normalized to a maximal saturation value of 0.025.

Let us denote by A^{V2}_{pql}, A^{V4}_{pql} and A^{IT}_{pql} the activity of the l-pool in a column with receptive fields at the retinal center c_{pq} in the module $V2$, $V4$, and IT module, respectively. Similarly, let us denote with A^{PP}_{ij} the activity of a pool in the PP module corresponding to the location ij in the visual field. The neurodynamical equations that regulate the temporal evolution of the whole system are given by the following set of coupled differential equations:

$$\tau\frac{\partial A^{VI}_{pqkl}(t)}{\partial t} = -A^{V1}_{pqkl} + \alpha F(A^{V1}_{pqkl}(t)) - \beta I^{\mathrm{inh},V1}_{pq}(t) + I^{V1}_{pqkl}(t) + \gamma_b I^{V1-\mathrm{PP}}_{pq}(t) + \lambda_1 I^{V1-V2}_{pqkl}(t) + I_0 + \nu, \tag{A.8}$$

$$\tau\frac{\partial A^{V2}_{pql}(t)}{\partial t} = -A^{V2}_{pql} + \alpha F(A^{V2}_{pql}(t)) - \beta I^{\mathrm{inh,V2}}_{pq}(t) + \gamma_b I^{V2-\mathrm{PP}}_{pq}(t) + \lambda_2 I^{V2-V4}_{pql}(t) + I_0 + \nu, \tag{A.9}$$

$$\tau\frac{\partial A^{V4}_{pql}(t)}{\partial t} = -A^{V4}_{pql} + \alpha F(A^{V4}_{pql}(t)) - \beta I^{\mathrm{inh,V4}}_{pq}(t) + \lambda_3 I^{V4-IT}_{pql}(t) + I_0 + \nu, \tag{A.10}$$

$$\tau\frac{\partial A^{\mathrm{IT}}_{pql}(t)}{\partial t} = -A^{\mathrm{IT}}_{pql} + \alpha F(A^{\mathrm{IT}}_{pql}(t)) - \beta I^{\mathrm{inh,IT}}_{pq}(t) + I^{\mathrm{IT},A}_{l} + I_0 + \nu, \tag{A.11}$$

$$\tau\frac{\partial A^{\mathrm{PP}}_{ij}(t)}{\partial t} = -A^{\mathrm{PP}}_{ij} + \alpha F(A^{\mathrm{PP}}_{ij}(t)) - \beta I^{\mathrm{inh,PP}}_{ij}(t) + \gamma_f I^{\mathrm{PP}-V1}_{ij}(t) + \gamma_f I^{\mathrm{PP}-V2}_{ij}(t) + I^{\mathrm{PP},A}_{ij} + I_0 + \nu. \tag{A.12}$$

The spatial attentional biasing couplings $I^{V1-\mathrm{PP}}_{pq}$, $I^{V2-\mathrm{PP}}_{pq}$, $I^{\mathrm{PP}-V1}_{pq}$ and $I^{\mathrm{PP}-V2}_{ij}$ due to the intermodular 'where' connections with the pools in the parietal module PP are given by

$$I^{VE-\mathrm{PP}}_{pq} = \sum_{i,j} w^{VE-\mathrm{PP}}_{pqij} F(A^{\mathrm{PP}}_{ijn}(t)) \quad \text{for } VE = V1, V2, \tag{A.13}$$

$$I^{\mathrm{PP}-V1}_{ij} = \sum_{pqkl} w^{V1-\mathrm{PP}}_{pqij} F(A^{V1}_{pqkl}(t)) \tag{A.14}$$

and

$$I^{\mathrm{PP}-V2}_{ij} = \sum_{pql} w^{V2-\mathrm{PP}}_{pqij} F(A^{V2}_{pql}(t)). \tag{A.15}$$

