EJN
European Journal of Neuroscience

European Journal of Neuroscience, pp. 1–13, 2008 doi:10.1111/j.1460-9568.2008.06486.x

Spatial scene representations formed by self-organizing learning in a hippocampal extension of the ventral visual system

Edmund T. Rolls, James M. Tromans and Simon M. Stringer
Department of Experimental Psychology, Centre for Computational Neuroscience, Oxford University, South Parks Road, Oxford OX1 3UD, UK

Keywords: hippocampus, inferior temporal visual cortex, invariant object recognition, spatial view cells

Abstract

We show in a unifying computational approach that representations of spatial scenes can be formed by adding an additional self-organizing layer of processing beyond the inferior temporal visual cortex in the ventral visual stream without the introduction of new computational principles. The invariant representations of objects by neurons in the inferior temporal visual cortex can be modelled by a multilayer feature hierarchy network with feedforward convergence from stage to stage, and an associative learning rule with a short-term memory trace to capture the invariant statistical properties of objects as they transform over short time periods in the world. If an additional layer is added to this architecture, training now with whole scenes that consist of a set of objects in a given fixed spatial relation to each other results in neurons in the added layer that respond to one of the trained whole scenes but do not respond if the objects in the scene are rearranged to make a new scene from the same objects. The formation of these scene-specific representations in the added layer is related to the fact that in the inferior temporal cortex and, we show, in the VisNet model, the receptive fields of inferior temporal cortex neurons shrink and become asymmetric when multiple objects are present simultaneously in a natural scene. This reduced size and asymmetry of the receptive fields of inferior temporal cortex neurons also provides a solution to the representation of multiple objects, and their relative spatial positions, in complex natural scenes.

Introduction

We propose and analyse a unifying hypothesis of the relation between the ventral visual system and hippocampal spatial representations. We show that principles that can account for the formation of neurons with invariant responses in the inferior temporal visual cortex can also account for the formation of neurons with responses specific to spatial views in the hippocampus and related areas.

Over successive stages the visual system develops neurons that respond with view-, size- and position (translation)-invariance to objects or faces (Desimone, 1991; Tanaka *et al.*, 1991; Rolls, 1992, 2000; Rolls & Deco, 2002). The inferior temporal visual cortex has neurons that respond to faces and objects with translation- (Kobatake & Tanaka, 1994; Tovee *et al.*, 1994; Ito *et al.*, 1995; Op De Beeck & Vogels, 2000), size- (Rolls & Baylis, 1986; Ito *et al.*, 1995), and view (Hasselmo *et al.*, 1989; Booth & Rolls, 1998)-invariance (Rolls, 2008b). It is crucially important that the visual system builds invariant representations for only then can one-trial learning about an object generalize usefully to other transforms of the same object (Rolls & Deco, 2002; Rolls, 2005). Building invariant representations of objects is a major computational issue, and the means by which the cerebral cortex solves this problem is a topic of great interest (Biederman, 1987; Rolls, 1992, 2008b; Ullman, 1996; Riesenhuber & Poggio, 1999; Rolls & Deco, 2002; Wiskott & Sejnowski, 2002; Rolls & Stringer, 2006b; Wyss *et al.*, 2006).

The concept that recognition memory (measured for example in delayed match-to-sample tasks with objects with overlapping feature subsets) may be implemented in areas added to the ventral visual stream beyond the inferior temporal visual cortex (Mishkin *et al.*, 1997; Bussey *et al.*, 2002, 2003, 2005; Bussey & Saksida, 2005, 2007; Buckley & Gaffan, 2006) has been investigated in previous work by adding a layer corresponding to the perirhinal cortex which forms nodes that respond to combinations of the inputs it receives by a form of competitive learning (Cowell *et al.*, 2006). In that connectionist model, the nodes respond to feature combinations and thus assist performance when this depends on objects being defined by feature combinations. However, there is no concept or representation of space in that approach as only features are represented, with no representation of the relative spatial position of features or objects. In the discussion of that connectionist model, it was suggested (Cowell *et al.*, 2006) that 'tasks like repeating-items DMS (delayed matching to sample) merely provide an additional degree of' (object) 'ambiguity that must be resolved by even more complex conjunctive representations in a hierarchy that extends throughout the ventral visual stream through perirhinal cortex and on into other structures such as the hippocampus'.

Correspondence: E. T. Rolls, as above.
E-mail: edmund.rolls@oxcns.org; url:http://www.oxcns.org

Received 2 March 2008, revised 20 August 2008, accepted 4 September 2008

In contrast, the present hypotheses and model address the formation of spatial representations such as those provided by spatial-view neurons (Georges-François *et al.*, 1999; Rolls & Xiang, 2006) in areas such as the parahippocampal cortex and hippocampus, show how they may be related to the representation of multiple objects simultaneously in the inferior temporal visual cortex, and address the fundamental issue of how inferior temporal cortex representations can display considerable translation invariance yet still support the formation of spatial representations in the hippocampus.

Materials and methods

Hypotheses on the formation of spatial representations in cortical areas beyond the inferior temporal visual cortex

It is now possible to propose a unifying hypothesis of the relation between the ventral visual system and hippocampal spatial representations (Rolls, 2008b). The spatial representations found in the primate parahippocampal cortex and hippocampus include spatial-view neurons. A spatial-view neuron responds when a primate looks at a part of a spatial scene that is characterized by a set of features or objects that are in a fixed spatial relation to each other (Rolls *et al.*, 1997a, 1998, 2005; Robertson *et al.*, 1998; Georges-François *et al.*, 1999; Rolls & Kesner, 2006; Rolls & Xiang, 2006).

Consider a computational architecture in which an additional layer is added to a feature hierarchy model of the ventral visual system, VisNet, as illustrated in Fig. 1. In the anterior inferior temporal visual cortex, which corresponds to the fourth layer of VisNet, neurons respond to objects or faces (Rolls, 2008a,b). However, in complex natural scenes the receptive fields of these neurons decrease in size (Trappenberg *et al.*, 2002; Rolls *et al.*, 2003), and several objects close to the fovea (within approximately 10°) can be represented because many object-tuned neurons have spatially asymmetric receptive fields with respect to the fovea (Aggelopoulos & Rolls, 2005). In complex natural scenes, inferior temporal cortex neurons thus convey information not only about objects but also about their spatial position. If the additional ('hippocampal') layer added to the four-layer VisNet architecture performs the same operation as previous layers, it is predicted to form, by its self-organizing learning when trained on scenes consisting of a fixed spatial arrangement of a set of objects, neurons that respond to combinations of objects in the scene with the positions of the objects relative spatially to each other incorporated into the representation. That is, it is predicted that spatial-view neurons will be formed in the added layer (cf. de Araujo *et al.*, 2001). This is tested here by investigating whether the neurons in the hippocampal layer respond to whole spatial scenes in which the relative positions of the objects in the scene are encoded. This is measured by comparing the responses of neurons in the hippocampal layer to the trained scenes with the responses of the same neurons using different scenes comprised of exactly the same sets of objects but in different spatial positions relative to each other. The hypothesis is also tested by comparing the responses of the hippocampal layer neurons to single objects, with the prediction being that the scene-specific hippocampal layer representations should be more responsive to the whole scenes than to the single objects that form part of the scenes.

The model also allowed us to test the prediction that the receptive fields of inferior temporal cortex neurons become small and spatially asymmetric in complex natural scenes because of competition introduced when several objects were present simultaneously, and the probabilistic diluted connectivity of the different neurons providing the potential for spatial asymmetry. In particular, the model as far as the inferior temporal visual cortex was trained on single objects, each shown in four different positions, to determine that many of the neurons showed object-selective but spatially invariant receptive fields. Then, without further training, four objects were presented simultaneously, and we measured whether the object-selective neurons

