Network: Computation in Neural Systems
March–June 2012; 23(1–2): 1–23

Learning separate visual representations of independently rotating objects

JAMES MATTHEW TROMANS, HECTOR J.I. PAGE, & SIMON M. STRINGER

Department of Experimental Psychology, University of Oxford, Experimental Psychology, South Parks Road, Oxford, OX1 3UD, UK

(Received 6 November 2011; accepted 25 November 2011)

Abstract
Individual cells that respond preferentially to particular objects have been found in the ventral visual pathway. How the brain is able to develop neurons that exhibit these object selective responses poses a significant challenge for computational models of object recognition. Typically, many objects make up a complex natural scene and are never presented in isolation. Nonetheless, the visual system is able to build invariant object selective responses. In this paper, we present a model of the ventral visual stream, VisNet, which can solve the problem of learning object selective representations even when multiple objects are always present during training. Past research with the VisNet model has shown that the network can operate successfully in a similar training paradigm, but only when training comprises many different object pairs. Numerous pairings are required for statistical decoupling between objects. In this research, we show for the first time that VisNet is capable of utilizing the statistics inherent in independent rotation to form object selective representations when training with just two objects, always presented together. Crucially, our results show that in a dependent rotation paradigm, the model fails to build object selective representations and responds as if the two objects are in fact one. If the objects begin to rotate independently, the network forms representations for each object separately.

Keywords: *Object recognition, rotating objects, inferior temporal cortex*

Introduction

The ventral visual system plays a significant role in the processing of visual form and ultimately object identification (Goodale and Milner 1992). Neurophysiological studies suggest that these regions are organised in a hierarchical fashion, where

Correspondence: James Matthew Tromans, BA, MSc, Oxford Center for Theoretical Neuroscience and Artificial Intelligence, Department of Experimental Psychology, University of Oxford, Experimental Psychology, South Parks Road, Oxford, OX1 3UD, UK. E-mail: james.tromans@psy.ox.ac.uk

ISSN 0954-898X print/ISSN 1361-6536 online/02/01-020001–23 © 2012 Informa Healthcare Ltd.
DOI: 10.3109/0954898X.2011.651520

neuronal responses become more complex and receptive field sizes increase towards the latter stages of the ventral stream (Gattass et al. 1988; Levitt et al. 1994; Gegenfurtner et al. 1997). Over successive stages, the visual system develops neurons that respond with view, size and position (translation) invariance to objects or faces (Hasselmo et al. 1989; Desimone 1991; Rolls et al. 1992; Perrett and Oram 1993; Kobatake and Tanaka 1994; Tovee et al. 1994; Ito et al. 1995; Booth and Rolls 1998; Op De Beeck and Vogels 2000).

Given these findings, a major computational challenge is to understand the fundamental mechanisms of how the ventral stream may actually form separate neuronal representations for individual objects, despite the fact that objects are always presented in complex natural scenes. This research investigates how the primate visual system may build invariant representations of individual objects when multiple objects are always present in a scene. How the visual system learns representations of individual objects instead of just the combined scene is an important question for natural vision. Recent research has successfully shown that it is possible for the VisNet model of the ventral visual stream to utilise the statistics in the training input to build separate invariant representations of objects used during training, despite the fact that pairs of objects were always presented (Stringer et al. 2007; Stringer and Rolls 2008). Many different object pairings were required to produce a training set where each object was combined with the other objects to create a selection of object pairs. Crucially, during presentation the features that make up an object occur together more frequently compared to the features that make up different objects. The frequency with which the objects are presented together during training denotes the level of correlation between features from different objects. However, the features that comprise individual objects are always seen together. It has been shown that a competitive network will operate usefully in this situation forming invariant representations of individual objects, rather than the combinations of objects seen during training, by a mechanism such as Continuous Transformation learning (Stringer and Rolls 2008).

The number of different object pairs presented to the network during training limits this training method. Although a training paradigm comprising many objects is ecologically valid, there may be other underlying mechanisms available to assist this type of object selective learning. In this paper we investigate, for the first time, whether a multi-layer hierarchical model of the ventral visual pathway is able to use independent rotation to form separate object selective representations when trained with just two realistic 3D rotating objects always co-present on the retina.

We investigate the effects of a competitive network and also a self-organising map (SOM) by introducing short-range excitatory connections and long-range inhibitory connections (von der Malsburg 1973; Kohonen 1982). This lateral connectivity uses a "Mexican-hat" profile and encourages spatially proximal neurons to develop similar response properties due to their mutual excitation while neurons that are relatively far apart will experience mutual inhibition. This drives competition to create separate pools of functionally distinct neurons.

