CHAPTER 14

The Neurophysiology and Computational Mechanisms of Object Representation

Edmund T. Rolls

14.1 Introduction

A concise description of the representation of objects and faces provided by inferior temporal cortex neurons in the primate (macaque) brain is followed by new findings about how this representation operates in natural scenes and allows a number of objects and their relative spatial position in a scene to be encoded. Then a computational approach to how the object recognition processes described are performed in the primate brain is discussed as well as the types of strategy that the human brain uses to solve the enormous computational problem of invariant object recognition in complex natural scenes (Rolls 2008b; Rolls and Deco 2002; Rolls and Stringer 2006b). This contribution aims to provide a closely linked neurophysiological and computational approach to object recognition and categorization. Other approaches are represented in this volume and elsewhere (Biederman 1987; Fukushima 1989; Riesenhuber and Poggio 2000; Serre et al. 2007), and are compared with current approaches (Rolls 2008b; Rolls and Deco 2002; Rolls and Stringer 2006b).

14.2 The Neurophysiology of Object Representation in the Inferior Temporal Visual Cortex

Some properties of the hierarchical organization of the primate ventral visual system that lead to the inferior temporal visual cortex (IT), where object representations are present (Rolls 2000; Rolls 2007b, 2008b; Rolls and Deco 2002), are shown in Figure 14.1. The receptive fields of neurons become larger related to the convergence from stage to stage, and the representation develops over the stages from features such as bars and edges, to combinations of features such as combinations of lines or colors in intermediate stages such as V4 (Hegde and Van Essen 2000; Ito and Komatsu 2004) to objects in IT. Neurons in IT provide a sparse distributed representation of objects, in that each neuron has a high firing rate to one or several objects, and gradually decreasing responses to other objects, with little or no response to most objects or faces (see

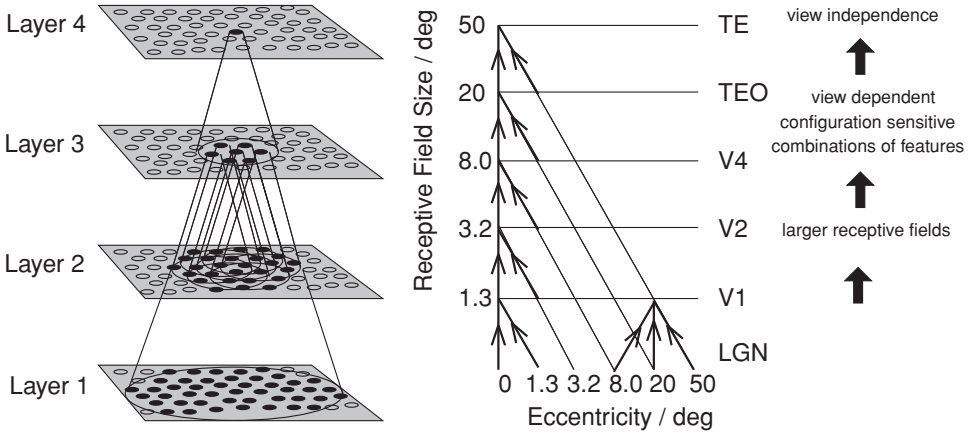


Figure 14.1. *Right,* Schematic diagram showing convergence achieved by the forward projections in the visual system, and the types of representation that may be built by competitive networks operating at each stage of the system from the primary visual cortex (V1) to the inferior temporal visual cortex (area TE) (see text). Lateral geniculate nucleus (LGN). Area TEO forms the posterior inferior temporal cortex. The receptive fields in the inferior temporal visual cortex (e.g., in the TE areas) cross the vertical midline (not shown). *Left,* Hierarchical network structure of VisNet, a feature hierarchy model of the processing in the visual pathways.

example in Fig. 14.2) (Franco et al. 2007; Rolls and Tovee 1995; Treves et al. 1999). Each neuron encodes information about objects that is independent of that carried by other neurons (up to for example 20–40 neurons), corresponding to response profiles to the set of stimuli that are uncorrelated, and providing an exponential increase in the number of objects that can be encoded with the number of neurons in the population (Franco et al. 2007; Rolls et al. 1997). Much of the information can be obtained from a population of neurons by using just a dot product (e.g., synaptically weighted) decoding of the responses (Rolls et al. 1997), which is the simplest type of neuronal decoding (Rolls 2008b). Most of the information from the firing of IT neurons is available if just the first few spikes occurring in a period of 20 or 40 ms after the onset of firing are used, facilitating the rapid transmission of information through the ventral visual system, with just approximately 15 ms per stage being sufficient (Fig. 14.1) (Rolls 2007a; Rolls et al. 2006; Rolls et al. 1994; Tovee and Rolls 1995). Moreover, most of the information is available in the spike counts from each neuron in a short period, rather than in stimulus-dependent correlations between neurons (at least in IT, where this has been examined during natural vision in a top-down attentionally biased search for an object in a complex scene (Aggelopoulos et al. 2005; Rolls 2008b)). This is an indication that stimulus-dependent synchrony (Singer 1999) is not necessary for binding (Rolls 2008b).

The evidence that IT neurons respond to and encode objects (with a sparse distributed representation), and not features, includes the following. Many IT neurons respond only to combinations of complex features, and not to the individual complex features themselves (Perrett et al. 1982; Rolls 2008b; Tanaka et al. 1990). Moreover, for IT neurons, the features need to be in the correct spatial configuration with respect to each other, and do not respond well to scrambled images in which the features have been moved to different positions with respect to each other (Rolls et al. 1994). Further, the

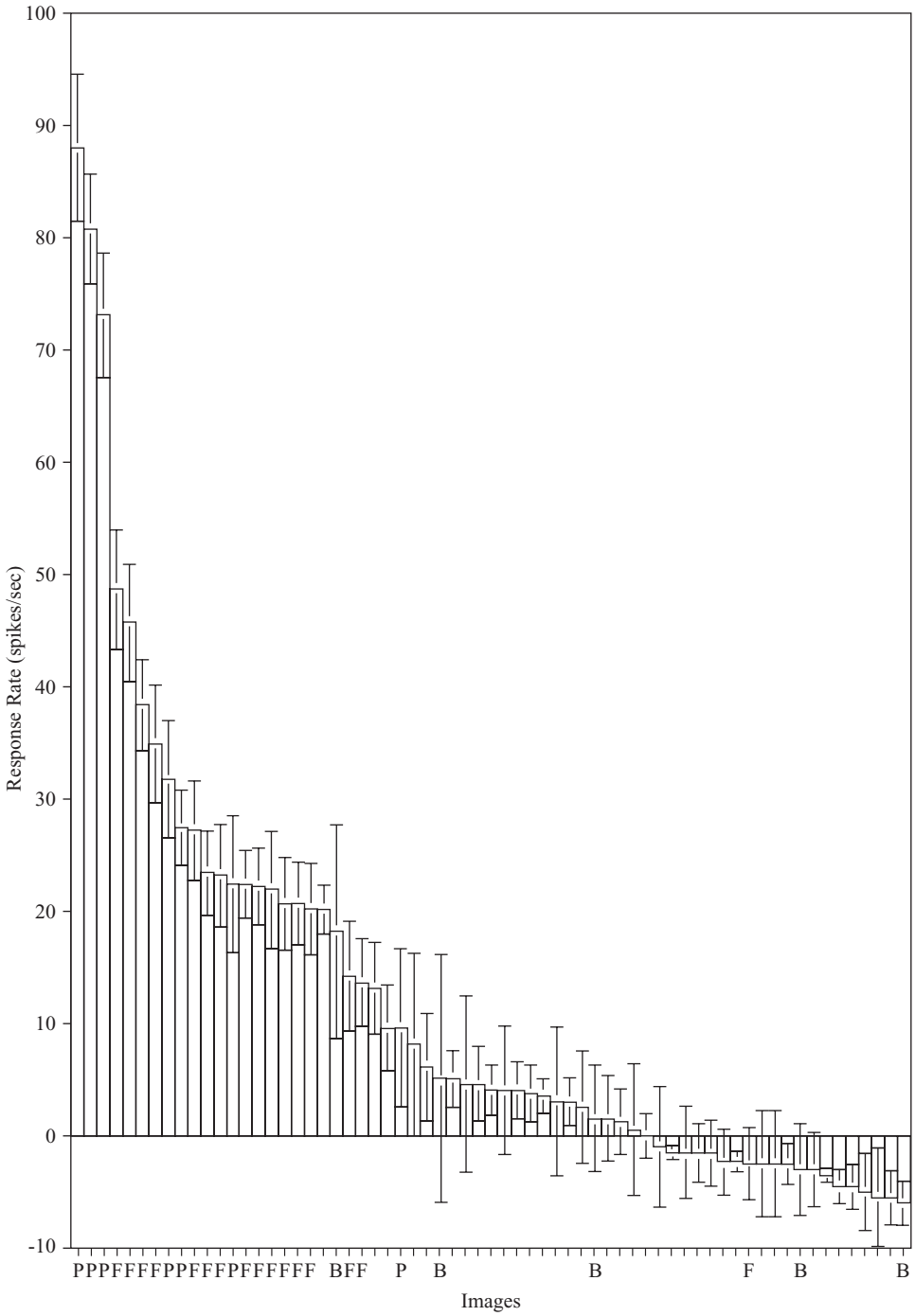


Figure 14.2. Firing-rate distribution of a single neuron in the temporal visual cortex to a set of 23 face (F) and 45 non-face images of natural scenes. The firing rate to each of the 68 stimuli is shown. *P* indicates a face-profile stimulus. *B* indicates a body-part stimulus, such as a hand. After Rolls and Tovee 1995a.

object (and face) representations are useful, in that they are rather invariant with respect to many transforms, including position on the retina, size, spatial frequency, and even view (Rolls 2000; Rolls 2007b, 2008a, b; Rolls and Deco 2002; Rolls and Stringer 2006b). The result is that a small population of neurons can represent which object of a very large number of different real-world objects is being viewed (Rolls 2008b). This is of fundamental importance for later stages of processing that can form memories about that object – for example, about what taste or other reinforcer is associated with the object, where the object is in a spatial scene, whether that object has been seen previously, and whether that object has been seen recently (Rolls 2008b). For this to operate correctly and usefully for objects, the representation in IT must generalize correctly to different transforms and so on of the object, and this is exactly what has been shown (Rolls 2008b).

If we go beyond what is required to represent normal real objects as found in the world with their only partly overlapping sets of features, to stimuli that require the ability to discriminate between stimuli composed of highly overlapping feature combinations in a low-dimensional feature space, then additional processing beyond IT in the perirhinal cortex may contribute to this type of discrimination (Rolls 2008b).

14.3 Outline of a Feature Hierarchy Model of the Computational Mechanisms in the Visual Cortex for Object and Face Recognition

The neurophysiological findings just described, and wider considerations on the possible computational properties of the cerebral cortex (Rolls 1989a, b, 1992, 2008b; Rolls and Treves 1998) lead to the following working hypotheses on object (including face) recognition by visual cortical mechanisms (Rolls 1992, 2008b; Rolls and Deco 2002).