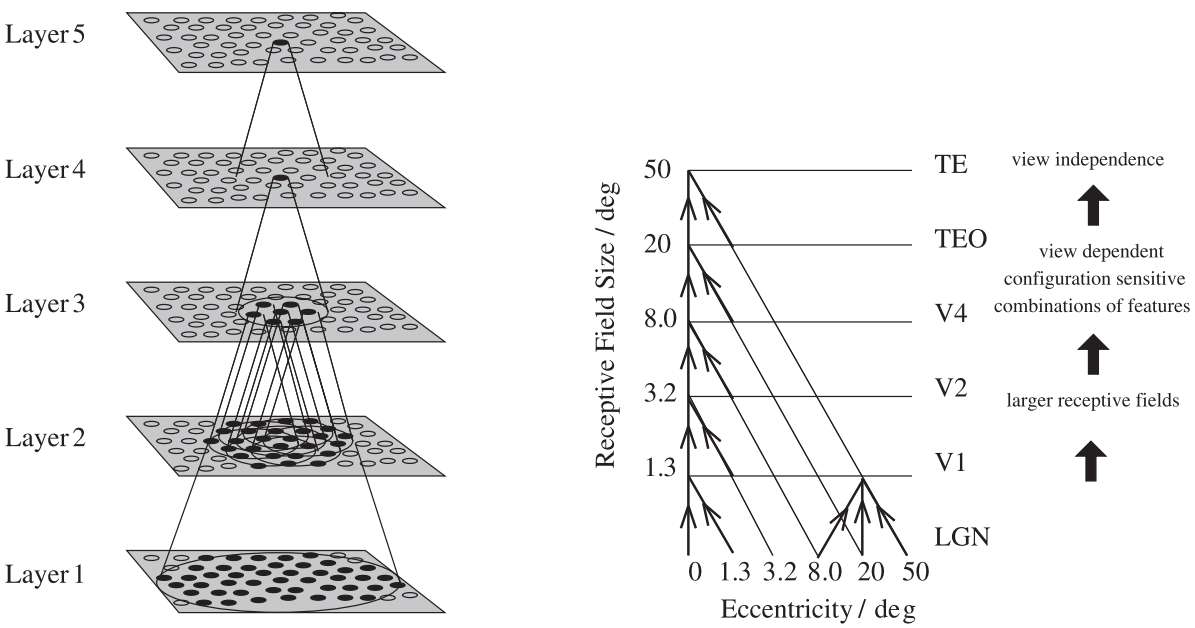

FIG. 1. Adding a fifth layer, corresponding to the parahippocampal gyrus–hippocampal system, after the inferior temporal visual cortex (corresponding to layer 4 in this diagram) may lead to the self-organization of spatial-view and/or place cells in layer 5 when whole scenes are presented (see text). Convergence in the visual system is shown in the earlier layers. (Right) As it occurs in the brain. V1, visual cortex area V1; TEO, posterior inferior temporal cortex; TE, inferior temporal cortex. (Left) In this diagram, layer 1 corresponds to V2, layer 2 to V4, layer 3 to posterior inferior temporal cortex (TEO), layer 4 to anterior inferior temporal cortex (TE) and layer 5 to the parahippocampal–hippocampal areas. Convergence through the network is designed to provide fourth-layer neurons with information from across the entire input retina.

showed less spatial invariance to the object to which they were tuned when four objects were being presented.

An important implication of the discovery that the receptive fields of individual inferior temporal cortex neurons can be large (e.g. 70° in diameter) when the effective object is presented against a blank background but much smaller, and spatially asymmetric, when the object is presented in a natural scene (Rolls *et al.*, 2003) or a scene with five different objects present (Aggelopoulos & Rolls, 2005) is that several objects can be represented in a scene, together with information about their position relative to the fovea, by the responses of inferior temporal cortex neurons. In a complex scene, different neurons are active to encode different objects with ensemble encoding (Rolls, 2008b), but each neuron that is active provides information about the spatial position of the object due to the asymmetry of the receptive fields, so that an object-selective neuron will only be firing in a complex scene if its object is in some but not other positions with respect to the fovea (Aggelopoulos & Rolls, 2005). This provides an important solution to the otherwise computationally difficult problem of representing multiple objects in a scene in a network with distributed representations (Mozer, 1991). The hypothesis presented for the mechanism for the reduction of the receptive field in complex scenes, and the asymmetry of the receptive field in complex scenes, is as follows (Rolls, 2008b). Consider that the forward connectivity from one layer of the visual system to the next (e.g. V4 to posterior inferior temporal cortex) has an approximately Gaussian shape as illustrated in Fig. 1 (left), with many connections to a neuron from a corresponding topographic position in visual space in the preceding layer and gradually fewer connections as one moves away from the corresponding position in the preceding layer. The result of this is that, when multiple objects are presented simultaneously, there will be larger feedback inhibition from the inhibitory neurons and this will reduce the size of the receptive field by making only the object when close to the centre of the receptive field produce sufficient activation to produce neuronal firing, as analyzed elsewhere (Trappenberg *et al.*, 2002; Deco & Rolls, 2004). Consider further that the forward connectivity will be sparse (or diluted), with the probability that a given neuron receives from a particular neuron in the preceding layer, even from a corresponding position, quite low, the order of 0.1 (Abeles, 1991; Braitenberg & Schütz, 1991; Rolls & Deco, 2002; Rolls, 2008b). The probabilistic nature of this connectivity will mean that by chance a given neuron may receive more connections from some parts of the generally Gaussian shape of the connectivity distribution than from other parts, and the result of this when the inhibition is raised with multiple stimuli present is that spatial asymmetries will become more evident (Rolls, 2008b). This hypothesis has not been tested computationally by simulation in a network with the appropriate diluted forward connectivity and, in the simulations described here, we perform this investigation, using VisNet which has probabilistic and approximately spatially Gaussian forward connectivity, as described below.

Normally, the layers of VisNet up to the inferior temporal visual cortex are trained sequentially, with layer one (corresponding to V2) trained first, and the last layer (corresponding to the inferior temporal visual cortex) trained last. The rationale for this is that there is little point in training the inferior temporal visual cortex until the early layers have been trained, so that their responses have become selective and stable. The biological correspondent and, we believe, the reason for this (Rolls, 2008b) is that there tends to be a critical period for plasticity in early visual cortical areas while the system becomes tuned to the environment (Hooks & Chen, 2007) whereas areas such as the inferior temporal visual cortex remain plastic throughout life so that the representations of new objects can be learned (Rolls *et al.*, 1989; Tovee *et al.*, 1996; Dolan *et al.*, 1997; Rolls, 2008b). This sequential training procedure was used in the current investigation, but we performed a check that if all the layers of VisNet as far as the inferior temporal visual cortex were trained simultaneously, neurons with similar properties were formed in the inferior temporal visual cortex layer, the procedure just being less efficient for the reasons given.

It is inherent in a system that learns to form invariant representations of objects that considerable training must be given, for it is only by being provided with different exemplars of the different objects (e.g. different positions, views, etc) that the system can learn to separate from the exemplars the inputs that correspond to one object in its different transforms from the inputs that correspond to different objects in their different transforms. Many trials of training are thus given to VisNet up to the stage of the inferior temporal visual cortex. However, hippocampal representations of new scenes and potentially of episodic memories must be capable of being formed quickly, in one or a few trials, as this is a property of the learning of new spatial scenes and episodic memories (Dere *et al.*, 2009; Rolls, 2009). We therefore tested whether, once VisNet had been trained to the inferior temporal cortex level on individual objects, the new learning of spatial scene representations in the hippocampal layer comprising of a particular arrangement of these objects could be learned in a small number of training trials.

Some of the crucial points we were able to show in relation to these hypotheses were the following. First, once a hierarchical architecture of the type described has been trained on single objects to form invariant representations of them, one can then obtain spatial scene representations by training the next layer on scenes consisting of particular spatial arrangements of the objects. Second, this spatial scene learning by the added layer can be fast. Third, when multiple objects are present in the scene, this produces shrinkage and asymmetry of the receptive field sizes in the object layer, which helps spatially specific scene representations to be formed. Fourth, the shrinkage and asymmetry of the receptive fields of the object layer neurons when multiple objects are present simultaneously provides a computational solution to the problem of how multiple objects can be represented simultaneously in the inferior temporal visual cortex, with the correct spatial positions of the objects encoded.

Experimental design

In this paper we describe a simulation to test the predictions of the hypothesis described in the preceding section, with VisNet simulations with conceptually a fifth layer added and appropriate fixed-object combinations in the training set to represent spatial views. In related work, a more artificial network trained by gradient ascent with a goal function that included forming relatively time-invariant representations and decorrelating the responses of neurons in the multilayer network, place-like cells were formed at the end of the network when the system was trained with a real or simulated robot moving through spatial environments (Wyss *et al.*, 2006), and view cells were formed when training on video sequences from a virtual-reality environment (Franzius *et al.*, 2007). In this paper we test whether spatial-view cells develop in the equivalent of a VisNet fifth layer if trained with spatial scenes. The utility of testing this with a VisNet-like architecture is that this architecture embodies a biologically plausible implementation based on neuronally plausible competitive learning and a short-term memory trace associative local learning rule.

The VisNet architecture

The model architecture (VisNet) implemented by Wallis & Rolls (1997) and Rolls & Milward (2000) that is used to investigate the

properties of object and scene learning in this paper is based on the following. (i) A series of hierarchical competitive networks with local graded lateral inhibition. (ii) Convergent connections to each neuron from a topologically corresponding region of the preceding layer, leading to an increase in the receptive field size of neurons through the visual processing areas. (iii) Synaptic plasticity based on a Hebb-like learning rule. Model simulations which incorporated these hypotheses with a modified associative learning rule to incorporate a short-term memory trace of previous neuronal activity were shown to be capable of producing stimulus-selective but translation- and view-invariant representations (Wallis & Rolls, 1997; Rolls & Milward, 2000; Rolls & Stringer, 2001, 2006b; Elliffe *et al.*, 2002; Stringer *et al.*, 2007).