The connections between pools in the ventral stream and pools in the PP module are specified such that topographically corresponding regions (in PP and in the ventral modules) are connected with maximal strength and the connections with neighboring regions decay with Gaussian modulation. The mutual (i.e., forward and back) connections between a pool A^{V1}_{pqkl} in VI, or A^{V2}_{pql} in $V2$ and a pool A^{PP}_{ij} in PP are therefore defined by

$$w^{VE-\mathrm{PP}}_{pqij} = \exp\left\{ -\frac{\mathrm{dist}(c_{pq}, c_{ij})^2}{2\sigma^2_{VE}} \right\}, \tag{A.16}$$

where c_{ab} corresponds to the $2D$-center in pixel retinal coordinates associated with the pool with space indices ab (in a ventral or PP module), and $\mathrm{dist}(c_1, c_2)$ is the Euclidean distance between centers c_1 and c_2. These connections mean that the VI pool A^{V1}_{pqkl} will have maximal amplitude when spatial attention is located at c_{pq} in the visual field, i.e. when the pool A^{PP}_{ij} in PP corresponding to $c_{ij} = c_{pq}$ is maximally activated. The same analysis hold for connections between pools in $V2$ and PP.

The feature based attentional top–down biasing terms I^{V1-V2}_{pqkl} due to the intermodular 'what' connections of pools between two immediate modules in the ventral stream are defined by

$$I^{V1-V2}_{pqkl} = \sum_{\tilde{p}=0}^{N_{V2}}\sum_{\tilde{q}=0}^{N_{V2}}\sum_{\tilde{l}=0}^{C} w^{V1-V2}_{pqkl\tilde{p}\tilde{q}\tilde{l}} F(A^{V2}_{\tilde{p}\tilde{q}\tilde{l}}(t)), \tag{A.17}$$

$$I^{V2-V4}_{pql} = \sum_{\tilde{p}=0}^{N_{V4}}\sum_{\tilde{q}=0}^{N_{V4}}\sum_{\tilde{l}=0}^{C} w^{V2-V4}_{pql\tilde{p}\tilde{q}\tilde{l}} F(A^{V4}_{\tilde{p}\tilde{q}\tilde{l}}(t)), \tag{A.18}$$

$$I^{V4-IT}_{pql} = \sum_{\tilde{p}=0}^{N_{\mathrm{IT}}}\sum_{\tilde{q}=0}^{N_{\mathrm{IT}}}\sum_{\tilde{l}=0}^{C} w^{V4-IT}_{pql\tilde{p}\tilde{q}\tilde{l}} F(A^{\mathrm{IT}}_{\tilde{p}\tilde{q}\tilde{l}}(t)), \tag{A.19}$$

where $w^{V1-V2}_{pqkl\tilde{p}\tilde{q}\tilde{l}}$, $w^{V2-V4}_{pql\tilde{p}\tilde{q}\tilde{l}}$ and $w^{V4-IT}_{pql\tilde{p}\tilde{q}\tilde{l}}$ are the connection strengths between the $V1$–$V2$, $V2$–$V4$ and $V4$–IT pools, respectively.

The local lateral inhibitory interactions $I^{\mathrm{inh},VE}_{pq}$ in modules in the ventral stream are given by

$$I_{pq}^{\mathrm{inh},VE} = \sum_{\tilde{p},\tilde{q},\tilde{k},\tilde{l}} w_{pq\tilde{p}\tilde{q}}^{\mathrm{inh},VE} F(A_{\tilde{p}\tilde{q}\tilde{k}\tilde{l}}^{VE}(t)) \quad \text{for } VE = V1, \tag{A.20}$$

$$I_{pq}^{\mathrm{inh},VE} = \sum_{\tilde{p},\tilde{q},\tilde{l}} w_{pq\tilde{p}\tilde{q}}^{\mathrm{inh},VE} F(A_{\tilde{p}\tilde{q}\tilde{l}}^{VE}(t)) \quad \text{for } VE = V2, V4, IT. \tag{A.21}$$

In the preceding two equations $w_{pq\tilde{p}\tilde{q}}^{\mathrm{inh},VE}$ expresses the lateral local connectivity between lateral nodes defined by

$$\begin{cases} w_{pq\tilde{p}\tilde{q}}^{\mathrm{inh},VE} = 1.0 & \text{if } p = \tilde{p} \text{ and } q = \tilde{q}, \\ w_{pq\tilde{p}\tilde{q}}^{\mathrm{inh},VE} = \delta \exp\left\{ -\frac{\mathrm{dist}(c_{pq}, c_{\tilde{p}\tilde{q}})^2}{\sigma_{VE}^2} \right\} & \text{else}, \end{cases} \tag{A.22}$$

where δ and σ control the amount and spread of lateral inhibition respectively. In our simulations, we use $\delta = 1$, $\sigma_{V1} = 16$, $\sigma_{V2} = 2$, $\sigma_{V4} = 1$ and $\sigma_{\mathrm{IT}} = 1$.