Cortical visual processing for object recognition is considered to be organized as a set of hierarchically connected cortical regions consisting at least of V1, V2, V4, posterior inferior temporal cortex (TEO), and inferior temporal cortex (TE, including TE3, Tea, Tem, TE2, and TE1), as shown schematically in Figure 14.1. There is convergence from each small part of a region to the succeeding region (or layer in the hierarchy) in such a way that the receptive field sizes of neurons (e.g., 1 degree near the fovea in V1) become larger by a factor of approximately 2.5 with each succeeding stage (and the typical parafoveal receptive field sizes found would not be inconsistent with the calculated approximations of, e.g., 8 degrees in V4, 20 degrees in TEO, and 50 degrees in inferior temporal cortex; Boussaoud et al. 1991) (see Fig. 14.1). Such zones of convergence would overlap continuously with each other. This connectivity would be part of the architecture by which translation-invariant representations are computed. Each layer is considered to act partly as a set of local self-organizing competitive neuronal networks with overlapping inputs. (The region within which competition would be implemented would depend on the spatial properties of inhibitory interneurons, and might operate over distances of 1–2 mm in the cortex.) These competitive nets operate by a single set of forward inputs leading to (typically nonlinear, e.g., sigmoid) activation of output neurons; of competition between the output neurons mediated by a set of feedback inhibitory interneurons that receive from

many of the principal (in the cortex, pyramidal) cells in the net and project back (via inhibitory interneurons) to many of the principal cells, which serves to decrease the firing rates of the less active neurons relative to the rates of the more active neurons; and then of synaptic modification by a modified Hebb rule, such that synapses to strongly activated output neurons from active input axons strengthen, and from inactive input axons weaken (Rolls 2008b; Rolls and Deco 2002; Rolls and Treves 1998).

Translation, size, and view invariance could be computed in such a system by utilizing competitive learning operating across short time scales to detect regularities in inputs when real objects are transforming in the physical world (Rolls 1992; Rolls 2000; Rolls 2008b; Rolls and Deco 2002; Wallis and Rolls 1997). The hypothesis is that because objects have continuous properties in space and time in the world, an object at one place on the retina might activate feature analyzers at the next stage of cortical processing, and when the object was translated to a nearby position, because this would occur in a short period (e.g., 0.5 s), the membrane of the postsynaptic neuron would still be in its "Hebb-modifiable" state (caused, e.g., by calcium entry as a result of the voltage-dependent activation of NMDA receptors, or by continuing firing of the neuron implemented by recurrent collateral connections forming a short-term memory), and the presynaptic afferents activated with the object in its new position would thus become strengthened on the still-activated postsynaptic neuron. It is suggested that the short temporal window (e.g., 0.5 s) of Hebb-modifiability helps neurons to learn the statistics of objects moving in the physical world, and at the same time to form different representations of different feature combinations or objects, as these are physically discontinuous and present less regular correlations to the visual system. Földiák (1991) has proposed computing an average activation of the postsynaptic neuron to assist with translation invariance. I also suggest that other invariances (e.g., size, spatial frequency, rotation, and view invariance) could be learned by similar mechanisms to those just described (Rolls 1992). It is suggested that the process takes place at each stage of the multiple-layer cortical processing hierarchy, so that invariances are learned first over small regions of space, and then over successively larger regions. This limits the size of the connection space within which correlations must be sought.

Increasing complexity of representations could also be built in such a multiple-layer hierarchy by similar competitive learning mechanisms. In order to avoid a combinatorial explosion, it is proposed that low-order combinations of inputs would be what is learned by each neuron. Evidence consistent with this suggestion that neurons are responding to combinations of a few variables represented at the preceding stage of cortical processing is that some neurons in V2 and V4 respond to end-stopped lines, to tongues flanked by inhibitory subregions, or to combinations of colors (Hegde and Van Essen 2000; Ito and Komatsu 2004); in posterior inferior temporal cortex to stimuli that may require two or more simple features to be present (Tanaka et al. 1990); and in the temporal cortical face-processing areas to images that require the presence of several features in a face (such as eyes, hair, and mouth) in order to respond (Perrett et al. 1982; Rolls 2008b; Tanaka et al. 1990; Yamane et al. 1988). It is an important part of this suggestion that some local spatial information would be inherent in the features that are being combined (Elliffe et al. 2002). For example, cells might not respond to the combination of an edge and a small circle unless they were in the correct spatial relation to each other. This is in fact consistent with the data of Tanaka et al. (1990) and with our data on

face neurons (Rolls et al. 1994), in that some face neurons require the face features to be in the correct spatial configuration, and not jumbled). The local spatial information in the features being combined would ensure that the representation at the next level would contain some information about the (local) arrangement of features. Further low-order combinations of such neurons at the next stage would include sufficient local spatial information so that an arbitrary spatial arrangement of the same features would not activate the same neuron, and this is the proposed, and limited, solution that this mechanism would provide for the feature binding problem (Elliffe et al. 2002).

It is suggested that view-independent representations could be formed by the same type of computation, operating to combine a limited set of views of objects. The plausibility of providing view-independent recognition of objects by combining a set of different views of objects has been proposed by a number of investigators (Koenderink and Van Doorn 1979; Logothetis et al. 1994; Poggio and Edelman 1990; Ullman 1996). Consistent with the suggestion that the view-independent representations are formed by combining view-dependent representations in the primate visual system is the fact that in the temporal cortical areas, neurons with view-independent representations of faces are present in the same cortical areas as neurons with view-dependent representations (from which the view-independent neurons could receive inputs) (Booth and Rolls 1998; Hasselmo et al. 1989; Perrett et al. 1987). This solution to "object-based" representations is very different from that traditionally proposed for artificial vision systems, in which the coordinates in 3-D space of objects are stored in a database, and general-purpose algorithms operate on these to perform transforms such as translation, rotation, and scale change in 3-D space (Ullman 1996), or a linked list of feature parts is used (e.g., Marr 1982). In the present, much more limited but more biologically plausible scheme, the representation would be suitable for recognition of an object and for linking associative memories to objects, but would be less good for making actions in 3-D space to particular parts of, or inside, objects, as the 3-D coordinates of each part of the object would not be explicitly available. It is therefore proposed that visual fixation is used to locate in foveal vision part of an object to which movements must be made, and that local disparity and other measurements of depth then provide sufficient information for the dorsal visual system and motor system to make actions relative to the small part of space in which a local, view-dependent representation of depth would be provided (c.f. Ballard 1990; Rolls 2008b; Rolls and Deco 2002).

14.4 A Computational Model of Invariant Visual Object and Face Recognition

To test and clarify the hypotheses just described about how the visual system may operate to learn invariant object recognition, we have performed simulations that implement many of the ideas just described, and that are consistent with, and based on, much of the neurophysiology summarized in the previous sections. The network simulated (VisNet) can perform object, including face, recognition in a biologically plausible way, and after training shows, for example, translation and view invariance (Rolls 2008b; Rolls and Deco 2002; Rolls and Milward 2000; Rolls and Stringer 2006b; Wallis and Rolls 1997; Wallis et al. 1993).

In the four-layer network, the successive layers correspond approximately to V2, V4, the posterior temporal cortex, and the anterior temporal cortex (see Fig. 14.1). The forward connections to a cell in one layer are derived from a topologically corresponding region of the preceding layer, using a Gaussian distribution of connection probabilities to determine the exact neurons in the preceding layer to which connections are made. This schema is constrained to preclude the repeated connection of any cells. Each cell receives 100 connections from the 32×32 cells of the preceding layer, with a 67% probability that a connection comes from within four cells of the distribution center. Figure 14.1 shows the general convergent network architecture used. Within each layer, lateral inhibition between neurons has a radius of effect just greater than the radius of feedforward convergence just defined. The lateral inhibition is simulated via a linear local contrast-enhancing filter active on each neuron. (Note that this differs from the global "winner-take-all" paradigm implemented by Földiák (1991). The cell activation is then passed through a nonlinear cell activation function, which also produces contrast enhancement of the firing rates (Rolls 2008b; Rolls and Deco 2002; Rolls and Milward 2000; Rolls and Stringer 2006b).

In order that the results of the simulation might be made particularly relevant to understanding processing in higher cortical visual areas, the inputs to layer 1 come from a separate input layer that provides an approximation to the encoding found in visual area 1 (V1) of the primate visual system.

The synaptic learning rule used can be summarized as follows:

$$\delta w_{ij} = k \cdot m_i \cdot r'_j$$

and

$$m_i^{\mathrm{t}} = (1 - \eta) r_i^{(\mathrm{t})} + \eta m_i^{(\mathrm{t}-1)}$$

where r'_j is the jth input to the neuron, r_i is the output of the ith neuron, w_{ij} is the jth weight on the ith neuron, η governs the relative influence of the trace and the new input (typically, 0.4–0.6), and $m_i^{(\mathrm{t})}$ represents the value of the ith cell's memory trace at time t. In the simulation the neuronal learning was bounded by normalization of each cell's dendritic weight vector, as in standard competitive learning (see Rolls 2008b; Rolls and Deco 2002; Rolls and Treves 1998).

To train the network to produce a translation invariant representation, one stimulus was placed successively in a sequence of nine positions across the input, then the next stimulus was placed successively in the same sequence of nine positions across the input, and so on through the set of stimuli. The idea was to enable the network to learn whatever was common at each stage of the network about a stimulus shown in different positions. To train on view invariance, different views of the same object were shown in succession, then different views of the next object were shown in succession, and so on. It has been shown that the network can learn to form neurons in the last layer of the network that respond to one of a set of simple shapes (such as "T, L and +") with translation invariance, or to a set of five to eight faces with translation, view, or size invariance, provided that the trace learning rule (not a simple Hebb rule, but see discussion of spatial transformation learning) is used (see Figs. 14.3 and 14.4) (Rolls and Deco 2002; Wallis and Rolls 1997).

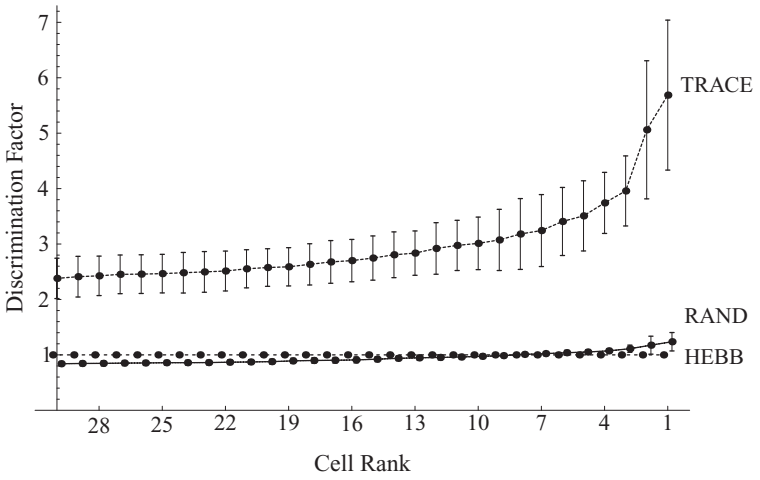


Figure 14.3. Comparison of VisNet network discrimination when trained with the trace learning rule, with a Hebb rule (no trace), and when not trained (Rand) on three stimuli (+, T, and L) at nine different locations. After Wallis and Rolls 1997.

There have been a number of investigations to explore this type of learning further. Rolls and Milward (2000) explored the operation of the trace learning rule used in the VisNet architecture, and showed that the rule operated especially well if the trace incorporated activity from previous presentations of the same object but received no contribution from the current neuronal activity being produced by the current exemplar of the object. The explanation for this is that this temporally asymmetric rule (the presynaptic term from the current exemplar, and the trace from the preceding exemplars) encourages neurons to respond to the current exemplar in the same way as they did to previous exemplars. It is of interest to consider whether intracellular processes related to Long-term potentiation (LTP) might implement an approximation of this rule, given that it is somewhat more powerful than the standard trace learning rule described before. Rolls and Stringer (2001) went on to show that part of the power of this type of trace rule can be related to gradient descent and temporal difference learning (Sutton and Barto 1998).