The model used here consists of a hierarchical series of four layers of competitive networks, corresponding to V2, V4, the inferior temporal cortex, and the parahippocampal areas and hippocampus. In the diagram shown in Fig. 1, layer 5, the parahippocampal–hippocampal layer, corresponds to layer 4 of the actual network simulated. The inferior temporal cortex corresponds to layer 3 of the network simulated. This is a little different from usual simulations with this architecture, in which layer 3 can be thought to correspond to the posterior inferior temporal visual cortex and layer 4 to the anterior inferior temporal cortex (Wallis & Rolls, 1997; Rolls & Stringer, 2006b). For these simulations, the posterior and anterior inferior temporal visual cortex were combined into a single inferior temporal visual cortex layer. The change was made for computational efficiency. When we refer to layer 3 in this paper, that indicates the inferior temporal visual cortex layer of the simulation. When we refer to layer 4 in the remainder of this paper, that indicates the parahippocampal–hippocampal layer of the simulation. The forward connections to individual cells are derived from a topologically corresponding region of the preceding layer, using a Gaussian distribution of connection probabilities. These distributions are defined by a radius which will contain ~67% of the connections from the preceding layer. The values used are given in Table 1. This diluted (incomplete, and probabilistically set up) feed-forward connectivity is important in some of the effects on the asymmetry of the receptive fields described in this paper that occur when inhibition is high in crowded scenes.

Before stimuli are presented to the network's input layer they are pre-processed by a set of input filters that accord with the general tuning profiles of simple cells in V1. The input filters used are computed by weighting the difference of two Gaussians by a third orthogonal Gaussian according to the following:

$$\Gamma_{xy}(\rho,\theta,f) = \rho\left[\mathrm{e}^{-\left(\frac{x\cos\theta+y\sin\theta}{\sqrt{2}/f}\right)^2} - \frac{1}{1.6}\mathrm{e}^{-\left(\frac{x\cos\theta+y\sin\theta}{1.6\sqrt{2}/f}\right)^2}\right]\mathrm{e}^{-\left(\frac{x\sin\theta-y\cos\theta}{3\sqrt{2}/f}\right)^2} \quad (1)$$

where f is the filter spatial frequency, θ is the filter orientation and ρ is the sign of the filter, i.e. ± 1. Individual filters are tuned to spatial frequency (0.0625 to 0.5 cycles per pixel); orientation (0° to 135° in steps of 45°); and sign (± 1). The number of layer 1 connections to each spatial frequency filter group is given in Table 2.

Table 1. Network dimensions showing the number of connections per neuron and the radius in the preceding layer from which 67% are received

	Dimensions	Number of connections	Radius
Layer 4	32 × 32	100	12
Layer 3	32 × 32	100	9
Layer 2	32 × 32	100	6
Layer 1	32 × 32	272	6
Retina	128 × 128 × 32	–	–

Table 2. Layer 1 connectivity

Frequency	0.5	0.25	0.125	0.0625
Number of connections	201	50	13	8

The numbers of connections from each spatial frequency set of filters are shown. The spatial frequency is in cycles per pixel.

The activation h_i of each neuron i in the network is set equal to a linear sum of the inputs y_j from afferent neurons j weighted by the synaptic weights w_{ij}. That is,

$$h_i = \sum_j w_{ij} y_j \quad (2)$$

where y_j is the firing rate of neuron j and w_{ij} is the strength of the synapse from neuron j to neuron i.

Within each layer, competition is graded rather than winner-take-all, and is implemented in two stages. First, to implement lateral inhibition the activations h of neurons within a layer are convolved with a spatial filter, I, where δ controls the contrast and σ controls the width, and a and b index the distance away from the centre of the filter

$$I_{a,b} = \begin{cases} -\delta\,\mathrm{e}^{-\frac{a^2+b^2}{\sigma^2}} & \text{if } a \neq 0 \text{ or } b \neq 0, \\ 1 - \sum_{\substack{a \neq 0 \\ b \neq 0}} I_{a,b} & \text{if } a = 0 \text{ and } b = 0 \end{cases} \quad (3)$$

The lateral inhibition parameters are given in Table 3. We note that lateral inhibition is a property of cortical, including visual, processing, and is implemented in the brain by inhibitory interneurons that operate within a localized area of the cortex (Rolls & Deco, 2002; Rolls, 2008b).

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

$$y = f^{\text{sigmoid}}(r) = \frac{1}{1+\mathrm{e}^{-2\beta(r-\alpha)}} \quad (4)$$

where r is the activation (or firing rate) after lateral inhibition, y is the firing rate after contrast enhancement, and α and β are the sigmoid threshold and slope respectively. The parameters α and β are constant within each layer, although α is adjusted to control the sparseness of the firing rates. For example, to set the sparseness to, say, 5%, the threshold is set to the value of the 95th percentile point of the activations within the layer. The parameters for the sigmoid activation function are shown in Table 4. We note that a nonlinear sigmoid activation function captures the threshold nonlinearity of real neurons, and the fact that their firing rates saturate at values that are typically in

Table 3. Lateral inhibition parameters

	Layer			
	1	2	3	4
Radius, σ	1.38	2.7	4.0	6.0
Contrast, δ	1.5	1.5	1.6	1.4

TABLE 4. Sigmoid parameters

	Layer			
	1	2	3	4
Percentile	99.2	98	98	91
Slope β	190	40	75	26

the order of 100 spikes/s in the visual system (Rolls & Deco, 2002; Rolls, 2008b).

Trace learning

We summarize next the trace-learning procedure developed and analyzed previously (Földiák, 1991; Rolls, 1992; Wallis & Rolls, 1997; Rolls & Milward, 2000; Rolls & Stringer, 2001). Trace learning utilizes the temporal continuity of objects in the world (over short time periods) to help the learning of invariant representations. The concept here is that on the short time scale, of e.g. a few seconds, the visual input is more likely to be from different transforms of the same object, rather than from a different object. A theory used to account for the development of view-invariant representations in the ventral visual system uses this temporal continuity in a 'trace-learning rule' (Wallis & Rolls, 1997; Rolls & Milward, 2000; Rolls & Stringer, 2001). The trace-learning mechanism relies on associative learning rules, which utilize a temporal trace of activity in the postsynaptic neuron (Földiák, 1991; Rolls, 1992). Trace learning encourages neurons to respond to input patterns which occur close together in time, which are likely to represent different transforms (views) of the same object.

The trace-learning rule (Földiák, 1991; Rolls, 1992; Wallis & Rolls, 1997; Rolls & Milward, 2000) encourages neurons to develop invariant responses to input patterns that tend to occur close together in time, because these are likely to be from the same object. The particular rule used (see Rolls & Milward, 2000) was

$$\delta w_j = \alpha \bar{y}^{\tau-1} x_j^{\tau} \tag{5}$$

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

$$\bar{y}^{\tau} = (1-\eta) y^{\tau} + \eta \bar{y}^{\tau-1} \tag{6}$$

and we have the following definitions: x_j, j^{th} input to the neuron; y, output from the neuron; $\bar{y}^{\tau}$,trace value of the output of the neuron at time step τ; α, learning rate (annealed to zero); w_j, synaptic weight between j^{th} input and the neuron; η, trace value (the optimal value varies with presentation sequence length). The parameter η may be set anywhere in the interval [0, 1], and for the simulations described here η was set to 0.8. A discussion of the good performance of this rule, and its relation to other versions of trace-learning rules, are provided by Rolls & Milward (2000) and Rolls & Stringer (2001).

Simulations: stimuli

The stimuli used to train the networks were computer-generated images of 3-D objects. The objects were created using OpenGL, which gives a maximum of control over all stimulus parameters and positions. OpenGL builds a 3-D representation of the objects and then is able to project different views onto a 2-D image. Lighting was mainly ambient with a diffuse light source added to allow different surfaces to be shown with different intensities as illustrated in Fig. 2, which illustrates the four objects used: (top left) a chair, (top right) a piano, (bottom left) a cabinet and (bottom right) a table.

Simulations: training and test procedure

For the purposes of the simulations described in this paper, layers 1–3 were used to simulate the ventral visual stream, with the layers corresponding to V2, V4 and the inferior temporal visual cortex. The connectivity of the architecture is such that, by layer 3, the connectivity allows information from most of the retina to influence any neuron in layer 3. We used objects placed in one of four quadrants of the input, and trained for invariant representation in layer 3 using the trace rule. For this training, one object was presented on a trial and the different transforms of each object were presented in permuted sequence so that the trace rule could use the temporal continuity to build invariant representations of the object, then another object was selected for training. One epoch was complete when each object had been selected once for training in all of its transforms. At each presentation the activation of individual neurons is calculated, then their firing rates are calculated, and then the synaptic weights are updated. In this manner the network is trained one layer at a time, starting with layer 1 and finishing with layer 3. The numbers of training epochs for layers 1–3 were 50, 100 and 100 respectively (and in each epoch there were four objects each presented in four locations sequentially). After training layers 1–3 in this way, we tested and confirmed that many neurons in layer 3 had translation-invariant representations of the objects, with many neurons responding when one object was presented during testing to one of the objects in all of its transforms.

Layer 4 of the network was treated here as the additional layer added to test processing beyond the inferior temporal visual cortex, and to represent processing in the hippocampus or cortical areas that precede the hippocampus such as the parahippocampal gyrus. It was trained after layers 1–3 had been trained. The training of layer 4 consisted of presenting spatial scenes, and allowing the same self-organizing learning principles to operate as for the earlier layers. Each spatial scene consisted of a particular arrangement of the four objects presented simultaneously. The training stimulus appeared for example as illustrated in Fig. 2. Each of the four spatial scenes was trained for 75 presentations. The training of layer 4 acts as a competitive network (Rolls & Deco, 2002; Rolls, 2008b). Testing was performed by testing whether different layer 4 neurons responded to one of the four different spatial scenes. The design of the scenes allowed for a rigorous test of the hypothesis about the formation of spatial scenes, for each spatial scene contained the identical set of features or objects, but these were arranged spatially differently in the different scenes. In real life, different spatial views might include some of the same features or landmarks, but would probably not overlap totally in the set of features or landmarks that define each spatial scene.