It was found that statistical decoupling of the learned output representations can occur between two objects through independent rotation, even when the objects are always presented together concurrently during training. Crucially, the features that comprise a view of an object are always presented together and these features are not regularly presented with a specific view of the alternate object. Independent rotation ensures that features between objects remain statistically decoupled. After learning, separate output representations for the two objects develop.

It was observed that a SOM network architecture is able to form output neurons that respond with complete view invariance to their preferred object. Although the competitive network architecture was also able to form neurons with object selective representations, they were not fully invariant over all views. Instead, these neurons primarily responded to small regions of contiguous views of their preferred object and a population of neurons is required to encode all the views of each object presented during training.

Method

Learning to respond to an object with transform invariance

A simplified example of the Continuous Transformation (CT) learning process is illustrated in Figure 1a and operates as follows. When an input pattern is presented during learning, as represented by three active neurons in the input layer (Figure 1, neurons 1, 2, and 3), it causes activity to propagate through the random feedforward connections to the output layer, where one of the neurons (Figure 1, neuron 8) wins the competition. The co-activation of neurons in the input and output layers causes their synaptic connection to become strengthed, according to a Hebbian learning rule,

$$\delta w_{ij} = \alpha y_i y_j \tag{1}$$

where δw_{ij} is the increment in the synaptic weight w_{ij}, y_i is the firing rate of the post-synaptic neuron i, y_j is the firing rate of the pre-synaptic neuron j, and α is the learning rate. To restrict and limit the growth of each neuron's synaptic weight vector, $\mathbf{w}_i$ for the ith neuron, its length is normalised at the end of each timestep during training as is usual in competitive learning (Hertz et al. 1991). Normalisation is required to ensure that the same set of neurons do not always win the competition. Neurophysiological evidence for synaptic weight normalization is provided by Royer and Pare (2003).

As the stimulus moves from position 1 to position 2 (shown in Figure 1b), it causes neurons 2, 3 and 4 to become active. Note that neuron 1 is no longer active. However, neurons 2 and 3 remain active and therefore increase the likelihood that the same neuron (neuron 8) in the output layer will become active and win the competition within the output layer. The synaptic connections are then strengthened once more and neuron 8 will now become associated with the activation of input neurons 2, 3 and 4. This process is continued, as shown in Figure 1c and is repeated as the object gradually shifts along the input layer.

In the case of a more complex model such as VisNet, input layer activation represents the output from the pre-processing input filter stage. The Hebbian learning rule only strengthens the synaptic connections from those V1 filters that are activated by the particular visual form of the object view currently presented. The weight vector of each of the first layer neurons gradually shifts during learning to point in the same direction as the V1 input pattern(s) to which it is learning to respond. After training, each neuron will respond in proportion to the similarity

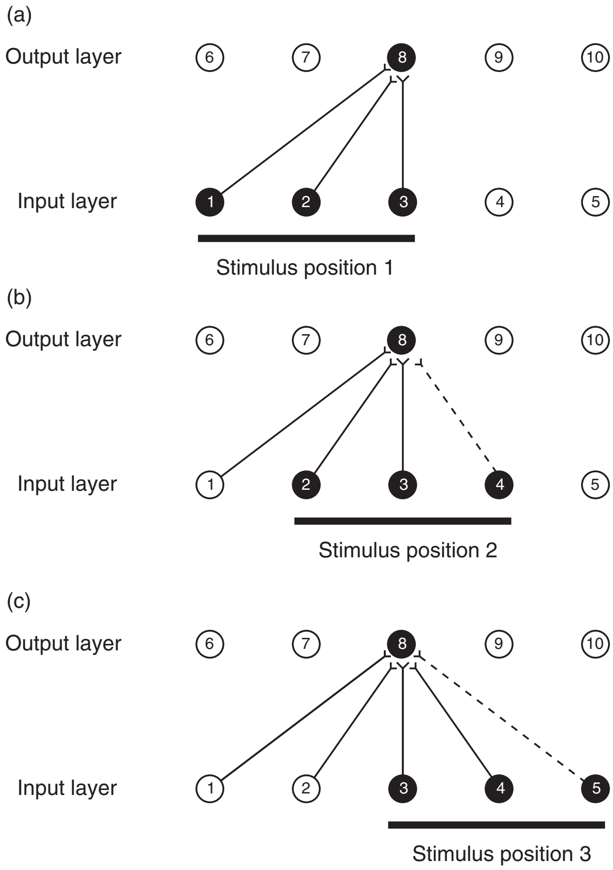

Figure 1. An illustration of how CT learning functions in a feed-forward one-layer network. Activation of overlapping neurons during the transformation of the object from position to position leads to the activation of the same neuron in the output layer. Connections are strengthened according to a Hebbian learning rule after each presentation of the stimulus.

between the current input pattern and the input pattern(s) the neuron learned to respond to during training; each first layer neuron computes its activation (Equation 3) according to the dot product of its weight vector and the current input pattern from the V1 input layer (Hertz et al. 1991).