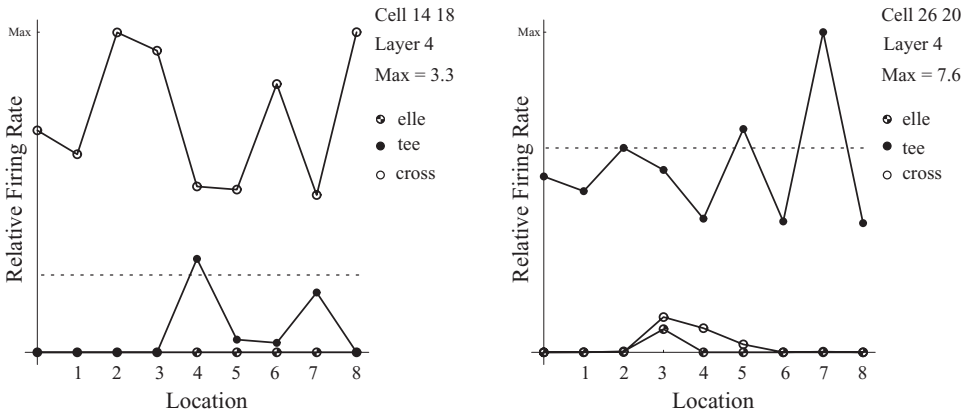


Figure 14.4. Response profiles for two fourth-layer neurons in VisNet (discrimination factors 4.07 and 3.62) in the L, T, and + invariance learning experiment. After Wallis and Rolls 1997.

Elliffe et al. (2002) examined the issue of spatial binding in this general class of hierarchical architecture studied originally by Fukushima (1980, 1989, 1991), and showed how by forming high spatial precision feature combination neurons early in processing, it is possible for later layers to maintain high precision for the relative spatial position of features within an object, yet achieve invariance for the spatial position of the whole object.

These results show that the proposed learning mechanism and neural architecture can produce cells with responses selective for stimulus identity with considerable position or view invariance (Rolls and Deco 2002). This ability to form invariant representations is an important property of the temporal cortical visual areas, for if a reinforcement association leading to an emotional or social response is learned to one view of a face, that learning will automatically generalize to other views of the face. This is a fundamental aspect of the way in which the brain is organized in order to allow this type of capability for emotional and social behavior (Rolls 1999; Rolls 2005). Further developments include operation of the system in a cluttered environment (Stringer and Rolls 2000); generalization from trained to untrained views of objects (Stringer and Rolls 2002); and a unifying theory of how invariant representations of optic flow produced by rotating or looming objects could be produced in the dorsal visual system (Rolls and Stringer 2006a) (sect. 14.10). The approach has also been extended to show that spatial continuity as objects gradually transform during training can be used with a purely associative learning rule to help build invariant representations in what has been termed continuous spatial transformation learning (Perry et al. 2006; Perry et al. 2009; Rolls and Stringer 2006b; Stringer et al. 2006). Further developments are considered after we introduce new neurophysiology on the representation of an object, and even several objects simultaneously, in complex natural scenes.

14.5 Neurophysiology of Object Representations in Complex Natural Scenes

Much of the neurophysiology of ventral stream visual processing has been performed with one feature, set of features, or object presented on a blank background, or in studies of attention two features may be presented on a blank background. How does the visual system operate in more realistic visual conditions when objects are presented in natural scenes? We learn much about computational aspects of natural vision from such investigations.

14.5.1 Object-Based Attention and Object Selection in Complex Natural Scenes

Object-based attention refers to attention to an object. For example, in a visual search task the object might be specified as what should be searched for, and its location must be found. In spatial attention, a particular location in a scene is pre-cued, and the object at that location may need to be identified.

To investigate how attention operates in complex natural scenes, and how information is passed from the IT to other brain regions to enable stimuli to be selected

from natural scenes for action, Rolls, Aggelopoulos, and Zheng (2003) analyzed the responses of inferior temporal cortex neurons to stimuli presented in complex natural backgrounds while performing a top-down object-based attentional search task. The monkey had to search for two objects on a screen; a touch of one object was rewarded with juice, and of another object was punished with saline. Neuronal responses to the effective stimuli for the neurons were compared when the objects were presented in the natural scene or on a plain background. It was found that the response of the neuron to objects at the fovea was hardly reduced when they were presented in natural scenes, and the selectivity of the neurons remained (see also Sheinberg and Logothetis 2001. However, the main finding was that the magnitudes of the responses of the neurons typically became much less in the real scene the further the monkey fixated in the scene away from the object – that is, the receptive fields became smaller in complex natural scenes (Fig. 14.5). It is proposed that this reduced translation invariance (i.e., invariance with respect to position on the retina) in natural scenes helps an unambiguous representation of an object that may be the target for action to be passed to the brain regions that receive from the primate IT. It helps with the binding problem, by reducing in natural scenes the effective receptive field of at least some IT neurons to approximately the size of an object in the scene. In a very similar task, in which one of two objects had to be selected against a complex background, it is found that almost all the information (>95% of the total information) is encoded by the firing rates of simultaneously recorded IT neurons, and that very little information is encoded by stimulus-dependent synchronization, which may therefore not be important for implementing feature binding (Aggelopoulos et al. 2005; Rolls 2008b).

It is also found that in natural scenes, the effect of object-based attention on the response properties of IT neurons is relatively small, as illustrated in Figure 14.3 (Rolls et al. 2003). The results summarized in Figure 14.5 for 5-degree stimuli show that the receptive fields were large (77.6 degrees) with a single stimulus in a blank background (top left), and were greatly reduced in size (to 22.0 degrees) when presented in a complex natural scene (top right). The results also show that there was little difference in receptive field size or firing rate in the complex background when the effective stimulus was selected for action (bottom right, 19.2 degrees), and when it was not (middle right, 15.6 degrees) (Rolls et al. 2003). (For comparison, the effects of attention against a blank background were much larger, with the receptive field increasing from 17.2 degrees to 47.0 degrees as a result of object-based attention, as shown in Fig. 14.5.) The computational basis for these relatively minor effects of object-based attention when objects are viewed in natural scenes is considered in section 14.6.

14.5.2 The Interface from Object Representations to Action

These findings on how objects are represented in natural scenes make the interface to memory and action systems simpler, in that what is at the fovea can be interpreted (e.g., by an associative memory in the orbitofrontal cortex or amygdala) partly independently of the surroundings, and choices and actions can be directed if appropriate to what is at the fovea (Ballard 1993; Rolls and Deco 2002). There thus may be no need to have the precise coordinates of objects in space represented in the IT and passed to the

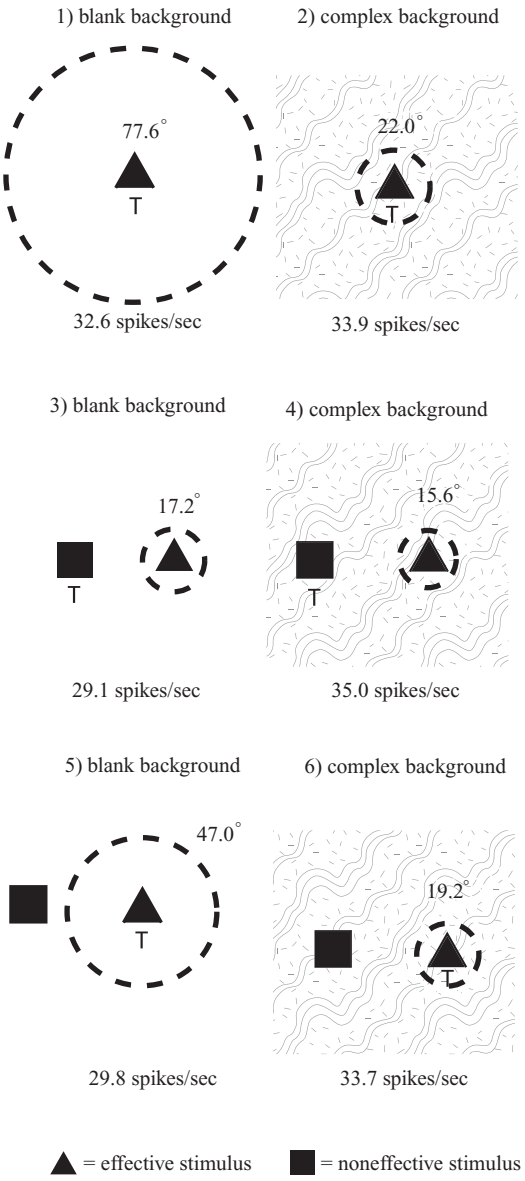


Figure 14.5. Summary of the receptive field sizes of inferior IT neurons to a 5-degree effective stimulus presented in either a blank background (blank screen) or in a natural scene (complex background). The stimulus that was a target for action in the different experimental conditions is marked by T. When the target stimulus was touched, a reward was obtained. The mean receptive field diameter of the population of neurons analyzed, and the mean firing rate in spikes/sec, is shown. The stimuli subtended 5 degrees × 3.5 degrees at the retina, and occurred on each trial in a random position in the 70 degrees × 55 degrees screen. The dashed circle is proportional to the receptive field size. *Top row*, Responses with one visual stimulus in a blank (left) or complex (right) background. *Middle row*, Responses with two stimuli, when the effective stimulus was not the target of the visual search. *Bottom row*, Responses with two stimuli, when the effective stimulus was the target of the visual search. After Rolls et al. 2003.

motor system for action to be directed at the target. Instead, given that the output of the IT in complex visual scenes primarily about an object at the fovea, the dorsal visual system may be able to initiate action to whatever is at the fovea (Rolls 2008b; Rolls and Deco 2006). The condition for an action to be performed is that the ventral visual system must have provided a representation of the object, and this must have been identified as a goal (i.e., as a rewarded object) by, for example, associative reward-based lookup in brain structures such as the orbitofrontal cortex and amygdala, where visual stimuli are interfaced to reward systems (Rolls 2005).

This approach to object representation is very different to that attempted in some artificial vision systems, in which identification of all the objects in a scene, and where they are in the scene, is attempted. The research described here shows that biological systems perform a much simpler task and provide representations in complex natural scenes primarily of objects at or close to the fovea. Consistent with this, much of our perception of the objects in a scene is a memory-based not a perceptual representation, as shown by change blindness. Change blindness refers to our inability to detect a change to objects in a scene that can occur if the scene is changed during, for example, a blink or saccade (Simons and Rensink 2005). (The blink or saccade means that motion, etc., cannot be used to detect the removal of the object.) The small receptive fields of IT neurons in complex scenes provides an explanation for change blindness (Rolls 2008c). Another phenomenon related to the small size of IT neurons in complex natural scenes is inattentional blindness. This is demonstrated, for example, when watching a basketball-passing event in which the instructions were to count the ball passes between the team members with white shirts; the subjects were less likely to report that a black gorilla was walking across the scene compared to participants counting passes between the team members with black shirts (Simons and Chabris 1999). It is proposed here that an important factor that contributes to inattentional and change blindness is the reduced diameter of the receptive fields of IT neurons that occurs in natural scenes. This would result in a failure to activate representations of objects (such as the gorilla that appears in the middle of the basketball-passing event) if they are not close to the fovea, in complex cluttered scenes. The instructions to count the number of ball passes between the players with black shirts would tend, by top-down biased competition effects, to facilitate representations of black stimuli, including the (unexpected) (black) gorilla. This facilitation, the neurophysiology shows (Fig. 14.5) (Rolls et al. 2003), consists of increasing somewhat the receptive field sizes of the neurons that are being biased (in this case, neurons that respond to black features). This, given many eye movements round the scene, would make it more likely that some of the black in the gorilla would activate some IT neurons. Thus, a combination of the reduced receptive field size of neurons in a complex natural scene and the effects of top-down biased competition would help account for inattentional blindness (Rolls 2008c).