Two information-theoretic measures of performance were used to assess the ability of the layer 3 neurons of the network to respond with view-invariance to individual stimuli or objects (see Rolls & Milward, 2000). A single-cell information measure was applied to individual cells in layer 3 and measures how much information is available from the response of a single cell about which stimulus was shown, independently of view. A multiple-cell information measure, the average amount of information that is obtained about which stimulus was shown from a single presentation of a stimulus from the responses of all the cells, enabled measurement of whether across a population of cells information about every object in the set was provided. Procedures for calculating the multiple-cell information measure are given in Rolls *et al.* (1997b) and Rolls & Milward (2000). In the

FIG. 2. The four objects used in the simulations: (top left) a chair, (top right) a piano, (bottom left) a cabinet and (bottom right) a table. A spatial scene would consist of all objects present simultaneously, as in this Figure. A different spatial scene might consist of the same four objects presented simultaneously, but in a different spatial arrangement.

experiments presented later, the multiple-cell information was calculated from only a small subset of the output cells. There were five cells selected for each stimulus, and these were the five cells which gave the highest single-cell information values for that stimulus.

The maximum single-cell information measure is

$$\text{Maximum single cell information} = \log_2(\text{Number of stimuli}), \quad (7)$$

where in this case the number of objects is four. This gives a maximum single-cell information measure of 2 bits.

Results

Layers 1–3 were trained with single objects using a trace-learning rule to build invariant representations in layer 3 across the four training locations for each object. Figure 3 shows that the representations, when tested with one object present at a time, were invariant. This is shown in Fig. 3a by the fact that a typical layer 3 cell with invariant representations responded to one of the objects (the chair) in all four locations, and to none of the other three objects in any location. Figure 3c shows that many single cells in layer 3 reached

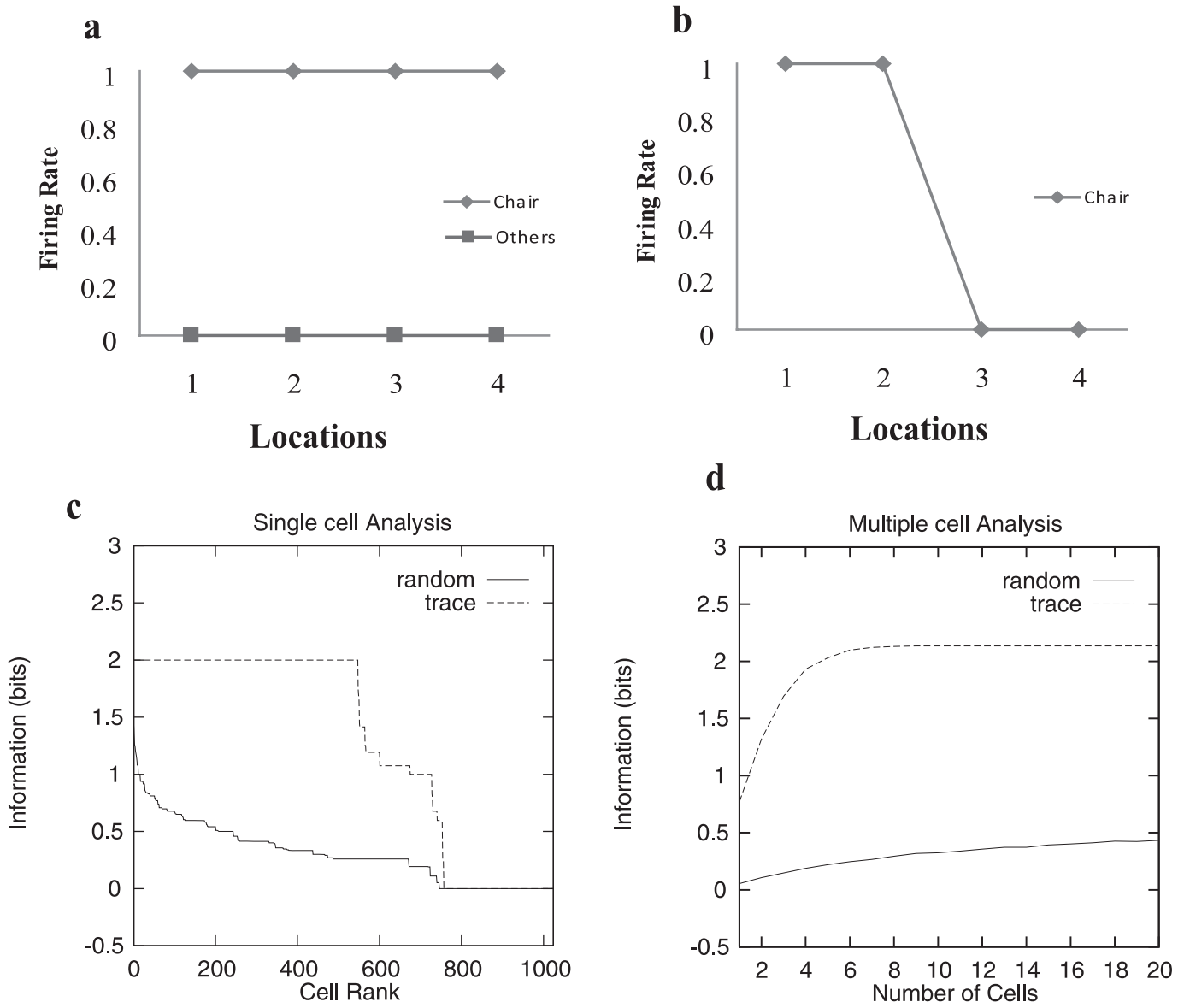

FIG. 3. (a) Firing rate for one layer 3 neuron with position invariant responses across four locations when tested with one object present in the scene at a time. The neuron responded to the chair in all four locations, and to none of the other three objects in any location. (b) Firing rate for the same layer 3 neuron with position invariant responses across two locations when tested with four objects present in the scene. The neuron responded to the chair in two of the four locations. (c) Single-cell information for layer 3 neurons when tested with one object present at a time. (d) Multiple-cell information for layer 3 neurons when tested with one object present at a time.

the maximum level of invariance for a cell trained with four stimuli, namely 2 bits, indicating that these cells responded to only one object but in every location, and to no other objects in any location. Figure 3d shows the multiple-cell information for layer 3 cells and indicates that, for every object, some cells were tuned to have invariant representations of that object. The results in Fig. 3c and d were obtained with one object present during testing. As noted in the 'Hypotheses' section, the above training was performed with successive training of layers 1–3. However, we reran the simulations with simultaneous training of layers 1–3 and obtained results that were very similar and, indeed, indistinguishable from those shown in Fig. 3d, showing that successive vs. simultaneous training of the layers of VisNet involved in invariant object recognition is not an important factor in its success.

After this training of layers 1–3, four spatial-view scenes were presented to the network and layer 4 was trained. Each spatial-view scene consisted of every object that had been trained previously present simultaneously in one spatial arrangement, as illustrated in Fig. 2. After the training, some layer 4 cells were activated by one of the trained spatial-view scenes and much less by the other spatial-view scenes. Figure 4 shows for each spatial scene the responses of the 36 neurons in layer 4 that were most responsive to that scene. For these neurons, the activations produced by the other scenes were 33% of those produced by the scene to which the neuron responded best ($P \ll 0.001$, Mann–Whitney U test). Thus the responses of these neurons were selective for a particular scene. For these neurons, the activations produced by the objects presented individually were 42% of those produced by the scene to which the neuron responded best

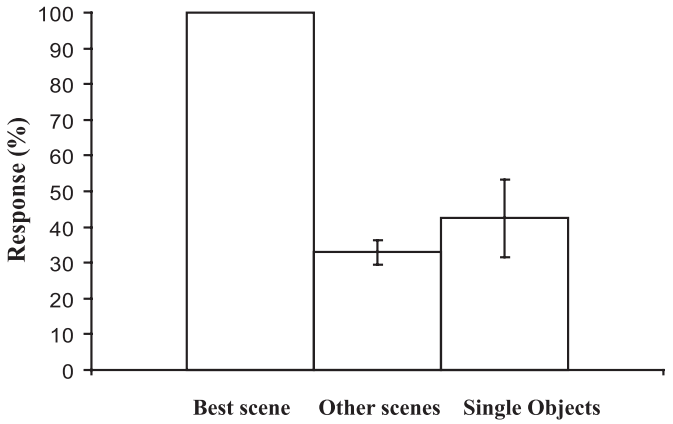

FIG. 4. Scene-specific responses of layer 4 neurons. For the 36 neurons most responsive to each scene, the activations produced in these neurons by the other scenes was 33%, and by single objects was 42%. The values shown are the means and standard errors across the different scene-specific neurons when tested with other scenes, and with the four objects presented one at a time in every position.