A more comprehensive description of Continuous Transformation learning and simulation results in the context of invariant object recognition is provided by Stringer et al. (2006) and Perry et al. (2006).

The VisNet model

The original model architecture (VisNet) implemented by Wallis and Rolls (1997) that is used to investigate the object selective properties learned by independent rotation during training is based on the following: (i) A series of hierarchical competitive networks with local graded inhibition. (ii) Convergent connections to

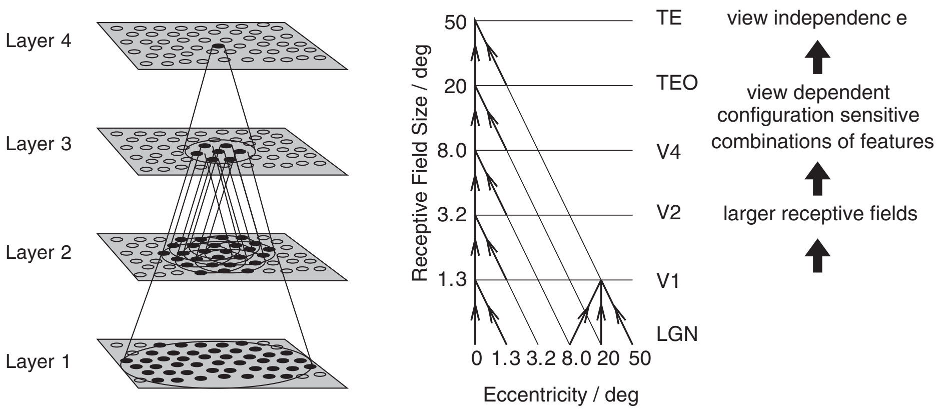

Figure 2. Left: Stylised image of the four layer network. Convergence through the network is designed to provide fourth layer neurons with information from across the entire input retina. Right: Convergence in the visual system V1: visual cortex area V1; TEO: posterior inferior temporal cortex; TE: anterior inferior temporal cortex (IT).

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	–	–

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.

In this paper, for the first time, we replaced the original competitive network architecture within each layer with a self-organising map (SOM). That is, some model simulations adjusted the local graded inhibition within a layer to incorporate short-range excitation and longer-range inhibition. This type of competition within a layer is known as a self-organising map, or SOM (von der Malsburg 1973; Kohonen 1982).

The original VisNet model consists of a hierarchical series of four layers of competitive networks, corresponding to V2, V4, the posterior inferior temporal cortex, and the anterior inferior temporal cortex, as shown in Figure 2. 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 approximately 67% of the connections from the preceding layer. The values used are given in Table 1.

Before the objects are presented to the network's input layer they are preprocessed by a set of input filters which accord with the general tuning profiles of

Table 2. Layer 1 connectivity. The numbers of connections from each spatial frequency set of filters are shown. The spatial frequency is in cycles per pixel.

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

simple cells in V1. The filters provide a unique pattern of filter outputs for each transform of each visual object, which is passed through to the first layer of VisNet. The input filters used are computed by weighting the difference of two Gaussians by a third orthogonal Gaussian according to the following:

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

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/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. Past neurophysiologcal research has shown that models based on difference-of-Gaussians functions are superior to those based on the Gabor function or the second differential of a Gaussian (Hawken and Parker 1987).

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,

$$h_i = \sum_j w_{ij} y_j \tag{3}$$

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.

Competitive network

The original VisNet model implemented a competitive network within each layer. Within each layer, competition is graded rather than winner-take-all, and is implemented in two stages. To implement lateral inhibition, the activation 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

$$I_{a,b} = \begin{cases} -\delta e^{-\frac{a^2+b^2}{\sigma^2}} & \text{if } a \neq 0 \text{ or } b \neq 0, \\ 1 - \sum\limits_{\substack{a\neq 0 \\ b\neq 0}} I_{a,b} & \text{if } a = 0 \text{ and } b = 0 \cdot \end{cases} \tag{4}$$

The lateral inhibition parameters are given in Table 3.