The main point here is that these phenomena are related to the small receptive fields of IT neurons in complex natural scenes, and these small receptive fields are fundamental to how the biological visual system operates, for it greatly simplifies the computational problem compared to attempting to analyze the whole scene. It also reduces the binding and segmentation problems. The computational mechanisms that produce this reduction in receptive field size in complex natural scenes are described in section 14.6.

14.5.3 The Representation of Information About the Relative Positions of Multiple Objects in a Scene

These experiments have been extended to address the issue of how several objects are represented in a complex scene. The issue arises because the relative spatial locations of objects in a scene must be encoded (this is possible even in short presentation times without eye movements (Biederman 1972) and has been held to involve some spotlight of attention), and because what is represented in complex natural scenes is primarily about what is at the fovea; however, we can locate more than one object in a scene even without eye movements. Aggelopoulos and Rolls (2005) showed that with five objects simultaneously present in the receptive field of IT neurons, although all the neurons responded to their effective stimulus when it was at the fovea, some could also respond to their effective stimulus when it was in a parafoveal position 10 degrees from the fovea. An example of such a neuron is shown in Figure 14.6. The asymmetry is much more evident in a scene with five images present (Fig. 14.6*A*) than when only one image is shown on an otherwise blank screen (Fig. 14.6*B*). Competition between different stimuli in the receptive field thus reveals the asymmetry in the receptive field of IT neurons.

This has been tested computationally in VisNet, and it has been shown that the receptive fields in VisNet become small and asymmetric in scenes in which multiple objects are present, with the underlying mechanism that asymmetries related to the probabilistic nature of the excitatory feedforward connections and lateral inhibitory connections are revealed when the competition is high owing to the presence of multiple objects in a scene (Rolls et al. 2008).

The asymmetry provides a way of encoding the position of multiple objects in a scene. Depending on which asymmetric neurons are firing, the population of neurons provides information (using a distributed representation of the type that a population of receiving neurons with mutual, lateral, inhibition can decode; see Rolls 2008b) to the next processing stage not only about which image is present at or close to the fovea, but where it is with respect to the fovea. This information is provided by neurons that have firing rates that reflect the relevant information, and stimulus-dependent synchrony is not necessary. Top-down attentional biasing input could, by biasing the appropriate neurons, facilitate bottom-up information about objects without any need to alter the time relations between the firing of different neurons. The exact position of the object with respect to the fovea, and effectively its spatial position relative to other objects in the scene, would then be made evident by the subset of asymmetric neurons firing.

This is the solution that these experiments indicate is used for the representation of multiple objects in a scene (Aggelopoulos and Rolls 2005), an issue that has previously been difficult to account for in neural systems with distributed representations (Mozer 1991) and for which "attention" has been a proposed solution.

14.6 Object Representation and Attention in Natural Scenes: A Computational Account

The results described in section 14.5 and summarized in Figure 14.5 show that the receptive fields of IT neurons were large (77.6 degrees) with a single stimulus in a

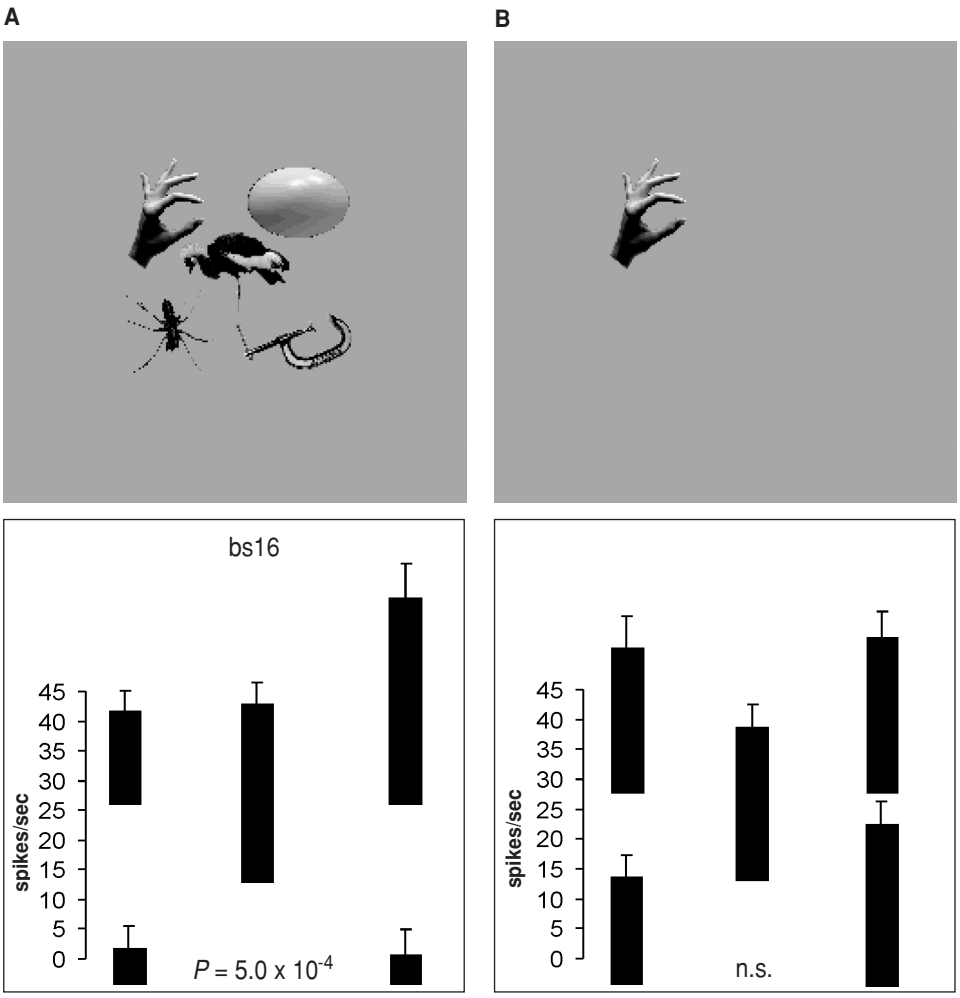


Figure 14.6. *A,* The responses (firing rate with the spontaneous rate subtracted, means ± sem) of one neuron when tested with five stimuli simultaneously present in the close (10 degree) configuration with the parafoveal stimuli located 10 degrees from the fovea. *B,* The responses of the same neuron when only the effective stimulus was presented in each position. The firing rate for each position is that when the effective stimulus for the neuron was in that position. The *P* value is that from the ANOVA calculated over the four parafoveal positions. After Aggelopoulos and Rolls 2005.

blank background (top left), and were greatly reduced in size (to 22 degrees) when presented in a complex natural scene (top right). The results also show that there was little difference in receptive field size or firing rate in the complex background when the effective stimulus was selected for action (bottom right), and when it was not (middle right) (Rolls et al. 2003).

Trappenberg, Rolls, and Stringer (2002) have suggested what underlying mechanisms could account for these findings, and they simulated a model to test the ideas. The model utilizes an attractor network which represents the inferior temporal visual cortex (implemented by the recurrent excitatory connections between IT neurons), and

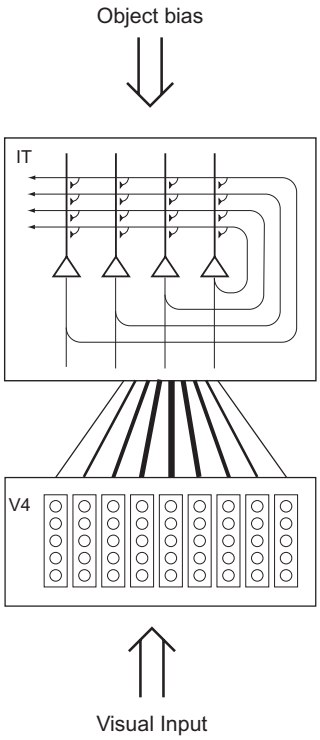


Figure 14.7. The architecture of the inferior temporal cortex (IT) model of Trappenberg et al. (2002) operating as an attractor network with inputs from the fovea given preferential weighting by the greater magnification factor of the fovea. The model also has a top-down, object-selective bias input. The model was used to analyze how object vision and recognition operate in complex natural scenes.

a neural input layer with several retinotopically organized modules representing the visual scene in an earlier visual cortical area such as V4 (Fig. 14.7). An attractor, or autoassociation, network is implemented by associatively modifiable connections between the neurons in the network. Each vector of neuronal firing rates represents a stimulus or memory, and is stored by associative synaptic modification between the neurons representing that stimulus or memory. When even a partial retrieval cue is provided that is similar to one of the patterns stored in the network, the network is attracted into the state with the neurons in the original pattern active and, thus, implements memory retrieval. The neurons continue firing stably with just that set of neurons active, thus implementing short-term memory, too. A description of the operation of these and other networks is provided elsewhere (Rolls 2008b). The attractor network aspect of the model produces the property that the receptive fields of IT neurons can be large in blank scenes by enabling a weak input in the periphery of the visual field to act as a retrieval cue for the object attractor. On the other hand, when the object is shown in a complex background, the object closest to the fovea tends to act as the retrieval cue for the attractor, because the fovea is given increased weight in activating the IT module because the magnitude of the input activity from objects at the fovea is greatest due to the cortical higher magnification factor of the fovea incorporated into the model. (The cortical magnification factor can be expressed as the number of

millimeters of cortex representing 1 degree of visual field. The cortical magnification factor decreases rapidly with increasing eccentricity from the fovea (Cowey and Rolls 1975; Rolls and Cowey 1970).) This results in smaller receptive fields of IT neurons in complex scenes, because the object tends to need to be close to the fovea to trigger the attractor into the state representing that object. In other words, if the object is far from the fovea in a cluttered scene, then the object will not trigger neurons in IT that represent it, because neurons in IT are preferentially being activated by another object at the fovea. This may be described as an attractor model in which the competition for which attractor state is retrieved is weighted towards objects at the fovea.

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

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

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

In natural scenes, the model is able to account for the neurophysiological data that the IT neuronal responses are larger when the object is close to the fovea, by virtue of the fact that objects close to the fovea are weighted by the cortical magnification factor. The model accounts for the larger receptive field sizes from the fovea of IT neurons in natural backgrounds if the target is the object being selected compared to when it is not selected (Rolls et al. 2003). The model accounts for this by an effect of top-down bias, which simply biases the neurons towards particular objects compensating for their decreasing inputs produced by the decreasing magnification factor modulation with increasing distance from the fovea. Such object-based attention signals could originate in the prefrontal cortex and could provide the object bias for the inferotemporal cortex (Renart et al. 2001; Renart et al. 2000; Rolls and Deco 2002). Important properties of the architecture for obtaining the results just described are the high magnification factor at the fovea and the competition between the effects of different inputs, implemented in the preceding simulation by the competition inherent in an attractor network.