($P \ll 0.001$, Mann–Whitney U test). Thus the responses of these neurons were selective for a particular scene and their responses could not be accounted for by responses to individual objects.

The layer 3 to layer 4 connectivity (as the other layers) was a competitive network and, for the results illustrated in Fig. 4, the scene was trained for 75 epochs (where each epoch consists of a single presentation of each of the training stimuli). However, the hypothesis is that this stage of training could be fast, to match with the fact that new spatial representations in the hippocampus, and even more a new episodic memory, can be formed fast (see the section on 'Hypotheses'). Accordingly, we reran the training of the layer 3 to layer 4 connectivity with just four epochs and found that the results were very similar to those already described and illustrated in Fig. 4. In particular, the mean activation for the untrained scenes with a different spatial arrangement of the same objects was 40.7 ± 0.8% of that to the trained scenes, and the mean activation to individual objects was 42.9 ± 0.5% of that to the trained scenes ($P \ll 0.001$ in both cases; Mann–Whitney U test), which can be compared with the data shown in Fig. 4. To compensate for the small number of epochs, the learning rate was increased by a factor of 10. (Four was the minimum number of epochs that could easily be run, and might correspond to one 4-s look at a scene or four 1-s glances at a scene, so this is fast learning.) We note that this training of the hippocampal layer can be fast computationally, because the aim is to build a conjunctive representation based on the neurons that are active in the inferior temporal visual cortex, layer 3 of this simulation, when a single scene is presented. In contrast, the training of invariant representations themselves in layers 1–3 of the network is inherently much slower computationally, because representations have to be extracted from the statistics provided by the exemplars of each object to form a representation of that object that is distinct from the representation formed of other objects from their exemplars by other neurons. However, the important point made in this paper is that, despite these facts, the same type of architecture and learning process can be used to learn the invariant representations of objects up to the inferior temporal visual cortex, and the spatial scene representations in the hippocampal/parahippocampal hippocampal areas, though the synaptic learning rate for the episodic or spatial learning is likely to benefit from being larger than that in the earlier, ventral visual stream layers, to help produce fast 'one-shot' learning vs. the slow incremental learning of invariant representations.

We now analyse how the scene-specific 'spatial-view cells' in layer 4 may be produced, given that there is considerable invariance in layer 3 cells for every object, as illustrated in Fig. 3. The hypothesis we tested was based on the discovery that although inferior temporal cortex neurons typically have large receptive fields, up to 70° in diameter, when tested with a single object presented against a blank background, the receptive fields of the same neurons shrink and become asymmetric when tested with several objects presented simultaneously within ∼10° of the fovea (Aggelopoulos & Rolls, 2005). These neurons may for example respond to the effective object for the neuron when the object is at the fovea or to the upper left or upper right of the fovea, but not when it is to the lower left or lower right of the fovea. The implication of this neurophysiology is that the lateral inhibition is stronger when several objects are presented than when one object is being presented, and that this reduces the receptive field size and reveals underlying asymmetries in the probabilistic forward connections received by each neuron, as illustrated in Fig. 1, and also in the probabilistic connectivity implemented through the inhibitory interneurons.

We tested this hypothesis in these simulations with VisNet by comparing the number of neurons that had complete invariance in that they responded to their effective object in all four locations, with the numbers that responded to fewer locations, when one object was present during testing compared with when all four objects were present during testing. Figure 3b shows an example of a layer 3 cell that responded to its effective stimulus, a chair, in two of four locations when four objects were present simultaneously, though it responded to its effective object in all four locations when only one object was present, as shown in Fig. 3a. Figure 5 shows that, when tested with one object present at a time, all the layer 3 neurons with best responses to a particular object responded to all four possible positions of the object. There were for each object, on average, 169 completely position-invariant neurons, as shown in Fig. 5. Figure 5 shows that, when tested with four objects present simultaneously, the same neurons with best responses to a particular object responded overall to fewer locations of the effective stimulus for the neuron. Reasonable numbers of neurons responded to their effective stimulus in two or three locations (on average 60 of the 169 neurons as shown in Fig. 5), but only one neuron on average responded to the effective stimulus in all four locations. This indicates that the receptive fields of VisNet neurons do shrink, and show somewhat less translation invariance, when several objects are presented simultaneously in a whole scene. It is this reduced translation invariance that will facilitate the ability of the layer 4 neurons to be able to learn to respond to one arrangement of the simultaneously presented objects but to have smaller responses to another scene with a different arrangement of the same four objects.

Figure 5 shows that, even when tested on whole scenes, for which scene-specific spatial-view neurons can be formed in layer 4 (corresponding to hippocampal areas), the layer 3 neurons (corresponding in these simulations to the inferior temporal visual cortex) show some invariant responsiveness. The degree of invariance shown may be useful for invariant object recognition, with neurons often responding to an object in more than one location. On the other hand, there is sufficient spatial asymmetry that the layer 3 neurons can provide sufficient spatial information when several objects are simultaneously present for layer 4 of the network to learn scene representations that are specific to the spatial arrangement of the objects in the scene.

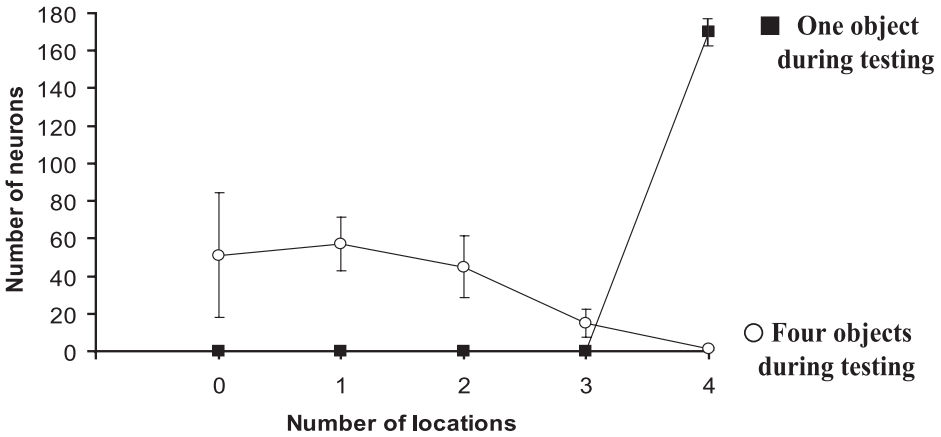

FIG. 5. Smaller receptive fields of layer 3 neurons when tested with all four objects presented simultaneously during testing compared to when one object is present during testing. When tested with one object present at a time, all the neurons with best responses to a particular object responded to all four possible positions of the object. There were for each object, on average, 169 completely position-invariant neurons, as shown in the plot labelled 'one object during testing'. The neurons did not respond to other objects. When tested with four objects present simultaneously, the same neurons with best responses to a particular object responded to overall fewer locations of the effective stimulus for the neuron.

Discussion

The results described here show that when the same self-organizing principles that can account for the formation of neurons with invariant responses in the inferior temporal visual cortex are applied in a further layer of computation trained by the same principles, but now including spatial scenes in the training data, then neurons with spatial-view-specific responses develop in the new added layer. This is thus a unifying computational hypothesis, for it shows that not only invariant representations can be developed with the functional architecture, but that spatial-view neurons can be formed by the same computational principles operating in a further layer or layers. Part of the attractiveness of this unifying computational hypothesis is that it makes the design of the brain by evolutionary processes relatively tractable in that adding another layer, rather than inventing new computational principles, is what is required.

The last layer in the conceptual design (layer 5 in the network illustrated in Fig. 1, and layer 4 in the network as simulated) is taken to correspond to systems such as the hippocampus and parahippocampal gyrus in which spatial-view cells are found. In fact, spatial-view cells are found in the primate (macaque) parahippocampal gyrus as well as the hippocampus, fields CA3 and CA1 (Rolls *et al.*, 1997a, 1998, 2005; Robertson *et al.*, 1998; Georges-François *et al.*, 1999; Rolls & Kesner, 2006; Rolls & Xiang, 2006). Given that there are connections from the temporal cortical visual areas to the cortical areas that overlie the hippocampus and in turn send projections to the hippocampus via the entorhinal cortex (Van Hoesen, 1982; Suzuki & Amaral, 1994; Lavenex & Amaral, 2000; Witter *et al.*, 2000; Lavenex *et al.*, 2004), it is possible that spatial-view cells are formed in these cortical areas that overlie the hippocampus. These representations may then be passed on to the entorhinal cortex, and thus into the hippocampus via the dentate granule cells. The dentate granule cells may by competitive learning help to make the representation more sparse (Rolls & Kesner, 2006; Rolls *et al.*, 2006; Rolls, 2008b). As the animal navigates through the environment and looks at different spatial views, different spatial-view cells would be formed.