The local lateral inhibitory interactions $I_{ij}^{\mathrm{inh,PP}}$ in the PP module in the dorsal stream are given by

$$I_{ij}^{\mathrm{inh,PP}} = \sum_{\tilde{i},\tilde{j}} w_{ij\tilde{i}\tilde{j}}^{\mathrm{inh,PP}} F(A_{\tilde{i}\tilde{j}}^{\mathrm{PP}}(t)), \tag{A.23}$$

$w_{ij\tilde{i}\tilde{j}}^{\mathrm{inh,PP}}$ being the lateral local connections between lateral nodes defined by

$$\begin{cases} w_{ij\tilde{i}\tilde{j}}^{\mathrm{inh,PP}} = 1.0 & \text{if } i = \tilde{i} \text{ and } j = \tilde{j}, \\ w_{ij\tilde{i}\tilde{j}}^{\mathrm{inh,PP}} = \delta \exp\left\{ -\frac{\mathrm{dist}(c_{ij}, c_{\tilde{i}\tilde{j}})^2}{\sigma_{\mathrm{PP}}^2} \right\} & \text{else}. \end{cases} \tag{A.24}$$

In the particular case of PP, the center c_{ij} coincides with the location ij in the retinal input matrix.

The external attentional spatially specific top–down bias $I_{ij}^{\mathrm{PP},A}$ is assumed to come from prefrontal area 46d, whereas the external attentional object-specific top–down bias $I_l^{\mathrm{IT},A}$, is assumed to come from prefrontal area 46v. Both of them are associated with working memory.

In our simulations, we use $\alpha = 0.95$, $\beta = 0.8$, $\gamma_f = 1$ $\gamma_b = 0.4$, $\lambda_1 = \lambda_2 = \lambda_3 = 0.4$, $\delta = 0.1$, $I_0 = 0.025$, and the standard deviation of the additive noise $\nu, \sigma_\nu = 0.02$. The values of the external bias $I_{ij}^{\mathrm{PP},A}$ and $I_l^{\mathrm{IT},A}$ are equal to 0.07 for the pools that eventually receive an external positive bias and otherwise are equal to zero. The choice of these parameters is uncritical and is based on biological parameters.

In the case of object-based attention, the bias in IT $I_l^{\mathrm{IT},A}$ is set so that only the pool l corresponding to the object to be attended to receives a positive bias, while the external attentional location-specific bias in PP $I_{ij}^{\mathrm{PP},A}$ is set equal to zero everywhere. The external attentional bias $I_l^{\mathrm{IT},A}$ drives the competition in the IT module so that the pool corresponding to the attended object wins.

In the case of space-based attention, the bias in PP $I_{ij}^{\mathrm{PP},A}$ is set so that only the pool associated with the spatial location where the object to be identified is receives a positive bias, i.e. a spatial region will be 'illuminated'. The other external bias $I_l^{\mathrm{IT},A}$ is zero everywhere. In this case, the dynamics evolves such that in PP only the pool associated with the top–down biased spatial location will win. This fact drives the competition in $V1$, $V2$, $V4$, and IT such that only the pools corresponding to features of the stimulus at that location will win, biasing the dynamics in IT such that only the pool identifying the class of the features at that position will remain active indicating the category of the object at that predefined spatial location.

A.2. The trace learning rule

During a learning phase each object is learned. This is done by training the connections between modules in the ventral stream (i.e., $w_{pqkl\tilde{p}\tilde{q}\tilde{l}}^{V1-V2}, w_{pql\tilde{p}\tilde{q}\tilde{l}}^{V2-V4}$), and $w_{pql\tilde{p}\tilde{q}\tilde{l}}^{V4-IT}$, by a Hebbian-like trace learning rule.