We have also been able to obtain similar results in a hierarchical feedforward network in which each layer operates as a competitive network (Deco and Rolls 2004). This network thus captures many of the properties of our hierarchical model of invariant visual object recognition in the ventral visual stream (Elliffe et al. 2002; Rolls 1992; Rolls and Deco 2002; Rolls and Milward 2000; Rolls and Stringer 2001, 2006a, b; Stringer et al. 2006; Stringer and Rolls 2000, 2002; Wallis and Rolls 1997), but also incorporates a foveal magnification factor and top-down projections with a dorsal visual stream so that attentional effects can be studied, as shown in Figure 14.8.

Deco and Rolls (2004) trained the network described shown in Figure 14.8 with two objects, and used the trace learning rule (Rolls and Milward 2000; Wallis and Rolls 1997) to achieve translation invariance. With this model, we were able to obtain similar effects to those already described, and in addition were able to make predictions about the interaction between stimuli when they were placed in different relative positions with respect to the fovea.

14.7 Learning Invariant Representations of an Object with Multiple Objects in the Scene and with Cluttered Backgrounds

The results of simulations of learning with an object in a cluttered background suggest that in order for a neuron to *learn* invariant responses to different transforms of a stimulus when it is presented during training in a cluttered background, some form of segmentation is required in order to separate the figure (i.e., the stimulus or object) from the background (Stringer and Rolls 2000). This segmentation might be performed using evidence in the visual scene about different depths, motions, colors, and so on of the object from its background. In the visual system, this might mean combining evidence represented in different cortical areas and might be performed by cross-connections between cortical areas to enable such evidence to help separate the representations of objects from their backgrounds in the form-representing cortical areas.

A second way in which training a feature hierarchy network in a cluttered natural scene may be facilitated follows from the finding that the receptive fields of IT neurons shrink from in the order of 70 degrees in diameter when only one object is present in

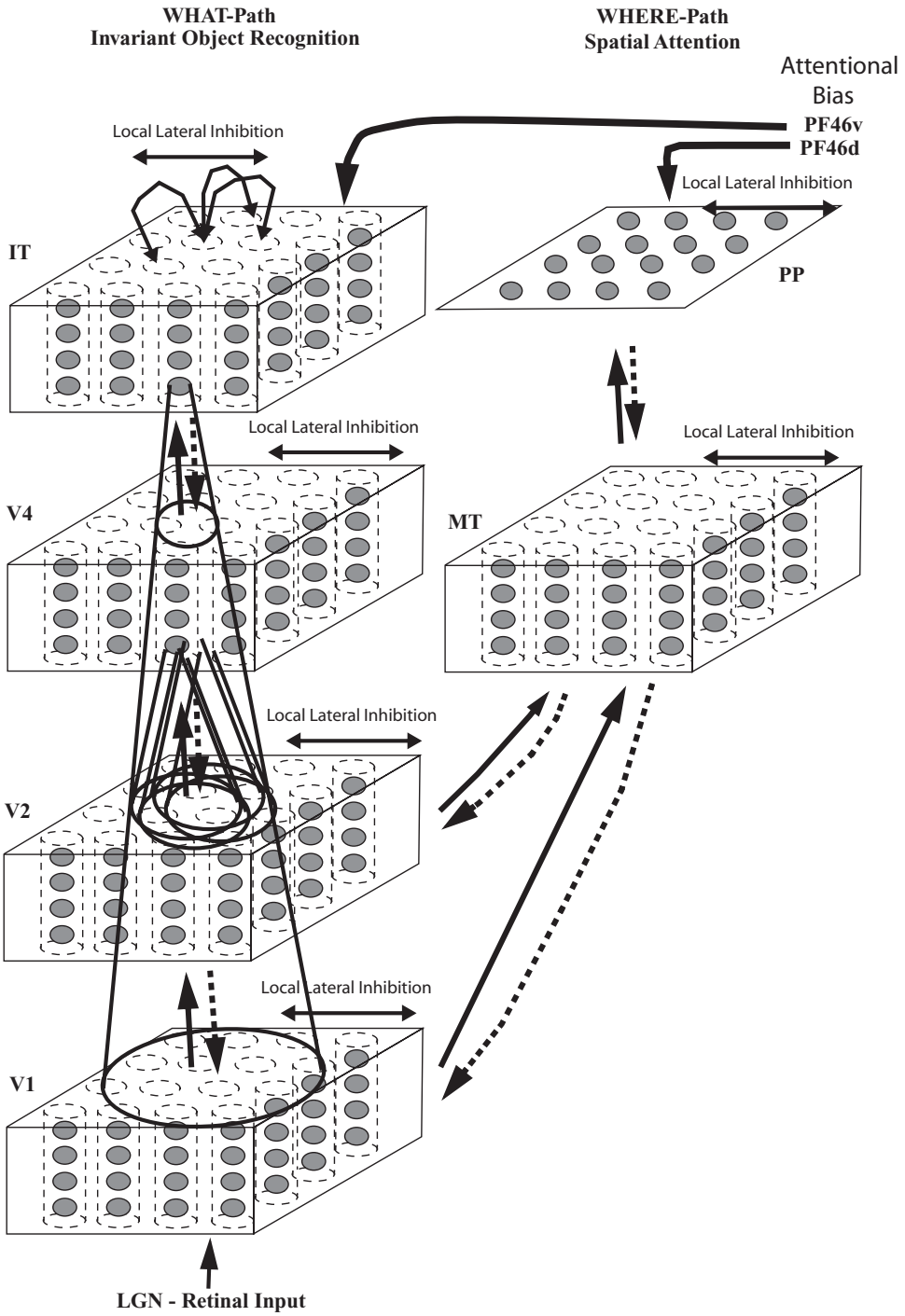


Figure 14.8. Cortical architecture for hierarchical and attention-based visual perception. The system is essentially composed of five modules structured such that they resemble the two known main visual paths of the mammalian visual cortex (MT). Information from the retinogeniculostriate pathway enters the visual cortex through area V1 in the occipital lobe and proceeds into two processing streams. The occipitotemporal stream leads ventrally through V2-V4 and IT (inferior temporal visual cortex), and is mainly concerned with object recognition. The occipitoparietal stream leads dorsally into PP (posterior parietal complex), and is responsible for maintaining a spatial map of an object's location. The solid lines with arrows between levels show the forward connections, and the dashed lines, the top-down backprojections. Short-term memory systems in the prefrontal cortex (PF46) apply top-down attentional bias to the object or spatial processing streams. After Deco and Rolls 2004.

a blank scene to much smaller values of as little as 5–10 degrees close to the fovea in complex natural scenes (Rolls et al. 2003) (see sects. 14.5 and 14.6). This allows primarily the object at the fovea to be represented in the IT and, it is proposed, for learning to be about this object, and not about the other objects in a whole scene.

Third, top-down spatial attention (Deco and Rolls 2004, 2005a; Deco and Rolls 2005b; Rolls and Deco 2002) could bias the competition towards a region of visual space in which the object to be learned is located.

Fourth, if object 1 is presented during training with different objects present on different trials, then the competitive networks that are part of VisNet will learn to represent each object separately, because the features that are part of each object will be much more strongly associated together, than are those features with the other features present in the different objects seen on some trials during training (Stringer and Rolls 2008; Stringer et al. 2007). It is a natural property of competitive networks that input features that co-occur frequently are allocated output neurons to represent the pattern as a result of the learning. Input features that do not co-occur frequently may not have output neurons allocated to them. This principle may help feature hierarchy systems to learn representations of individual objects, even when other objects with some of the same features are present in the visual scene, but with different other objects on different trials. With this fundamental and interesting property of competitive networks, it has now become possible for VisNet to self-organize invariant representations of individual objects, even though each object is always presented during training with at least one other object present in the scene (Stringer and Rolls 2008; Stringer et al. 2007).

14.8 A Biased Competition Model of Object and Spatial Attentional Effects on the Representations in the Visual System

So far, the models have been operating mainly in a feedforward, bottom-up way. In this section, I consider a computational account of how top-down influences of attention operate by biased competition to modulate the representations in the visual system.

Visual attention exerts top-down influences on the processing of sensory information in the visual cortex; therefore, it is intrinsically associated with interactions between cortical areas. Elucidating the neural basis of visual attention is an excellent paradigm for understanding the basic mechanisms of intercortical neurodynamics. Recent developments in cognitive neuroscience allow a more direct study of the neural mechanisms underlying attention in humans and primates. In particular, the work of Chelazzi, Miller, Duncan, and Desimone (1993) has led to a promising account of attention termed the "biased-competition hypothesis" (Desimone and Duncan 1995; Reynolds and Desimone 1999). According to this hypothesis, attentional selection operates in parallel by biasing an underlying competitive interaction between multiple stimuli in the visual field toward one stimulus or another, so that behaviorally relevant stimuli are processed in the cortex while irrelevant stimuli are filtered out. Thus, attending to a stimulus at a particular location or with a particular feature biases the underlying neural competition in a certain brain area in favor of neurons that respond to the location, or the features, of the attended stimulus. As a result of the competition, neurons that represent features without a top-down bias have reduced activity.

Neurodynamical models for biased competition have been proposed and successfully applied in the context of attention and working memory. In the context of attention, Usher and Niebur (1996) introduced an early model of biased competition. Deco and Zihl (2001) extended Usher and Niebur's model to simulate the psychophysics of visual attention by visual search experiments in humans. Their neurodynamical formulation is a large-scale hierarchical model of the visual cortex whose global dynamics is based on biased-competition mechanisms at the neural level. Attention then appears as an emergent effect related to the dynamical evolution of the whole network. This large-scale formulation, using a simplified version of the architecture shown in Figure 14.8, has been able to simulate and explain in a unifying framework, visual attention in a variety of tasks and at different cognitive neuroscience experimental measurement levels (Deco and Rolls 2005a); single-cells (Deco and Lee 2002; Rolls and Deco 2002), fMRI (Corchs and Deco 2002), psychophysics (Deco and Rolls 2005a; Rolls and Deco 2002), and neuropsychology (Deco and Rolls 2002). In the context of working memory, further developments (Deco and Rolls 2003; Rolls 2008b) managed to model in a unifying form attentional and memory effects in the prefrontal cortex, integrating single-cell and fMRI data, and different paradigms in the framework of biased competition.

In particular, Deco and Rolls (2005b) extended previous concepts of the role of biased competition in attention by providing the first analysis at the integrate-and-fire neuronal level, which allows the neuronal nonlinearities in the system to be explicitly modeled, in order to investigate realistically the processes that underlie the apparent gain modulation effect of top-down attentional control. In the integrate-and-fire implementation, the synaptic currents that lead to activation of the neuron, and then the generation of a spike when a threshold is reached, are modeled, producing a system that can model many of the nonlinearities in the system and also the effects of the probabilistic spiking of the neurons in the network on how it operates, as described elsewhere (Rolls 2008b; Rolls and Deco 2002). In the integrate-and-fire model, the competition is implemented realistically by the effects of the excitatory neurons on the inhibitory neurons and their return inhibitory synaptic connections. This was also the first integrate-and-fire analysis of top-down attentional influences in vision that explicitly models the interaction of several different brain areas. Part of the originality of the model is that in the form in which it can account for attentional effects in V2 and V4 in the paradigms of Reynolds, Chelazzi, and Desimone (1999) in the context of biased competition, the model with the same parameters effectively makes predictions that show that the "contrast gain" effects in MT (Martinez-Trujillo and Treue 2002) can be accounted for by the same model. For example, the top-down attentional modulation effects are most evident when the bottom-up input is weak (e.g., has low contrast), because the top-down effects themselves must never be so strong that they dominate perception. In addition, the top-down modulation can appear as a nonlinear multiplication effect with the bottom-up input, although the processes involved include only linear summation within the neurons of bottom-up and top-down synaptic inputs, and the threshold nonlinearity of neurons involved in whether an action potential is generated (Deco and Rolls 2005b). These detailed and quantitative analyses of neuronal dynamical systems are an important step towards understanding the operation of complex processes such as top-down attention, which necessarily involve the interaction of several brain areas.