Because of the overlapping fields of adjacent spatial-view neurons, and hence their coactivity as the animal navigates, recurrent collateral associative connections at the next stage of the system, CA3, could form a continuous attractor representation of the environment. Part of the utility of forming a continuous spatial attractor representation in CA3 is that if it operates as a single network due to its widespread recurrent collateral connections (Rolls, 1996, 2008b; Rolls & Treves, 1998; Rolls & Kesner, 2006) then any pair of different landmarks that happened to be close in a spatial environment could be associated together by coactivity in the recurrent collateral connections, even if quite different neurons in the network represented the different landmarks. This type of associatively might be harder to build in the neocortex where the recurrent collateral connections are short-range. We thus have a hypothesis for how the spatial representations are formed as a natural extension of the hierarchically organised competitive networks in the ventral visual system. The expression of such spatial representations in CA3 may be particularly useful for associating those spatial representations with other inputs, such as objects or rewards, and thus in episodic memory (Rolls & Xiang, 2005, 2006; Rolls *et al.*, 2005; Rolls & Kesner, 2006; Rolls, 2008b).

Part of the mechanism by which layer 3 neurons can support the formation of scene-specific neurons in layer 4 is that the receptive fields of the layer 3 neurons, though in many cases large and fully position invariant when tested with one object, become smaller, responding to fewer locations of the effective object, when tested with simultaneously presented objects, as in natural scenes (see Fig. 5). At the same time, Fig. 5 shows that, even when tested on whole scenes with four objects present simultaneously, the layer 3 neurons show some invariant responsiveness, which may be useful for invariant object recognition, yet sufficient spatial asymmetry that they can provide sufficient spatial information when several objects are simultaneously present for layer 4 of the network to learn scene representations that are specific to the spatial arrangement of the objects or landmarks in the scene. Part of the concept described in this paper (which is consistent with the neurophysiological evidence of Aggelopoulos & Rolls, 2005) is that the last layer of the unimodal ventral visual system corresponding with the inferior temporal visual cortex has receptive fields that, due to competition and the sparseness of the representation, decrease in size and become asymmetric with respect to the fovea, in a complex spatial scene when more than one object is present simultaneously. For this reason, the inferior temporal cortex cannot support spatial scene representations in which several objects in the correct spatial position must be represented. It is partly for this reason that an additional layer to the hierarchy, identified with

the parahippocampal gyrus–hippocampus, with additional convergence and an appropriate level of competition, is needed to provide a conjunctive, scene, representation of what is represented in the inferior temporal visual cortex.

The simulations described here provide a computational account of the mechanism by which the receptive fields of inferior temporal cortex neurons become smaller in crowded scenes. The simulations show that this is related to the diluted, probabilistic, feedforward connectivity which has an approximately Gaussian spatial distribution from neurons in the preceding layer (see Materials and methods and Fig. 1). If the inhibition in such a layer of neurons is increased because other neurons become active due to different objects being presented simultaneously then the Gaussian spatial profile will tend to make the receptive field shrink, and the fact that the connectivity is incomplete will mean that there may be more connections present from neurons in the preceding layer in one direction with respect to another, with this producing asymmetry in the receptive field shape that becomes particularly evident when the competitive inhibition is strong. With less competitive inhibition, when one object is present at a time, the smaller number of connections in a given direction are nevertheless sufficient to produce receptive fields that respond to the effective object for a neuron in all positions of the object, as illustrated in Fig. 5. In this context, recent neurophysiology shows how there is some spatial information in object-tuned neurons in the ventral visual system in scenes with multiple objects present simultaneously (Aggelopoulos & Rolls, 2005), and the results described here indicate how this could arise computationally.

The findings described here complement and are supported by results obtained with a more artificial network trained by gradient ascent with a goal function that included forming relatively time-invariant representations and decorrelating the responses of neurons in the multilayer hierarchical network (Wyss *et al.*, 2006; Franzius *et al.*, 2007). In their simulations, place-like cells were formed at the end of the network when the system was trained with a real or simulated robot moving through spatial environments (Wyss *et al.*, 2006), and view cells were formed when training on video sequences from a virtual-reality environment (Franzius *et al.*, 2007). The functional architecture of these studies was more abstract, in that the goal function was prescribed, and gradient ascent to optimize the goal function was used. The approach taken in this paper is different in that it starts with the details of the architecture, including convergent probabilistic feedforward connectivity as illustrated in Fig. 1 and competitive learning implemented by a local associative (Hebbian) synaptic modification rule with a short-term memory trace component and mutual inhibition between neurons, and shows that by adding an additional (fifth) layer with the same architecture, and then training it on whole scenes, scene-specific spatial-view cells can be produced. The test for spatial-view neurons is particularly rigorous in the present work because the same objects, features and/or landmarks are present in the different scenes, and each scene is formed by a particular arrangement of the relevant objects, features and/or landmarks. In the work with more abstract training (Wyss *et al.*, 2006; Franzius *et al.*, 2007), the visual inputs were derived from a robot moving in the world or from images of the world, and so different spatial scenes might be composed of different objects or landmarks, which is an easier problem more like that of object recognition in which the objects differ from each other by having at least largely nonoverlapping sets of features. The present approach also shows how invariant representations in the inferior temporal visual cortex are related to the need for some spatial information to be represented if they are to provide the basis for spatial scene learning, for the present results show that the spatial selectivity of inferior temporal cortex neurons becomes more evident in complex natural scenes with several objects or landmarks present simultaneously.

The present results are also relevant to understanding the finding that areas beyond the inferior temporal cortex may be especially important when the different objects to be discriminated have many overlapping features (Bussey *et al.*, 2002, 2003, 2005; Bussey & Saksida, 2005, 2007; Buckley & Gaffan, 2006). Objects as represented in the inferior temporal visual cortex are everyday objects (such as those illustrated in Fig. 2 and in Rolls, 2008b, fig. 2.11) which typically overlap in relatively few features (Rolls, 2008b). It is proposed that part of the solution to learning to discriminate difficult objects, in which there are many features in the different objects that are in common, is that by adding an additional layer of processing to the inferior temporal visual cortex, such as perirhinal and/or parahippocampal areas, combinations of the features with their relative spatial position encoded by the inferior temporal visual cortex neurons (because of the asymmetry of their receptive fields described here and by Aggelopoulos & Rolls, 2005) can allow neurons to be formed in the added layer that depend on new combinations of features in which the relative spatial position is part of the new representation formed (Rolls, 2008b). Thus the present hypothesis provides a computational account for how a cortical area beyond the inferior temporal visual cortex can be important for perceptual discrimination when the objects consist of overlapping sets of features, and also how the spatial arrangement of the features can be incorporated into the representation that may be required for the discrimination, which has not been accounted for by previous models.

We note that in a review paper by Bussey & Saksida (2007) they refer to the concept that the hippocampus is an extension of the ventral visual system (Squire, 1992; Mishkin *et al.*, 1997) useful for solving 'object ambiguity' when objects are repeated in for example a repeating-items Delayed Match-to-sample task, and that they summarize their view as follows: "Thus perirhinal cortex, which contains complex conjunctive representations specifying unique objects, protects…from interfering feature ambiguity. In the case of repeating items, however, not only are the features repeatedly presented, the objects are repeatedly presented. As a result, the representations in perirhinal cortex are not enough to protect…from interference. One might say that an additional level of ambiguity, 'object ambiguity', has been created. Now, the resolution of ambiguity at this level would require an additional, more rostral layer, containing conjunctive representations of an even higher degree of complexity than those found in perirhinal cortex." They suggest that this functionality, the resolution of object ambiguity when objects are repeated, may be provided by the hippocampus. However, they do not address the central theme of the present paper, that the parahippocampal gyrus–hippocampus may actually form spatial-view representations, by combining representations of objects that include some information about the positions of the objects due to the spatially asymmetric receptive fields of inferior temporal cortex neurons in crowded scenes. Moreover, part of the central theme of the present paper is that the computations that allow these conjunctive representations to be formed require no more than adding a further layer of feed-forward competitive learning to the existing hierarchy, and this added layer we identify generically with the parahippocampal gyrus–hippocampus. Although we have modelled this further conjunctive learning provided by the parahippocampal cortex–hippocampus in this paper by competitive learning, which is a type of computation that can be implemented by cortical areas and by the dentate gyrus (Rolls & Treves, 1998; Rolls & Deco, 2002; Rolls *et al.*, 2006; Rolls, 2008b), we note that the CA3 recurrent collateral system of the hippocampus could also contribute to the same functionality, as it includes

competition and associative learning, as well as to associations between objects and places (Rolls & Treves, 1998; Rolls & Kesner, 2006; Rolls & Xiang, 2006; Rolls, 2008b).

The fact that the receptive fields of inferior temporal cortex neurons become smaller and asymmetric with respect to the fovea in scenes with several objects situated close to the fovea (Aggelopoulos & Rolls, 2005) provides a solution to the representation of multiple objects in a scene, which is an important issue in hierarchically convergent object recognition systems with distributed representations (Mozer, 1991). By having object-selective neurons with different spatial asymmetries, the representation provided by a population of neurons provides information not just about what objects are present in a scene but also about their relative spatial position with respect to the fovea (Aggelopoulos & Rolls, 2005). Indeed, this type of encoding could account not only for how we can see several objects in a scene in their correct spatial position with respect to the fovea, but also for how we are able to see two versions of the same object at different positions in a scene (Aggelopoulos & Rolls, 2005). The contribution of the present paper to this issue is that it shows that these asymmetries can arise in a hierarchical feedforward network with probabilistic forward connectivity, which provides the basis for asymmetries in the receptive fields to emerge when several objects are sufficiently close so that inhibition through inhibitory neurons (themselves with probabilistic connectivity) can reveal the underlying asymmetry produced by the probabilistic connectivity.