Self-organising map

In this paper, we have also run simulations with a self-organising map (SOM) implemented within each layer. In the case of the SOM architecture, short-range

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. Example lateral inhibition and excitation parameters for the SOM

Layer	1	2	3	4
Excitatory radius, σ_E	2.1	1.65	1.2	1.8
Excitatory contrast, δ_E	5.35	33.15	117.57	120.12
Inhibitory radius, σ_I	4.14	8.1	12.0	18.0
Inhibitory contrast, δ_I	1.5	1.5	1.6	1.4

excitation and long-range inhibition forms a Mexican-hat spatial profile and is constructed as a difference of two Gaussians as follows:

$$I_{a,b} = -\delta_I e^{\left[-\frac{a^2+b^2}{\sigma_I^2}\right]} + \delta_E e^{\left[-\frac{a^2+b^2}{\sigma_E^2}\right]} \tag{5}$$

Here, to implement the SOM, the activations h_i of neurons within a layer are convolved with a spatial filter, I, where δ_I controls the inhibitory contrast and δ_E controls the excitatory contrast. The width of the inhibitory radius is controlled by σ_I while the width of the excitatory radius is controlled by σ_E. a and b index the distance away from the centre of the filter. The lateral inhibition and excitation parameters are given in Table 4.

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

$$y = f^{sigmoid}(r) = \frac{1}{1 + e^{-2\beta(r-\alpha)}} \tag{6}$$

where r is the activation (or firing rate) after applying the lateral inhibition or SOM filter, 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 within each layer of neurons to control the sparseness of the firing rates. The sparseness a of the firing within a layer can be defined, by extending the binary notion of the proportion of neurons that are firing, as

$$a = \frac{\left(\sum_{i=1}^{N} y_i/N\right)^2}{\sum_{i=1}^{N} y_i^2/N} \tag{7}$$

where y_i is the firing rate of the ith neuron in the set of N neurons (Rolls and Treves 1990, 1998). For the simplified case of neurons with binarised firing rates $= 0/1$, the sparseness is the proportion $\in [0, 1]$ of neurons that are active. For example, to set the sparseness to, say, 4%, the threshold is set to the value of the 96th percentile point of the activations within the layer. The parameters for the sigmoid activation function are shown in Table 5. These are general robust values found to operate well for this experiment. They are similar to the standard VisNet sigmoid parameter

Table 5. Sigmoid parameters

Layer	1	2	3	4
Percentile	96	96	96	96
Slope β	190	40	75	26

values which were previously optimised to provide reliable performance (Stringer et al. 2006, 2007; Stringer and Rolls 2008).

Stimuli

To train the VisNet model we used three 3D rotating visual stimuli to create three separate pairings. Each pairing was run as a separate experiment and the objects used to create the pairings comprise a Cube, Dodecahedron and custom designed Hectoid. We used three separate pairings with three very different objects to ensure that our results were robust. Figure 3a shows the Cube and the Hectoid, Figure 3b shows the Cube and Dodecahedron and Figure 3c shows the Dodecahedron and the Hectoid. In each experiment, the same stimuli were also used during testing and were continuously rotating over 360°. Each object was designed and created in Electric Rain's vector based 3D modelling software, Swift 3D. Extra measures were taken to ensure that image outlines and borders were clearly defined to avoid any artefacts that may occur due to the size of the VisNet retina. Furthermore, ambient lighting with diffused light sources were used to ensure that different surfaces were shown with different illuminations.

Training procedure

The three pairings were presented to the network in separate experiments in exactly the same manner. The following section describes the process using the Cube and the Hectoid, but the same can also be said for the Cube and the Dodecahedron as well as the Dodecahedron and the Hectoid.

In order to simulate independent rotation we presented different views of object one, the Cube, with different views of object 2, the Hectoid. Specifically, we presented each view of the Cube with ten equally spaced views of the Hectoid. This process was repeated for the Hectoid, where every view was presented with ten equally spaced different views of the Cube. We ensured that each view of any given object was presented as frequently as any other view.

For example, in 'block 1' of training we presented a specific view of the Cube at an angle of 1° with ten views of the Hectoid at 1°, 37°, 73°, 109°, 145°, 181°, 217°, 252°, 289°, 325° respectively. In 'block 2' of training we presented a specific view of the Cube at an angle of 2° with ten views of the Hectoid at 2°, 38°, 74°, 110°, 146°, 182°, 218°, 253°, 290°, 326° respectively. This was repeated for every view of the Cube, therefore creating 360 Cube training 'blocks'. We also created 360 similar Hectoid training 'blocks'. In total, we presented 7200 object pairings to the network, and together these comprise one epoch. We explored the performance of the network with a SOM and competitive network architecture within each layer.