They are being extended to provide neurally plausible models of decision-making and action selection (Deco and Rolls 2003; Deco and Rolls 2005d, 2006; Rolls 2008b).

In relation to representation in the brain, the impact of these findings is that they show details of the mechanisms by which representations can be modulated by attention, and moreover can account for many phenomena in attention using models in which the firing rate of neurons is represented, and in which stimulus-dependent neuronal synchrony is not involved (Rolls 2007b, 2008b).

The top-down back-projection pathways between adjacent cortical areas that implement the attentional effects in this model are weak relative to the forward (bottom-up) inputs. Consistent with this, the back-projection synapses end on the apical dendrites of pyramidal cells in the preceding cortical area, quite far from the cell body, where they might be expected to be sufficient to dominate the cell firing when there is no forward input close to the cell body (i.e., during memory recall, which may be one of the functions of these back-projection pathways, given the very large number of connections and their associative modifiability) (Rolls 1989a, 2008b; Treves and Rolls 1994). In contrast, when there is forward input to the neuron, activating synapses closer to the cell body than the back-projecting inputs, this would tend to electrically shunt the back-projection effects received on the apical dendrites, accounting for their relatively small but useful biasing effect. The associative modifiability is useful for setting up the connectivity required not only for memory recall, but also for top-down attentional effects to influence the correct neurons (Rolls 2008b).

14.9 Decision-making in Perception

When the surfaces of a Necker cube flip from back to front, it is as if an internal model of the 3-D structure of the cube is influencing the representation of the depth of the different surfaces of the cube (Gregory 1970, 1998; Helmholtz 1857; Rolls 2008b). When two objects are presented to the visual system in rivalry, first one is seen, and then there is a probabilistic flip to the other object being seen (Maier et al. 2005). Again, an internal representation of each object appears to be influencing visual processing so that first the whole of one object, and then of the other object, is seen, and not a combination of the features of both. A recent model of probabilistic decision-making in the brain (Deco and Rolls 2006) contributes to our understanding of these perceptual phenomena, for it is suggested that the underlying computational mechanism may be similar, providing a unifying approach to these aspects of brain processing (Rolls 2008b).

The architecture of the model is that of an attractor network and is within the theoretical framework utilized by Wang (2002), which is based on a neurodynamical model first introduced by Brunel and Wang (2001) and which has been recently extended and successfully applied to explain several experimental paradigms (Deco and Rolls 2002; Deco and Rolls 2003, 2005b; Deco et al. 2004; Deco et al. 2005; Rolls and Deco 2002; Szabo et al. 2004). In this framework, we model probabilistic decision-making by an attractor network of interacting integrate-and-fire neurons with spiking activity organized into a discrete set of populations. For a binary decision, there are two populations of neurons, each one corresponding to one of the decisions. Each population has its own biasing input, f1 and f2. The network starts with spontaneous activity, and if the biases

are equal, what is effectively a biased competition network eventually falls into one of the attractor states (i.e., with one of the populations of neurons firing with a high rate), with a probability that each attractor wins being 0.5. The probabilistic settling of the network is due to the inherent noise in the finite size network due to the Poisson-like firing of the neurons (Deco and Rolls 2006). The attractor that wins represents the decision. If one of the biases is stronger than the other, then the probability that the network will reach that decision increases. Because the model is a short-term memory network, the system can integrate information over long time periods, of hundreds of milliseconds, before a decision is reached. As the biases become more unequal, the reaction time of the decision made by the network decreases. As the biases are both increased, the magnitude of the difference between them for a decision to be reached must be increased in proportion, that is $\Delta I/I =$ a constant. The network implements Weber's law because as I is increased, the activity of the inhibitory feedback neurons in the integrate-and-fire network increases linearly, and these produce divisive inhibition on the excitatory cells that form the attractor, resulting in a need for ΔI to increase in proportion to the divisive inhibition, resulting in $\Delta I/I =$ a constant (Deco and Rolls 2006).

The decision-making network was tested (Deco and Rolls 2006) against neurophysiological data on decision-making for vibrotactile frequency discrimination. However, it is proposed here that the same type of decision-making network could be implemented in many brain areas, to account for many types of probabilistic decision-making. In the context of visual perception, it is proposed that the two attractor states might represent the two alternative interpretations of which side of a Necker cube is closer, or which binocularly presented image is being seen. Then with some adaptation in the excitatory neurons of the synapses between them, which has been modeled (Deco and Rolls 2005c; Deco and Rolls 2005d), the firing rate in the currently active attractor would gradually decrease, allowing the other attractor to spontaneously become active in a probabilistic way that depends on the number of spikes that happen to be generated with Poisson-like statistics by the different neurons in the different attractors. In this way, the higher-level representation of the object (e.g., the cube or the image) would spontaneously and probabilistically flip, and this higher-level representation would bias the lower-level representations by top-down influences so that first one edge and then the other edge of the cube would be biased to appear close, or the features in one image versus the other would be biased by the top-down competitive influence in pattern (Maier et al. 2005) and even binocular rivalry.

It is thus proposed that this model of decision-making (Deco and Rolls 2006) might account for many decision-making processes in the brain, including those involved in the interpretation of visual images and the recognition of objects. In this sense, what is seen or emphasized at the lower level is biased or pre-empted by the state of the higher-level representations, themselves determined probabilistically. We may note that in fact the decision may not be taken only in the high-level network but could be distributed throughout the system of interconnected networks, with their feedforward and top-down feedback connections all contributing to the decision-making (Deco and Rolls 2006), and all perhaps to the probabilistic change of state according to the extent to which the coupled neurons at different levels of the network show adaptation and by their probabilistic spiking contribute to the noise in the system.

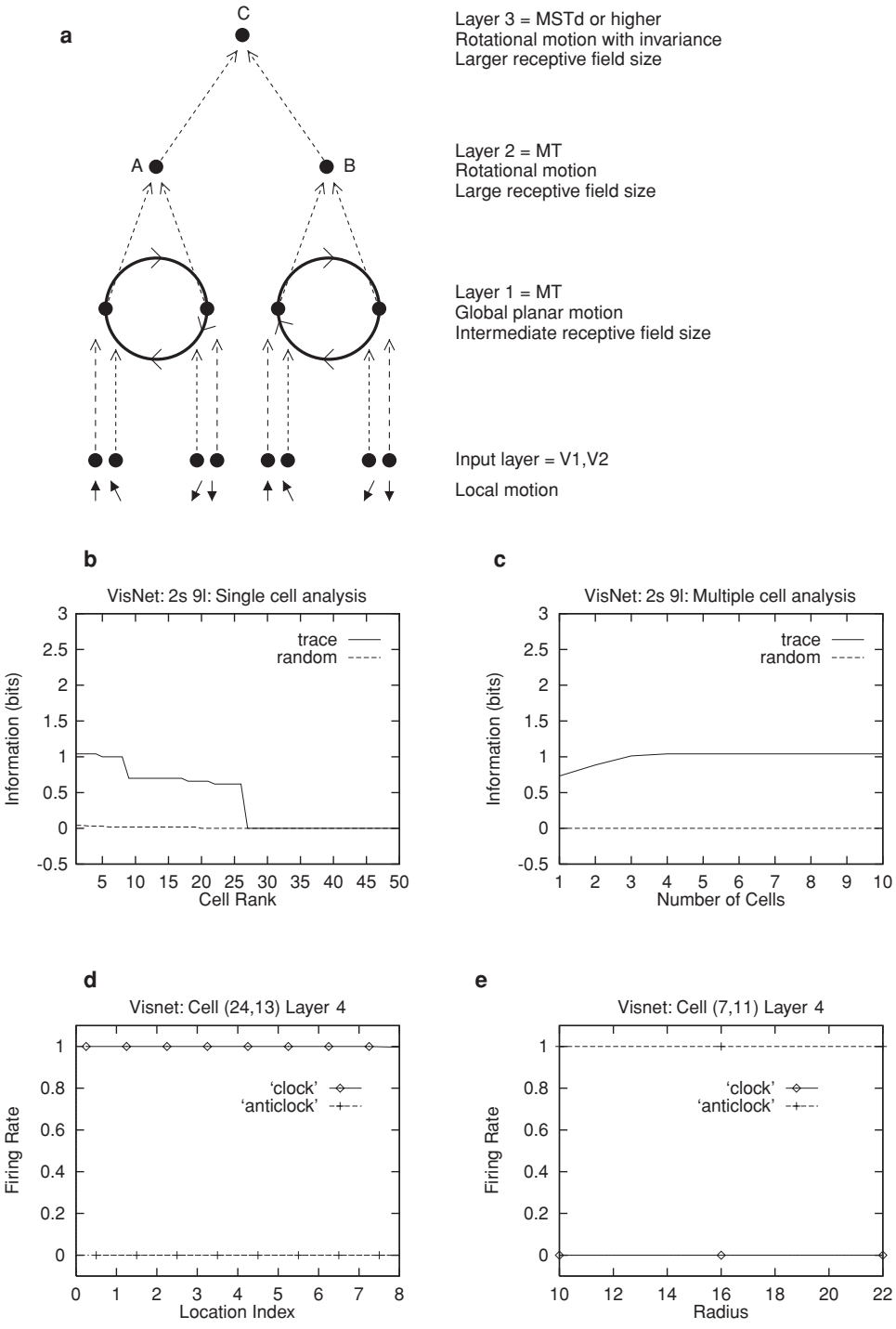
14.10 Invariant Global Object Motion in the Dorsal Visual System

A key issue in understanding the cortical mechanisms that underlie motion perception is how we perceive the motion of objects such as a rotating wheel invariantly with respect to position on the retina and size. For example, we perceive the wheel shown in Figure 14.9a rotating clockwise independently of its position on the retina. This occurs even though the local motion for the wheel in the different positions may be opposite. How could this invariance of the visual motion perception of objects arise in the visual system? Invariant motion representations are known to be developed in the cortical dorsal visual system. Motion-sensitive neurons in V1 have small receptive fields (in the range of 1–2 degrees at the fovea) and can therefore not detect global motion; this is part of the aperture problem (Wurtz and Kandel 2000). Neurons in MT, which receives inputs from V1 and V2, have larger receptive fields (e.g., 5 degrees at the fovea) and are able to respond to planar global motion, such as a field of small dots in which the majority (in practice as little as 55%) move in one direction, or to the overall direction of a moving plaid, the orthogonal grating components of which have motion at 45 degrees to the overall motion (Newsome et al. 1989; Wurtz and Kandel 2000). Further on in the dorsal visual system, some neurons in macaque visual area MST (but not MT) respond to rotating flow fields or looming with considerable translation invariance (Geesaman and Andersen 1996; Graziano et al. 1994).