It is an interesting part of the hypothesis described that, because spatial views and places are defined by the relative spatial positions of fixed landmarks (such a buildings), slow learning of such representations over a number of trials might be useful, so that the neurons come to represent spatial views or places and do not learn to represent a random collection of moveable objects seen once in conjunction. This enables us to make a clear distinction between representations of objects and representations of scenes. Objects are typically moveable, and can appear in different places in a spatial environment. The features within an object are associated with each other to form a representation of the object, but strong associations are not made between those features and scenes (or other objects) because statistically the associations are stronger between features within an object than between the features in a spatial scene or in another object (Stringer *et al.*, 2007; Stringer & Rolls, 2008). In contrast, the landmarks, features or objects that are part of a spatial scene are seen in the same overall spatial relationship to each other because of the fixed nature of a scene. An example is that when looking at a scene in a room the floor, walls and ceiling are always in a fixed relationship to each other, and the walls are not sometimes seen below the floor. It is this type of scene-learning that we have shown could be implemented by adding an additional layer to the ventral visual stream architecture.

In conclusion, we believe that it is an interesting and unifying hypothesis that an effect of adding an additional layer to VisNet-like ventral stream visual cortical processing might with training in a natural environment lead to the self-organization, using the same principles as in the ventral visual stream, of spatial-view representations in parahippocampal or hippocampal areas. This hypothesis helps to bring together in a unifying framework (Rolls, 2008b) neurophysiological studies on invariant object and face representations at the end of the unimodal ventral visual stream in the inferior temporal visual cortex (Rolls, 1984, 2000, 2007; Baylis *et al.*, 1985, 1987; Rolls & Baylis, 1986; Hasselmo *et al.*, 1989; Tovee *et al.*, 1994, 1996; Rolls *et al.*, 1997b,c, 2003; Booth & Rolls, 1998; Hölscher *et al.*, 2003; Aggelopoulos & Rolls, 2005; Aggelopoulos *et al.*, 2005; Franco *et al.*, 2007) and how they may be formed computationally (Rolls, 1992; Wallis & Rolls, 1997; Rolls & Milward, 2000; Stringer & Rolls, 2000; Rolls & Stringer, 2001; Elliffe *et al.*, 2002; Rolls & Deco, 2002; Stringer & Rolls, 2002; Trappenberg *et al.*, 2002; Deco & Rolls, 2004; Perry *et al.*, 2006; Rolls & Deco, 2006; Rolls and Stringer, 2006a,b; Stringer *et al.*, 2006, 2007), with neurophysiological studies on the representation of space in the primate hippocampus (Rolls & O'Mara, 1995; Rolls *et al.*, 1997a, 1998, 2005; Robertson *et al.*, 1998; Georges-François *et al.*, 1999; Rolls, 1999; Rolls & Xiang, 2005, 2006) and how this may be formed computationally (Rolls, 1996; Rolls & Treves, 1998; de Araujo *et al.*, 2001; Rolls *et al.*, 2002, 2006; Stringer *et al.*, 2004, 2005; Rolls & Stringer, 2005; Rolls & Kesner, 2006).

Acknowledgement

This research was supported by the Wellcome Trust.

References

Abeles, M. (1991) *Corticonics - Neural Circuits of the Cerebral Cortex*. Cambridge University Press, New York.

Aggelopoulos, N.C. & Rolls, E.T. (2005) Natural scene perception: inferior temporal cortex neurons encode the positions of different objects in the scene. *Eur. J. Neurosci.*, **22**, 2903–2916.

Aggelopoulos, N.C., Franco, L. & Rolls, E.T. (2005) Object perception in natural scenes: encoding by inferior temporal cortex simultaneously recorded neurons. *J. Neurophysiol.*, **93**, 1342–1357.

de Araujo, I.E.T., Rolls, E.T. & Stringer, S.M. (2001) A view model which accounts for the spatial fields of hippocampal primate spatial view cells and rat place cells. *Hippocampus*, **11**, 699–706.

Baylis, G.C., Rolls, E.T. & Leonard, C.M. (1985) Selectivity between faces in the responses of a population of neurons in the cortex in the superior temporal sulcus of the monkey. *Brain Res.*, **342**, 91–102.

Baylis, G.C., Rolls, E.T. & Leonard, C.M. (1987) Functional subdivisions of the temporal lobe neocortex. *J. Neurosci.*, **7**, 330–342.

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

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

Braitenberg, V. & Schütz, A. (1991) *Anatomy of the Cortex*. Springer-Verlag, Berlin.

Buckley, M.J. & Gaffan, D. (2006) Perirhinal cortical contributions to object perception. *Trends Cogn. Sci.*, **10**, 100–107.

Bussey, T.J. & Saksida, L.M. (2005) Object memory and perception in the medial temporal lobe: an alternative approach. *Curr. Opin. Neurobiol.*, **15**, 730–737.

Bussey, T.J. & Saksida, L.M. (2007) Memory, perception, and the ventral visual-perirhinal-hippocampal stream: thinking outside of the boxes. *Hippocampus*, **17**, 898–908.

Bussey, T.J., Saksida, L.M. & Murray, E.A. (2002) Perirhinal cortex resolves feature ambiguity in complex visual discriminations. *Eur. J Neurosci.*, **15**, 365–374.

Bussey, T.J., Saksida, L.M. & Murray, E.A. (2003) Impairments in visual discrimination after perirhinal cortex lesions: testing 'declarative' vs. 'perceptual-mnemonic' views of perirhinal cortex function. *Eur. J. Neurosci.*, **17**, 649–660.

Bussey, T.J., Saksida, L.M. & Murray, E.A. (2005) The perceptual-mnemonic/feature conjunction model of perirhinal cortex function. *Q. J. Exp. Psychol. B*, **58**, 269–282.

Cowell, R.A., Bussey, T.J. & Saksida, L.M. (2006) Why does brain damage impair memory? A connectionist model of object recognition memory in perirhinal cortex. *J. Neurosci.*, **26**, 12186–12197.

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

Dere, E., Easton, A., Nadel, L. & Huston, J.P. (2009) *Handbook of Episodic Memory*. Elsevier, Amsterdam.

Desimone, R. (1991) Face-selective cells in the temporal cortex of monkeys. *J. Cogn. Neurosci.*, **3**, 1–8.

Dolan, R.J., Fink, G.R., Rolls, E.T., Booth, M., Holmes, A., Frackowiak, R.S.J. & Friston, K.J. (1997) How the brain learns to see objects and faces in an impoverished context. *Nature*, **389**, 596–599.