We implemented here the original trace learning rule used in the simulations of Wallis & Rolls (1997) which is given by

$$\delta w_{ij} = \alpha \bar{y}_i^\tau x_j^\tau, \tag{A.25}$$

where x_j^τ is the jth input to the pool at time step τ, y_i is the output of the ith pool, and w_{ij} is the jth weight on the ith pool. The trace $\bar{y}_i^\tau$ is updated according to

$$\bar{y}_i^\tau = (1-\eta) y_i^\tau + \eta \bar{y}_i^{\tau-1}. \tag{A.26}$$

The parameter $\eta \in [0,1]$ controls the relative contributions to the trace $\bar{y}_i^\tau$ from the instantaneous firing rate y_i^τ at time step τ and the trace at the previous time step $\bar{y}_i^{\tau-1}$ where for $\eta = 0$ we have $\bar{y}_i^\tau = y_i^\tau$ and Eq. (A.25) becomes the standard Hebb rule

$$\delta w_j = \alpha y^\tau x_j^\tau. \tag{A.27}$$

References

Abbott, L. F. (1991). Realistic synaptic inputs for model neural networks. *Network, 2*, 245–258.

Amit, D. J., & Tsodyks, M. V. (1991). Quantitative study of attractor neural network retrieving at low spike rates I. Substrate—spikes, rates and neuronal gain. *Network, 2*, 259–273.

Andersen, R.A., Snyder, L.H., Bradley, D.C., & J., X. (1997). Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience 20*, 303–330.

Bushnell, C., Goldberg, M., & Robinson, D. (1981). Behavioral enhancement of visual responses in monkey cerebral cortex. I. Modulation in posterior parietal cortex related to selective visual attention. *Journal Neurophysiology, 46*, 755–772.

Chelazzi, L. (1998). Serial attention mechanisms in visual search: A critical look at the evidence. *Psychological Research, 62*, 195–219.

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.

Daugman, J. (1988). Complete discrete 2*D*-Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustic, Speech, and Signal Processing, 36*, 1169–1179.

De Valois, R. L., & De Valois, K. K. (1988). *Spatial Vision*. New York: Oxford University Press.

Deco, G. (2001). Biased competition mechanisms for visual attention. In S. Wermter, J. Austin, & D. Willshaw (Eds.), *Emergent Neural Computational Architectures Based on Neuroscience* (pp. 114–126). Heidelberg: Springer.

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., & Zihl, J. (2001). Top–down selective visual attention: A neurodynamical approach. *Visual Cognition, 8*, 119–140.

Desimone, R., Albright, T. D., Gross, C. G., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience, 4*, 2051–2062.

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

Duncan, J. (1980). The locus of interference in the perception of simultaneous stimuli. *Psychological Review, 87*, 272–300.

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

Duncan, J., & Humphreys, G. (1989). Visual search and stimulus similarity. *Psychological Review, 96*, 433–458.

Duncan, J., Humphreys, G., & Ward, R. (1997). Competitive brain activity in visual attention. *Current Opinion in Neurobiology, 7*, 255–261.

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.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1*, 1–47.

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/TR91, University of Cambridge, Department of Engineering.

Gerstner, W. (2000). Population dynamics of spiking neurons: Fast transients, asynchronous states, and locking. *Neural Computation, 12*, 43–89.

Gross, C. G., Desimone, R., Albright, T. D., & Schwartz, E. L. (1985). Inferior temporal cortex and pattern recognition. *Experimental Brain Research, 11*(Suppl.), 179–201.

Haxby, J. V., Horwitz, B., Ungerleider, L., Maisog, J. M., Pietrini, P., & Grady, C. L. (1994). The functional organization of human extrastriate cortex: A PET-rCBF study of selective attention to faces and locations. *Journal of Neuroscience, 14*, 6336–6353.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology, 160*, 106–154.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*, 1489–1506.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Review Neuro science, 2*, 194–203.

Kandel, E. R., Schwartz, J. H., & Jessel, T. H. (2000). *Principles of Neural Science* (fourth ed.). New York: McGraw-Hill.

Kastner, S., De Weerd, P., Desimone, R., & Ungerleider, L. (1998). Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. *Science, 282*, 108–111.

Kastner, S., Pinsk, M., De Weerd, P., Desimone, R., & Ungerleider, L. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron, 22*, 751–761.

Lee, T. S. (1996). Image representation using 2*D* Gabor wavelets. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*(10), 959–971.