In a unifying hypothesis with the design of the ventral cortical visual system, Rolls and Stringer (2006a) proposed that the dorsal visual system uses a hierarchical feedforward network architecture (V1, V2, MT, MSTd, parietal cortex) with training of the connections with a short-term memory trace associative synaptic modification rule to capture what is invariant at each stage. The principal difference from VisNet used to the model the ventral visual system is that the input filtering that for the ventral visual system uses difference of Gaussian filtering to produce V1 "oriented bar" simple cell-like receptive fields is replaced for the dorsal visual system by filtering of the image for the local direction and velocity of the optic flow (see Fig. 14.9). Simulations showed that the proposal is computationally feasible, in that invariant representations of the motion flow fields produced by objects self-organize in the later layers of the architecture (see Fig. 14.9). The model produces invariant representations of the motion flow fields produced by global in-plane motion of an object, in-plane rotational motion, looming versus receding of the object, and object-based rotation about a principal axis (Rolls and Stringer 2006a). Thus, the dorsal and ventral visual systems may share some similar computational principles.

14.11 Conclusion

An approach to the computations involved in object recognition that is very closely linked to the neurophysiology of object recognition has been described here and in more detail elsewhere (Rolls 2008b; Rolls and Deco 2002; Rolls and Stringer 2006b). The theory of how different types of invariance are learned is generic, and indeed the


a
C
A
B
Layer 3 = MSTd or higher
Rotational motion with invariance
Larger receptive field size
Layer 2 = MT
Rotational motion
Large receptive field size
Layer 1 = MT
Global planar motion
Intermediate receptive field size
Input layer = V1,V2
Local motion
b
VisNet: 2s 9l: Single cell analysis
trace
random
Information (bits)
Cell Rank
c
VisNet: 2s 9l: Multiple cell analysis
trace
random
Information (bits)
Number of Cells
d
Visnet: Cell (24,13) Layer 4
'clock'
'anticlock'
Firing Rate
Location Index
e
Visnet: Cell (7,11) Layer 4
'clock'
'anticlock'
Firing Rate
Radius

same architecture and model were used for all the investigations of translation, view, rotation, size, lighting, and object motion invariance described. In all cases, spatial and/or temporal continuity across the different transforms of individual objects was what allowed the architecture to learn invariant representations. The unifying nature of the overall approach is illustrated by the fact that the same architecture was able to form a model of invariant object motion representations in the dorsal visual system just by using local motion inputs instead of the oriented spatial filter inputs normally used for simulations of processing in the ventral visual system (Rolls and Stringer 2006a). Further, by adding top-down biasing inputs to the feedforward architecture, one can model the operation of top-down attentional effects (Deco and Rolls 2004; Rolls 2008b).

Although the architecture is generic, this does not, of course, preclude the local self-organization of topographic maps that provide some segregation of different types of processing. Indeed this is seen to arise from the short-range lateral excitatory and inhibitory connections that are characteristic of the neocortex and to provide the advantage of minimizing the lengths of the connections between neurons that need to exchange information, an important factor in the size of the brain (Rolls 2008b).

An important attribute of the architecture is the inhibitory connections that are important for implementing the competition between neurons. In conjunction with the probabilistic feedforward connectivity, these contribute to making the receptive fields asymmetric in complex scenes when there is increased competition, and thus support representations of multiple objects in a scene and their relative spatial position (Rolls et al. 2008).

The generic nature of the architecture has been further extended by the concept that adding a further layer or layers beyond the fourth layer of VisNet (which corresponds to the anterior inferior temporal visual cortex as shown in Fig. 14.1) and operating by the same principles provides a computational approach to understanding spatial scene representations in areas beyond the ventral visual stream, such as the parahippocampal cortex and hippocampus (Rolls et al. 2008). If such a fifth layer is trained with the same principles, but now with several objects present in a particular spatial arrangement, then the fifth layer learns by the same type of competitive learning to form neurons that respond to spatial views. In other words the fifth-layer neurons respond to the scene

←

Figure 14.9. *a*, Two rotating wheels at different locations rotating in opposite directions. The local flow field is ambiguous. Clockwise or counterclockwise rotation can only be diagnosed by a global flow computation, and it is shown how the network is expected to solve the problem to produce position invariant global motion-sensitive neurons. One rotating wheel is presented at any one time, but the need is to develop a representation of the fact that in the case shown the rotating flow field is always clockwise, independent of the location of the flow field. *b*, Single-cell information measures showing that some layer-4 neurons have perfect performance of 1 bit (clockwise vs. anticlockwise) after training with the trace rule, but not with random initial synaptic weights in the untrained control condition. *c*, The multiple cell information measures show that small groups of neurons have perfect performance. *d*, Position invariance illustrated for a single cell from layer 4, which responded only to the clockwise rotation, and for every one of the nine positions. *e*, Size invariance illustrated for a single cell from layer 4, which after training three different radii of rotating wheel, responded only to anticlockwise rotation, independently of the size of the rotating wheels. After Rolls and Stringer 2006a.

that is a combination of objects, and respond better than when the same objects are shown in a different spatial arrangement (Rolls et al. 2008). This is possible because the neurons in layer 4 become asymmetric in the presence of multiple objects in a scene. The fifth-layer neurons thus have properties like those of spatial view cells in the parahippocampal areas and hippocampus (Rolls 2008b; Rolls and Kesner 2006; Rolls and Xiang 2006).

The general approach is supported by other investigations. It has been shown, for example, that a somewhat comparable feedforward hierarchical feature combination architecture (although with a series of "simple" and "complex" cell layers in which a MAX function is used in the "complex" layers) can learn invariant representations (Riesenhuber and Poggio 2000; Serre et al. 2007). It has also been shown in an architecture trained by gradient descent that temporal continuity is an important principle that can allow invariant representations of objects to be learned from exposure to world-like series of images (Franzius et al. 2007; Wyss et al. 2006).

The overall approach to the computational architecture involved in forming invariant representations of objects is aimed to incorporate what is known about visual object processing in the brain and to provide a way to explore the computational bases of this processing. There are many issues that it would be of interest to explore further; one is training with more real-world–like sequences of images. The simulations performed so far have been with precisely produced image sets that allow particular hypotheses to be investigated and tested, in which typically one parameter is varied, such as spatial position. It would be interesting to extend this to image sequences in which several parameters might be altering simultaneously, such as spatial position and view. Training with images drawn from the natural world would be of interest, and this would allow the matching of the image statistics to the dynamical properties of the real visual system in the brain to be investigated. For example, do the slow times constants of NMDA receptors provide the system with sufficiently slow temporal properties to capture the invariant properties of objects seen in the real world given the rates at which objects transform, or is additional slowness needed, such as that that could be provided by the short-term memory attractor properties instantiated by the cortical recurrent collateral connections (Rolls 2008b)? Another aspect of the architecture that it would be of interest to explore further is its capacity for representing many objects and many transforms of each. The concept is that the capacity should be sufficiently high, given the very large number of neurons in the ventral visual system, and the fact that many stages of processing contribute to the invariant representations, with relatively local invariance of feature combinations in early layers, and object invariance in the final layers. However, the simulations performed have been on a small scale, with 1024 neurons in each of four layers, and it would be interesting to investigate how the system scales up (Rolls 2008b).

Acknowledgments

The author has worked on some of the investigations described here with N. Aggelopoulos, P. Azzopardi, G.C. Baylis, H. Critchley, G. Deco, P. Földiák, L. Franco, M. Hasselmo, J. Hornak, M. Kringelbach, C.M. Leonard, T.J. Milward, D.I. Perrett, S.M.

Stringer, M.J. Tovee, T. Trappenberg, A. Treves, J. Tromans, and G.M. Wallis. Their collaboration is sincerely acknowledged. Discussions on inattentional blindness with Rebekah White were very helpful. Different parts of the research described were supported by the Medical Research Council, PG8513790; by a Human Frontier Science Program grant; by an EC Human Capital and Mobility grant; by the MRC Oxford Interdisciplinary Research Centre in Cognitive Neuroscience; and by the Oxford McDonnell-Pew Centre in Cognitive Neuroscience.

Bibliography

Aggelopoulos NC, Franco L, Rolls ET. 2005. Object perception in natural scenes: encoding by inferior temporal cortex simultaneously recorded neurons. *J Neurophysiol* 93: 1342–1357.

Aggelopoulos NC, Rolls ET. 2005. Natural scene perception: inferior temporal cortex neurons encode the positions of different objects in the scene. *Eur J Neurosci* 22: 2903–2916.

Ballard DH. 1990. Animate vision uses object-centred reference frames. In *Advanced neural computers*, ed. R Eckmiller, 229–236. Amsterdam: North-Holland.

Ballard DH. 1993. Subsymbolic modelling of hand-eye coordination. In *The simulation of human intelligence*, ed. DE Broadbent, 71–102. Oxford: Blackwell.

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

Biederman I. 1987. Recognition-by-components: a theory of human image understanding. *Psychol Rev* 94: 115–147.

Booth MCA, Rolls ET. 1998. View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cereb Cortex* 8: 510–523.

Boussaoud D, Desimone R, Ungerleider LG. 1991. Visual topography of area TEO in the macaque. *J Comp Neurol* 306: 554–575.

Brunel N, Wang XJ. 2001. Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. *J Comput Neurosci* 11: 63–85.

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

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

Cowey A, Rolls ET. 1975. Human cortical magnification factor and its relation to visual acuity. *Exp Brain Rese* 21: 447–454.

Deco G, Lee TS. 2002. A unified model of spatial and object attention based on inter-cortical biased competition. *Neurocomput* 44–46: 775–781.

Deco G, Rolls ET. 2002. Object-based visual neglect: a computational hypothesis. *Euro J Neurosci* 16: 1994–2000.

Deco G, Rolls ET. 2003. Attention and working memory: a dynamical model of neuronal activity in the prefrontal cortex. *Eur J Neurosci* 18: 2374–2390.

Deco G, Rolls ET. 2004. A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Res* 44: 621–644.

Deco G, Rolls ET. 2005a. Attention, short-term memory, and action selection: a unifying theory. *Prog Neurobiol* 76: 236–256.

Deco G, Rolls ET. 2005b. Neurodynamics of biased competition and co-operation for attention: a model with spiking neurons. *J Neurophysiol* 94: 295–313.

Deco G, Rolls ET. 2005c. Sequential memory: a putative neural and synaptic dynamical mechanism. *J Cognit Neurosci* 17: 294–307.

Deco G, Rolls ET. 2005d. Synaptic and spiking dynamics underlying reward reversal in orbitofrontal cortex. *Cereb Cortex* 15: 15–30.

Deco G, Rolls ET. 2006. Decision-making and Weber's Law: a neurophysiological model. *Eur J Neurosci* 24: 901–916.

Deco G, Rolls ET, Horwitz B. 2004. "What" and "where" in visual working memory: a computational neurodynamical perspective for integrating fMRI and single-neuron data. *J Cognit Neurosci* 16: 683–701.

Deco G, Rolls ET, Zihl J. 2005. A neurodynamical model of visual attention. In *Neurobiology of attention*, ed. L Itti, G Rees, and J Tsotos, 593–599. San Diego: Elsevier.

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

Desimone R, Duncan J. 1995. Neural mechanisms of selective visual attention. *Ann Rev Neurosci* 18: 193–222.

Elliffe MCM, Rolls ET, Stringer SM. 2002. Invariant recognition of feature combinations in the visual system. *Biol Cyber* 86: 59–71.