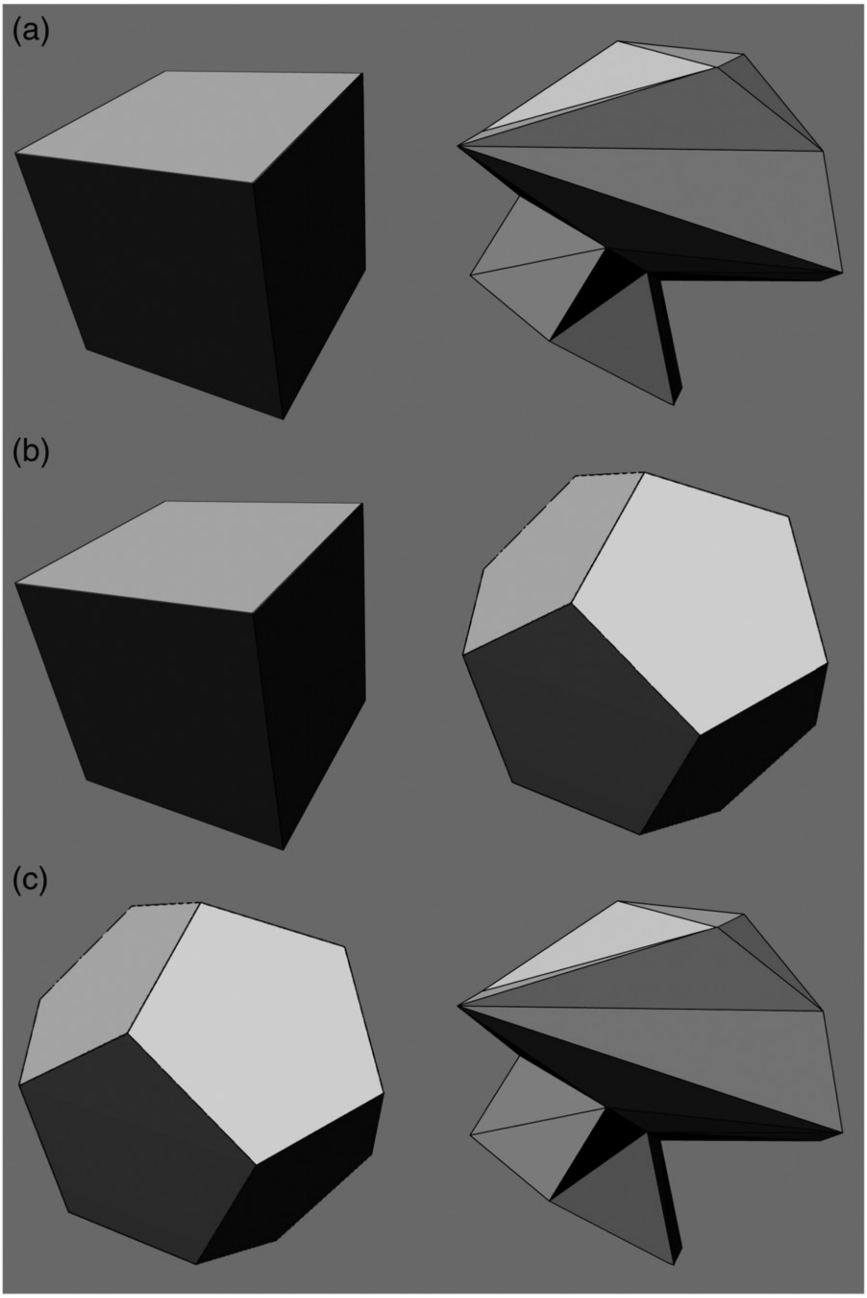

Figure 3. Example images of the three pairs of objects used to train the model. (a) Cube (left) and Hectoid (right); (b) Cube (left) and Dodecahedron (right); (c) Dodecahedron (left) and Hectoid (right). Each pair was presented as a separate experiment and the same objects were used during training and testing.

In all experiments, the learning rate of the model was set to 0.01 and the sparseness was set to 0.04. Other values were explored (not presented) and receive comment in the Discussion.

Testing procedure

To test the network's performance, for each experiment we presented each of the relevant objects in isolation. For example, all 360 views of either the Cube or the Hectoid were presented separately, and we recorded the firing rates from the output (fourth) layer. Results are displayed as polar plots and are collected from VisNet's output layer. The same can be said for the experiments with the Cube and the Dodecahedron pairing as well as the Dodecahedron and the Hectoid pairing.