Elliffe, M.C.M., Rolls, E.T. & Stringer, S.M. (2002) Invariant recognition of feature combinations in the visual system. *Biol. Cybern.*, **86**, 59–71.
Földiák, P. (1991) Learning invariance from transformation sequences. *Neural Comput.*, **3**, 194–200.
Franco, L., Rolls, E.T., Aggelopoulos, N.C. & Jerez, J.M. (2007) Neuronal selectivity, population sparseness, and ergodicity in the inferior temporal visual cortex. *Biol. Cybern.*, **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.
Georges-François, P., Rolls, E.T. & Robertson, R.G. (1999) Spatial view cells in the primate hippocampus: allocentric view not head direction or eye position or place. *Cereb. Cortex*, **9**, 197–212.
Hasselmo, M.E., Rolls, E.T., Baylis, G.C. & 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.
Hölscher, C., Rolls, E.T. & Xiang, J.-Z. (2003) Perirhinal cortex neuronal activity related to long-term familiarity memory in the macaque. *Eur. J. Neurosci.*, **18**, 2037–2046.
Hooks, B.M. & Chen, C. (2007) Critical periods in the visual system: changing views for a model of experience-dependent plasticity. *Neuron*, **56**, 312–326.
Ito, M., Tamura, H., Fujita, I. & Tanaka, K. (1995) Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J. Neurophysiol.*, **73**, 218–226.
Kobatake, E. & Tanaka, K. (1994) Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophysiol.*, **71**, 856–867.
Lavenex, P. & Amaral, D.G. (2000) Hippocampal-neocortical interaction: a hierarchy of associativity. *Hippocampus*, **10**, 420–430.
Lavenex, P., Suzuki, W.A. & Amaral, D.G. (2004) Perirhinal and parahippocampal cortices of the macaque monkey: intrinsic projections and interconnections. *J. Comp. Neurol.*, **472**, 371–394.
Mishkin, M., Suzuki, W.A., Gadian, D.G. & Vargha-Khadem, F. (1997) Hierarchical organization of cognitive memory. *Philos. Trans R. Soc. Lond.*, **352**, 1461–1467.
Mozer, M. (1991) *The Perception of Multiple Objects: A Connectionist Approach*. MIT Press, Cambridge, MA.
Op De Beeck, H. & Vogels, R. (2000) Spatial sensitivity of macaque inferior temporal neurons. *J. Comp. Neurol.*, **426**, 505–518.
Perry, G., Rolls, E.T. & Stringer, S.M. (2006) Spatial vs temporal continuity in view invariant visual object recognition learning. *Vision Res.*, **46**, 3994–4006.
Riesenhuber, M. & Poggio, T. (1999) Hierarchical models of object recognition in cortex. *Nat. Neurosci.*, **2**, 1019–1025.
Robertson, R.G., Rolls, E.T. & Georges-François, P. (1998) Spatial view cells in the primate hippocampus: effects of removal of view details. *J. Neurophysiol.*, **79**, 1145–1156.
Rolls, E.T. (1984) Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces. *Hum. Neurobiol.*, **3**, 209–222.
Rolls, E.T. (1992) Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philos. Trans R. Soc. Lond. B*, **335**, 11–21.
Rolls, E.T. (1996) A theory of hippocampal function in memory. *Hippocampus*, **6**, 601–620.
Rolls, E.T. (1999) Spatial view cells and the representation of place in the primate hippocampus. *Hippocampus*, **9**, 467–480.
Rolls, E.T. (2000) Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron*, **27**, 205–218.
Rolls, E.T. (2005) *Emotion Explained*. Oxford University Press, Oxford.
Rolls, E.T. (2007) The representation of information about faces in the temporal and frontal lobes. *Neuropsychologia*, **45**, 125–143.
Rolls, E.T. (2008a) Face processing in different brain areas, and critical band masking. *J. Neuropsychol.*, **2**, 325–360.
Rolls, E.T. (2008b) *Memory, Attention, and Decision-Making: A Unifying Computational Neuroscience Approach*. Oxford University Press, Oxford.
Rolls, E.T. (2009) The primate hippocampus and episodic memory. In Dere, E., Easton, A., Nadel, L. & Huston, J.P. (Eds), *Handbook of Episodic Memory*. Elsevier, Amsterdam, pp. 415–436.
Rolls, E.T. & Baylis, G.C. (1986) Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. *Exp. Brain Res.*, **65**, 38–48.
Rolls, E.T. & Deco, G. (2002) *Computational Neuroscience of Vision*. Oxford University Press, Oxford.
Rolls, E.T. & Deco, G. (2006) Attention in natural scenes: neurophysiological and computational bases. *Neural Netw.*, **19**, 1383–1394.
Rolls, E.T. & Kesner, R.P. (2006) A computational theory of hippocampal function, and empirical tests of the theory. *Prog. Neurobiol.*, **79**, 1–48.
Rolls, E.T. & Milward, T. (2000) A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Comput.*, **12**, 2547–2572.
Rolls, E.T. & O'Mara, S.M. (1995) View-responsive neurons in the primate hippocampal complex. *Hippocampus*, **5**, 409–424.
Rolls, E.T. & Stringer, S.M. (2001) Invariant object recognition in the visual system with error correction and temporal difference learning. *Netw. Comput. Neural Syst.*, **12**, 111–129.
Rolls, E.T. & Stringer, S.M. (2005) Spatial view cells in the hippocampus, and their idiothetic update based on place and head direction. *Neural Netw.*, **18**, 1229–1241.
Rolls, E.T. & Stringer, S.M. (2006a) Invariant global motion recognition in the dorsal visual system: a unifying theory. *Neural Comput.*, **19**, 139–169.
Rolls, E.T. & Stringer, S.M. (2006b) Invariant visual object recognition: a model, with lighting invariance. *J. Physiol. - Paris*, **100**, 43–62.
Rolls, E.T. & Treves, A. (1998) *Neural Networks and Brain Function*. Oxford University Press, Oxford.
Rolls, E.T. & Xiang, J.-Z. (2005) Reward-spatial view representations and learning in the hippocampus. *J. Neurosci.*, **25**, 6167–6174.
Rolls, E.T. & Xiang, J.-Z. (2006) Spatial view cells in the primate hippocampus, and memory recall. *Rev. Neurosci.*, **17**, 175–200.
Rolls, E.T., Baylis, G.C., Hasselmo, M.E. & Nalwa, V. (1989) The effect of learning on the face-selective responses of neurons in the cortex in the superior temporal sulcus of the monkey. *Exp. Brain Res.*, **76**, 153–164.
Rolls, E.T., Robertson, R.G. & Georges-François, P. (1997a) Spatial view cells in the primate hippocampus. *Eur. J. Neurosci.*, **9**, 1789–1794.
Rolls, E.T., Treves, A. & Tovee, M.J. (1997b) 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, E.T., Treves, A., Tovee, M.J. & Panzeri, S. (1997c) Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. *J. Comput. Neurosci.*, **4**, 309–333.
Rolls, E.T., Treves, A., Robertson, R.G., Georges-François, P. & Panzeri, S. (1998) Information about spatial view in an ensemble of primate hippocampal cells. *J. Neurophysiol.*, **79**, 1797–1813.
Rolls, E.T., Stringer, S.M. & Trappenberg, T.P. (2002) A unified model of spatial and episodic memory. *Proc. R. Soc. Lond. B*, **269**, 1087–1093.
Rolls, E.T., Aggelopoulos, N.C. & Zheng, F. (2003) The receptive fields of inferior temporal cortex neurons in natural scenes. *J. Neurosci.*, **23**, 339–348.
Rolls, E.T., Xiang, J.-Z. & Franco, L. (2005) Object, space and object-space representations in the primate hippocampus. *J. Neurophysiol.*, **94**, 833–844.
Rolls, E.T., Stringer, S.M. & Elliot, T. (2006) Entorhinal cortex grid cells can map to hippocampal place cells by competitive learning. *Netw. Comput. Neural Syst.*, **17**, 447–465.
Squire, L.R. (1992) Memory and the hippocampus: a synthesis from findings with rats, monkeys and humans. *Psychol. Rev.*, **99**, 195–231.
Stringer, S.M. & Rolls, E.T. (2000) Position invariant recognition in the visual system with cluttered environments. *Neural Networks*, **13**, 305–315.
Stringer, S.M. & Rolls, E.T. (2002) Invariant object recognition in the visual system with novel views of 3D objects. *Neural Comput.*, **14**, 2585–2596.
Stringer, S.M. & Rolls, E.T. (2008) Learning transform invariant object recognition in the visual system with multiple stimuli present during training. *Neural Networks*, **21**, 888–903.
Stringer, S.M., Rolls, E.T. & Trappenberg, T.P. (2004) Self-organising continuous attractor networks with multiple activity packets, and the representation of space. *Neural Networks*, **17**, 5–27.
Stringer, S.M., Rolls, E.T. & Trappenberg, T.P. (2005) Self-organizing continuous attractor network models of hippocampal spatial view cells. *Neurobiol. Learn. Mem.*, **83**, 79–92.
Stringer, S.M., Perry, G., Rolls, E.T. & Proske, J.H. (2006) Learning invariant object recognition in the visual system with continuous transformations. *Biol. Cybern.*, **94**, 128–142.
Stringer, S.M., Rolls, E.T. & Tromans, J. (2007) Invariant object recognition with trace learning and multiple stimuli present during training. *Netw. Comput. Neural Syst.*, **18**, 161–187.
Suzuki, W.A. & Amaral, D.G. (1994) Perirhinal and parahippocampal cortices of the macaque monkey - cortical afferents. *J. Comp. Neurol.*, **350**, 497–533.
Tanaka, K., Saito, H., Fukada, Y. & Moriya, M. (1991) Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *J. Neurophysiol.*, **66**, 170–189.
Tovee, M.J., Rolls, E.T. & Azzopardi, P. (1994) Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert macaque. *J. Neurophysiol.*, **72**, 1049–1060.

Tovee, M.J., Rolls, E.T. & Ramachandran, V.S. (1996) Rapid visual learning in neurones of the primate temporal visual cortex. *Neuroreport*, **7**, 2757–2760.
Trappenberg, T.P., Rolls, E.T. & Stringer, S.M. (2002) Effective size of receptive fields of inferior temporal cortex neurons in natural scenes. In Dietterich, T.G., Becker, S. & Ghahramani, Z. (Eds), *Advances in Neural Information Processing Systems 14*. MIT Press, Cambridge, MA, pp. 293–300.
Ullman, S. (1996) *High-Level Vision: Object Recognition and Visual Cognition*. Bradford/MIT press, Cambridge, MA.
Van Hoesen, G.W. (1982) The parahippocampal gyrus. New observations regarding its cortical connections in the monkey. *Trends Neurosci.*, **5**, 345–350.
Wallis, G. & Rolls, E.T. (1997) Invariant face and object recognition in the visual system. *Prog. Neurobiol.*, **51**, 167–194.
Wiskott, L. & Sejnowski, T.J. (2002) Slow feature analysis: unsupervised learning of invariances. *Neural Comput.*, **14**, 715–770.
Witter, M.P., Wouterlood, F.G., Naber, P.A. & Van Haeften, T. (2000) Anatomical organization of the parahippocampal-hippocampal network. *Ann. N.Y. Acad. Sci.*, **911**, 1–24.
Wyss, R., Konig, P. & Verschure, P.F. (2006) A model of the ventral visual system based on temporal stability and local memory. *PLoS Biol.*, **4**, e120.