Földiák P. 1991. Learning invariance from transformation sequences. *Neural Comput* 3: 194–200.

Franco L, Rolls ET, Aggelopoulos NC, Jerez JM. 2007. Neuronal selectivity, population sparseness, and ergodicity in the inferior temporal visual cortex. *Biol Cyber* 96: 547–560.

Franzius M, Sprekeler H, Wiskott L. 2007. Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLoS Comput Biol* 3: e166.

Fukushima K. 1980. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cyber* 36: 193–202.

Fukushima K. 1989. Analysis of the process of visual pattern recognition by the neocognitron. *Neural Netw* 2: 413–420.

Fukushima K. 1991. Neural networks for visual pattern recognition. *IEEE Trans* 74: 179–190.

Geesaman BJ, Andersen RA. 1996. The analysis of complex motion patterns by form/cue invariant MSTd neurons. *J Neurosci* 16: 4716–4732.

Graziano MSA, Andersen RA, Snowden RJ. 1994. Tuning of MST neurons to spiral motions. *J Neurosci* 14: 57–64.

Gregory RL. 1970. *The intelligent eye*. New York: McGraw-Hill.

Gregory RL. 1998. *Eye and brain*. Oxford: Oxford University Press.

Hasselmo ME, Rolls ET, Baylis GC, Nalwa V. 1989. Object-centred encoding by face-selective neurons in the cortex in the superior temporal sulcus of the the monkey. *Exp Brain Res* 75: 417–429.

Hegde J, Van Essen DC. 2000. Selectivity for complex shapes in primate visual area V2. *J Neurosci* 20: RC61.

Helmholtz Hv. 1857. *Handbuch der physiologischen optik*. Leipzig: Voss.

Ito M, Komatsu H. 2004. Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. *J Neurosci* 24: 3313–3324.

Koenderink JJ, Van Doorn AJ. 1979. The internal representation of solid shape with respect to vision. *Biol Cyber* 32: 211–217.

Logothetis NK, Pauls J, Bülthoff HH, Poggio T. 1994. View-dependent object recognition by monkeys. *Curr Biol* 4: 401–414.

Maier A, Logothetis NK, Leopold DA. 2005. Global competition dictates local suppression in pattern rivalry. *J Vision* 5: 668–677.

Marr D. 1982. *Vision*. San Francisco: WH Freeman.

Martinez-Trujillo J, Treue S. 2002. Attentional modulation strength in cortical area MT depends on stimulus contrast. *Neuron* 35: 365–370.

Mozer M. 1991. *The perception of multiple objects: a connectionist approach*. Cambridge: MIT Press.

Newsome WT, Britten KH, Movshon JA. 1989. Neuronal correlates of a perceptual decision. *Nature* 341: 52–54.

Perrett D, Mistlin A, Chitty A. 1987. Visual neurons responsive to faces. *Trends Neurosci* 10: 358–364.

Perrett DI, Rolls ET, Caan W. 1982. Visual neurons responsive to faces in the monkey temporal cortex. *Exp Brain Res* 47: 329–342.

Perry G, Rolls ET, Stringer SM. 2006. Spatial vs. temporal continuity in view invariant visual object recognition learning. *Vision Res* 46: 3994–4006.

Perry G, Rolls ET, Stringer SM. 2009. Continuous transformation learning of translation invariant representations. (in press).

Poggio T, Edelman S. 1990. A network that learns to recognize three-dimensional objects. *Nature* 343: 263–266.

Renart A, Moreno R, de la Rocha J, Parga N, Rolls ET. 2001. A model of the IT-PF network in object working memory which includes balanced persistent activity and tuned inhibition. *Neurocomputing* 38–40: 1525–1531.

Renart A, Parga N, Rolls ET. 2000. A recurrent model of the interaction between the prefrontal cortex and inferior temporal cortex in delay memory tasks. In *Advances in neural information processing systems*, ed. SA Solla, TK Leen, and K-R Mueller, 171–177. Cambridge: MIT Press.

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

Reynolds JH, Chelazzi L, Desimone R. 1999. Competitive mechanisms subserve attention in macaque areas V2 and V4. *J Neurosci* 19: 1736–1753.

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

Rolls ET. 1989a. Functions of neuronal networks in the hippocampus and neocortex in memory. In *Neural models of plasticity: experimental and theoretical approaches*, ed. JH Byrne, and WO Berry, 240–265. San Diego: Academic Press.

Rolls ET. 1989b. The representation and storage of information in neuronal networks in the primate cerebral cortex and hippocampus. In *The computing neuron*, ed. R Durbin, C Miall, and G Mitchison, 125–129. Wokingham, England: Addison-Wesley.

Rolls ET. 1992. Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philos Trans R Soc Lond B* 335: 11–21.

Rolls ET. 1999. *The brain and emotion*. Oxford: Oxford University Press.

Rolls ET. 2000. Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron* 27: 205–218.

Rolls ET. 2005. *Emotion explained*. Oxford: Oxford University Press.

Rolls ET. 2007a. A computational neuroscience approach to consciousness. *Neural Netw* 20: 962–982.

Rolls ET. 2007b. The representation of information about faces in the temporal and frontal lobes. *Neuropsychol* 45: 125–143.

Rolls ET. 2008a. Face processing in different brain areas, and critical band masking. *J Neuropsychol* 2: 325–360.

Rolls ET. 2008b. *Memory, attention, and decision-making: a unifying computational neuroscience approach*. Oxford: Oxford University Press.

Rolls ET. 2008c. Top-down control of visual perception: attention in natural vision. *Perception* 37: 333–354.

Rolls ET, Aggelopoulos NC, Zheng F. 2003. The receptive fields of inferior temporal cortex neurons in natural scenes. *J Neurosci* 23: 339–348.

Rolls ET, Cowey A. 1970. Topography of the retina and striate cortex and its relationship to visual acuity in rhesus monkeys and squirrel monkeys. *Exp Brain Res* 10: 298–310.

Rolls ET, Deco G. 2002. *Computational neuroscience of vision*. Oxford: Oxford University Press.

Rolls ET, Deco G. 2006. Attention in natural scenes: neurophysiological and computational bases. *Neural Netw* 19: 1383–1394.

Rolls ET, Franco L, Aggelopoulos NC, Perez JM. 2006. Information in the first spike, the order of spikes, and the number of spikes provided by neurons in the inferior temporal visual cortex. *Vision Res* 46: 4193–4205.

Rolls ET, Kesner RP. 2006. A computational theory of hippocampal function, and empirical tests of the theory. *Prog Neurobiol* 79: 1–48.

Rolls ET, 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 Comput* 12: 2547–2572.

Rolls ET, Stringer SM. 2001. Invariant object recognition in the visual system with error correction and temporal difference learning. *Network: Comput Neural Syst* 12: 111–129.

Rolls ET, Stringer SM. 2006a. Invariant global motion recognition in the dorsal visual system: a unifying theory. *Neural Comput* 19: 139–169.

Rolls ET, Stringer SM. 2006b. Invariant visual object recognition: a model, with lighting invariance. *J Physiol Paris* 100: 43–62.

Rolls ET, Tovee MJ. 1995. Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *J Neurophysiol* 73: 713–726.

Rolls ET, Tovee MJ, Purcell DG, Stewart AL, Azzopardi P. 1994. The responses of neurons in the temporal cortex of primates, and face identification and detection. *Exp Brain Res* 101: 473–484.

Rolls ET, Treves A. 1998. *Neural networks and brain function*. Oxford: Oxford University Press.

Rolls ET, Treves A, Tovee MJ. 1997. The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. *Exp Brain Res* 114: 177–185.

Rolls ET, Tromans J, Stringer SM. 2008. Spatial scene representations formed by self-organizing learning in a hippocampal extension of the ventral visual system. *Euro J Neurosci* 28: 2116–2127.

Rolls ET, Xiang J-Z. 2006. Spatial view cells in the primate hippocampus, and memory recall. *Rev Neurosci* 17: 175–200.

Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T. 2007. Robust object recognition with cortex-like mechanisms. *IEEE Tr Pattern Anal Mach Intell* 29: 411–426.

Sheinberg DL, Logothetis NK. 2001. Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. *J Neurosci* 21: 1340–1350.

Simons DJ, Chabris CF. 1999. Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception* 28: 1059–1074.

Simons DJ, Rensink RA. 2005. Change blindness: past, present, and future. *Trends Cogn Sci* 9: 16–20.

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

Stringer SM, Perry G, Rolls ET, Proske JH. 2006. Learning invariant object recognition in the visual system with continuous transformations. *Biol Cyber* 94: 128–142.

Stringer SM, Rolls ET. 2000. Position invariant recognition in the visual system with cluttered environments. *Neural Netw* 13: 305–315.

Stringer SM, Rolls ET. 2002. Invariant object recognition in the visual system with novel views of 3D objects. *Neural Comput* 14: 2585–2596.

Stringer SM, Rolls ET. 2008. Learning transform invariant object recognition in the visual system with multiple stimuli present during training. *Neural Netw* 21: 888–903.

Stringer SM, Rolls ET, Tromans J. 2007. Invariant object recognition with trace learning and multiple stimuli present during training. *Network: Comput Neural Syst* 18: 161–187.

Sutton RS, Barto AG. 1998. *Reinforcement learning*. Cambridge: MIT Press.

Szabo M, Almeida R, Deco G, Stetter M. 2004. Cooperation and biased competition model can explain attentional filtering in the prefrontal cortex. *Euro J Neurosci* 19: 1969–1977.

Tanaka K, Saito C, Fukada Y, Moriya M. 1990. Integration of form, texture, and color information in the inferotemporal cortex of the macaque. In *Vision, memory and the temporal lobe*, ed. E Iwai, and M Mishkin, 101–109. New York: Elsevier.

Tovee MJ, Rolls ET. 1995. Information encoding in short firing rate epochs by single neurons in the primate temporal visual cortex. *Visual Cogn* 2: 35–58.

Trappenberg TP, Rolls ET, Stringer SM. 2002. Effective size of receptive fields of inferior temporal cortex neurons in natural scenes. In *Advances in neural information processing systems 14*, ed. TG Dietterich, S Becker, Z Ghahramani, 293–300. Cambridge: MIT Press.

Treves A, Panzeri S, Rolls ET, Booth M, Wakeman EA. 1999. Firing rate distributions and efficiency of information transmission of inferior temporal cortex neurons to natural visual stimuli. *Neural Comput* 11: 611–641.

Treves A, Rolls ET. 1994. A computational analysis of the role of the hippocampus in memory. *Hippocampus* 4: 374–391.

Ullman S. 1996. *High-level vision: object recognition and visual cognition*. Cambridge: Bradford/ MIT Press.

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

Wallis G, Rolls ET. 1997. Invariant face and object recognition in the visual system. *Prog Neurobiol* 51: 167–194.

Wallis G, Rolls ET, Földiák P. 1993. Learning invariant responses to the natural transformations of objects. In *International joint conference on neural networks*, vol 2: 1087–1090.

Wang XJ. 2002. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron* 36: 955–968.

Wurtz RH, Kandel ER. 2000. Perception of motion depth and form. In *Principles of neural science*, ed. ER Kandel, JH Schwartz, and TM Jessell, 548–571. New York: McGraw-Hill.

Wyss R, Konig P, Verschure PF. 2006. A model of the ventral visual system based on temporal stability and local memory. *PLoS Biol* 4: e120.

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