Information measures

The network's ability to recognise a specific object during testing is assessed using two measures of information theory: single and multiple cell information. Full details on the application of these measures to VisNet are given by Stringer et al. (2006). These measures reflect the extent to which cells respond invariantly to an object over a number of different views (transforms), but respond differently to different objects. The single cell information measure is applied to individual cells in layer 4, and measures how much information is available from the response of a single cell about the stimlus that was presented. The single cell information measure for each cell shows the maximum amount of information that the cell conveys about any one object. This is computed using the following formula with details provided by Rolls et al. (1997) and Rolls and Milward (2000). The object-specific information $I(s, R)$ is the amount of information the set of responses R has about a specific object s, and is given by

$$I(s, R) = \sum_{r \in R} P(r|s) \log_2 \frac{P(r|s)}{P(r)}, \tag{8}$$

where r is an individual response from the set of responses R. However, the single cell information measure cannot give a complete assessment of VisNet's performance with respect to invariant object recognition. If all output cells learned to respond to the same object then there would in fact be relatively little information available about the set of objects S, and single cell information measures alone would not reveal this. To address these issues, we also calculate a multiple cell information measure, which assesses the amount of information that is available about the whole set of objects from a population of neurons.

Procedures for calculating the multiple cell information measure are described by Rolls et al. (1997) and Rolls and Milward (2000). In brief, from a single presentation of an object, we calculate the average amount of information obtained from the responses of all the cells regarding which object is shown. This is achieved through a decoding procedure that estimates which object s' gives rise to the particular firing rate response vector on each trial. A probability table of the real objects s and the decoded objects s' is then constructed. From this probability table, the mutual information is calculated as

$$I(S, S') = \sum_{s,s'} P(s, s') \log_2 \frac{P(s, s')}{P(s)P(s')}. \tag{9}$$

Multiple cell information values are calculated for the subset of cells which, according to the single cell analysis, have the most information about which object is shown. In particular, the multiple cell information is calculated from the first five cells for each object that had the most single cell information about that object. This results in a population of 10 cells given that there are two objects in the simulations presented below. Previous research (Stringer and Rolls 2000) found this to be a sufficiently large subset to demonstrate that invariant representations of each object presented during testing were formed, and that each object could be uniquely identified.

VisNet simulation results

In each experiment we trained the model with one of the three different pairings: Cube and Hectoid; Cube and Dodecahedron; Dodecahedron and Hectoid. Each pairing was presented as a separate experiment. In each experiment the objects within each pairing either rotated dependently together in lock-step during training or rotated independently during training. In the case of the Cube and the Hectoid pairing, we did this for both a self-organising map (SOM) architecture and a competitive network architecture. For the other two pairings, we only explored the network's performance with a self-organising map (SOM) architecture. Three different objects pairs were used in order to ensure that our results were robust.

We present the following results, showing that the network was unable to form separate representations of the objects when trained together in the dependent rotation condition. This was the case for both the SOM and the competitive network architectures. However, in the independent rotation condition, the network was able to form object selective representations for each of the objects, despite the fact that the objects were always presented rotating together during training. The effect of the SOM compared to the competitive network is explored in the context of this novel result.

Main results with SOM architecture

Figure 4 shows cell response polar plots of a 8 × 8 subset of the cells in the output layer after training using a SOM network architecture with the Cube and the Hectoid. Both the dependent and independent paradigms are presented. The figure is split into four quadrants forming two rows and two columns. Each quadrant comprises a subset of neurons from the model's output layer where 8 × 8 neurons are represented by 64 polar plots. Each degree in the polar plot represents an associated view of the test object. Firing rates are represented by the distance from the centre point. The left hand column comprises two plots that show the results of testing with the Cube and the right hand column comprises two plots that show