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

Martinez, A. F., DiRusso, L., Anllo-Vento, L., Sereno, M. L., Buxton, R., & Hillyard, S. A. (2001). Putting spatial attention on the map: Timing and localization of stimulus selection processes in styriate and extrastriate visual areas. *Vision Research, 41*, 1437–1457.

McAdams, C., & Maunsell, J. H. R. (1999). Effects of attention on orientation-tuning functions of single neurons in macaque cortical area v4. *Journal of Neuroscience, 19*, 431–441.

Miller, E., Gochin, P., & Gross, C. (1993). Suppression of visual responses of neurons in inferior temporal cortex of the awake macaque by addition of a second stimulus. *Brain Research, 616*, 25–29.

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science, 229*, 782–784.

Motter, B. C. (1993). Focal attention produces spatially selective processing in visual cortical areas *VI*, *V*2, and *V*4 in the presence of competing stimuli. *Journal of Neurophysiology, 70*, 909–919.

Motter, B. (1994a). Neural correlates of attentive selection for colours or luminance in extrastriate area V4. *Journal of Neuroscience, 14*, 2178–2189.

Motter, B. C. (1994b). Neural correlates of attentive selection for colours or luminance in extrastriate area V4. *Journal of Neuroscience, 14*, 2190–2192.

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.

Perrett, D. L., Rolls, E. T., & Caan, W. (1982). Visual neurons responsive to faces in the monkey temporal cortex. *Experimental Brain Research, 47*, 329–342.

Pollen, D., & Ronner, S. (1981). Phase relationship between adjacent simple cells in the visual cortex. *Science, 212*, 1409–1411.

Posner, M., Walker, J., Friedrich, F., & Rafal, B. (1984). Effects of parietal injury on covert orienting of attention. *Journal of Neuroscience, 4*, 1863–1874.

Renart, A., Parga, N., & Rolls, E. T. (1999a). Associative memory properties of multiple cortical modules. *Network, 10*, 237–255.

Renart, A., Parga, N., & Rolls, E. T. (1999b). Backprojections in the cerebral cortex: Implications for memory storage. *Neural Computation, 11*, 1349–1388.

Reynolds, J., & Desimone, R. (1999). The role of neural mechanisms of attention in solving the binding problem. *Neuron, 24*, 19–29.

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience, 3*(Suppl.), 1199–1204.

Rolls, E. T. (1984). Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces. *Human Neurobiology, 3*, 209–222.

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

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

Rolls, E. T., Aggelopoulos, N. C., & Zheng, F. (2004). The receptive fields of inferior temporal cortex neurons in natural scenes. *Journal of Neuroscience, 23*, 339–348.

Rolls, E. T., & Deco, G. (2002). *Computational Neuroscience of Vision*. Oxford: Oxford University Press.

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

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

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

Rolls, E. T., & Tovee, M. J. (1995). The responses of single neurons in the temporal visual cortical areas of the macaque when more than one stimulus is present in the visual field. *Experimental Brain Research, 103*, 409–420.

Rolls, E. T., & Treves, A. (1998). *Neural Networks and Brain Function*. Oxford: Oxford University Press.

Salinas, E., & Abbott, L. F. (1997). Invariant visual responses from attentional gain fields. *Journal of Neurophysiology, 77*, 3267–3272.

Spitzer, H., Desimone, R., & Moran, J. (1988). Increased attention enhances both behavioral and neuronal performance. *Science, 240*, 338–340.

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.

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 T. G. Dietterich, S. Becker, & Z. Gharamani (Eds.), *Advances in Neural Information Processing Systems* (vol. 14, pp. 293–300). Cambridge, MA: MIT Press.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. Ingle, M. A. Goodale, & R. Mansfield (Eds.), *Analysis of Visual Behaviour*. Cambridge, MA: MIT Press.

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., Felleman, D., DeYoe, E., Olavarria, J., & Knierim, J. (1990). Modular and hierarchical organization of extrastriate visual cortex in the macaque monkey. *Cold Spring Harbor Symposia on Quantitative Biology, 55*, 679–696.

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

Wallis, G., Rolls, E. T., & Foldiak, P. (1993). Learning invariant responses to the natural transformations of objects. *International Joint Conference on Neural Networks, 2*, 1087–1090.

Wilson, H., & Cowan, J. (1972). Excitatory and inhibitory interactions in localised populations of model neurons. *Biophysics Journal, 12*, 1–24.