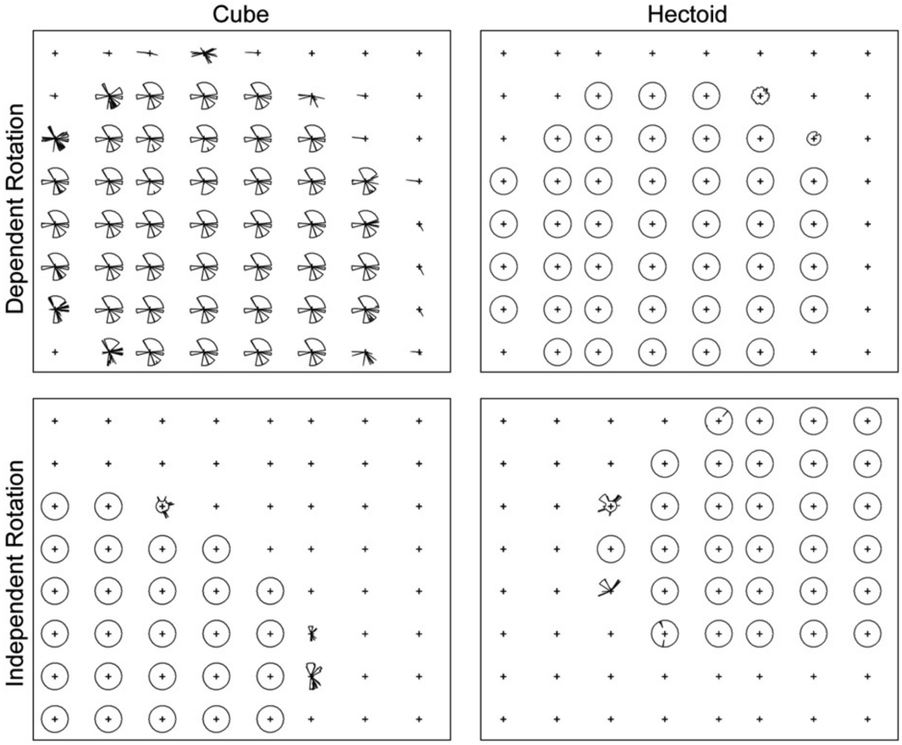

Figure 4. Cube-Hectoid pairing results with SOM network architecture. The figure shows the responses of a 8 × 8 subset of output layer cells after training with the dependent and independent rotation paradigms and testing with the Cube and Hectoid presented individually. The figure is divided into four quadrants. The left-hand column presents two plots comprising the cell responses when tested on the Cube and the right-hand column presents two plots comprising cell responses when tested with the Hectoid. The top row presents results after the dependent rotation training paradigm, and the bottom row presents results after the independent rotation training paradigm. Crucially, for the purposes of comparison, the same set of cells are shown within each training paradigm (rows) when tested with the Cube or the Hectoid. In the dependent rotation paradigm, cells learn to respond to both objects as if they were one with almost complete view invariance, and fail to build object specific responses. In the independent rotation paradigm, object selective responses form, that are also view invariant. Specifically, when the Cube and the Hectoid rotate independently during training, the cells in the lower left corner of the 8 × 8 array learn to respond to the Cube and the cells in the top right corner learn to respond to the Hectoid.

results of testing with the Hectoid. The top row of plots comprises results from the dependent rotation training paradigm and the bottom row comprises results from the independent rotation training condition. Within each paradigm (dependent or independent rotation), the same cells are plotted when testing on both the Cube and the Hectoid.

It is clear that in the dependent rotation condition, cell response profiles exhibit a large degree of overlap between the two objects. That is, cells that respond invariantly to the Hectoid also respond in-kind to the Cube, as if they were the same object. The network has not formed object selective representations, failing to build separate representations for the Cube and the Hectoid. However, in the

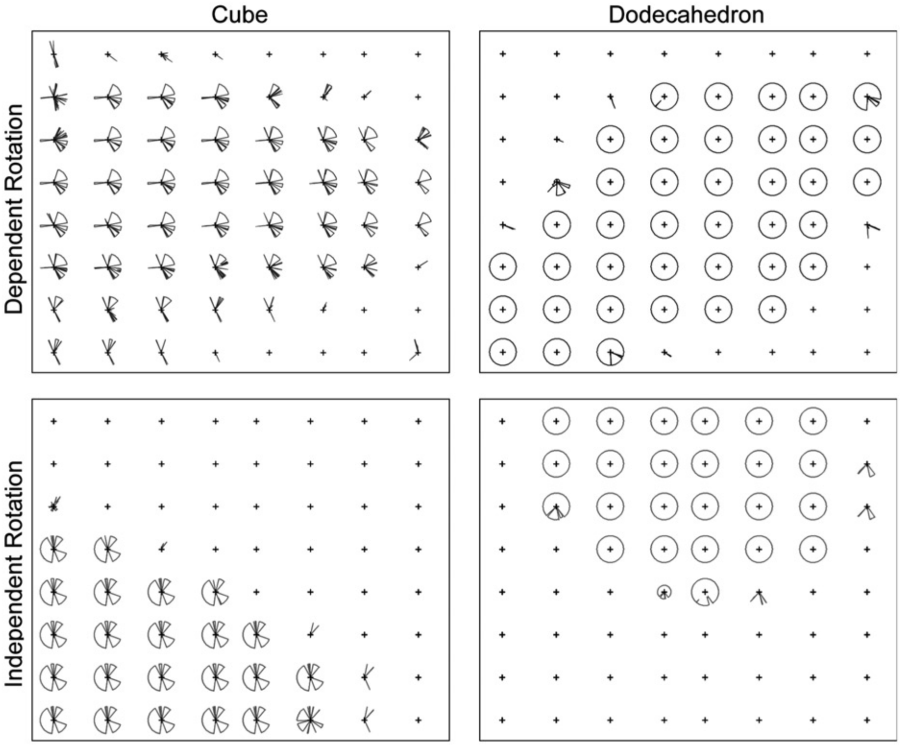

Figure 5. Cube-Dodecahedron pairing results with SOM network architecture. Conventions as in Figure 4. This figure shows that in the dependent case, the network has failed to build object selective responses while in the independent case, the network has learned to develop object selective responses. Although the cells responding preferentially to the Cube do so for the majority of its views, they do not represent all 360 views. Instead, a separate pool of cells (not shown in this figure) respond to the remaining views.

independent rotating paradigm, most individual output cells have learned to respond invariantly to either the Cube or the Hectoid (but not both), despite the fact that the objects were always presented together during training. Crucially, the network has formed object selective representations for each object where cells respond to their preferred object and not to the alternative object.

Figures 5 and 6 show the same results but for the Cube and the Dodecahedron, and the Dodecahedron and the Hectoid, respectively. It can clearly be seen that the response plots reflect the results found with the Cube and the Hectoid. That is, in the dependent rotation condition VisNet fails to develop object selective representations, while in the case of the independent rotation, cells in the output layer learn to respond exclusively to their preferred object. In the case of the Cube and the Dodecahedron, cells that respond preferentially to the Cube did not develop complete invariance and instead there is another cluster of cells that respond to the remaining views (not shown).

Information analysis plots are provided for when the SOM network was trained with object pairs rotating independently and rotating dependently. Information results with the Cube-Hectoid, Cube-Dodecahedron and Dodecahedron-Hectoid pairings are presented in Figures 7–9, respectively.

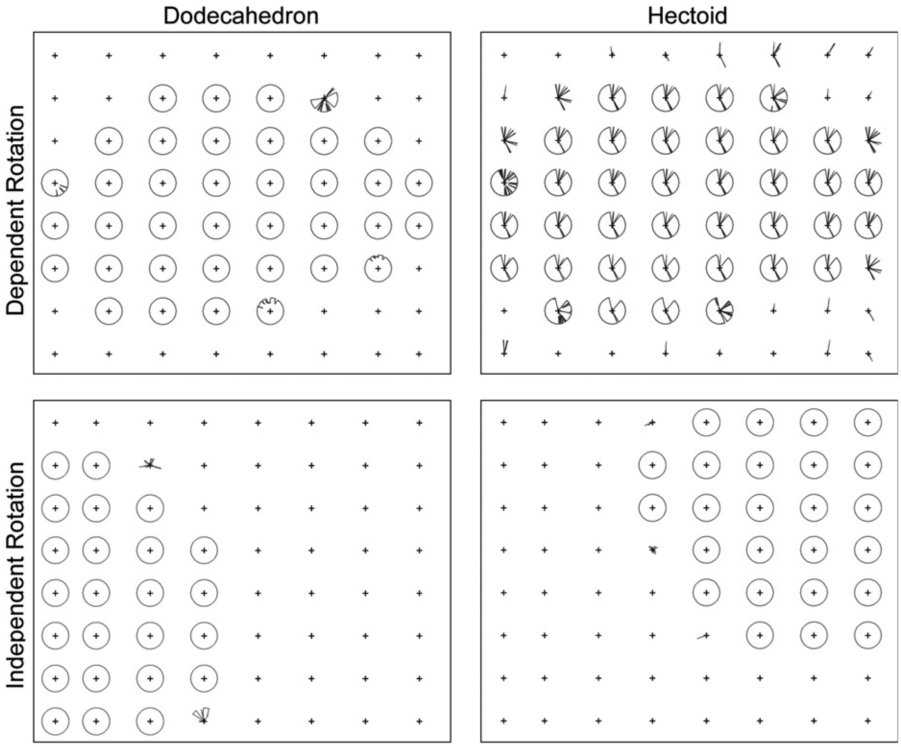

Figure 6. Dodecahedron-Hectoid pairing results with SOM network architecture. Conventions as in Figure 4. This figure shows that in the dependent case, the network has failed to build object selective responses while in the independent case, the network has learned to develop object selective responses.

In all three figures, the left-hand plot shows the single cell information analysis. This was calculated to confirm whether the network had developed neurons that responded exclusively to their preferred object over all 360 views. In the case of the Cube-Hectoid pairing, it may be observed that in the case of independent rotation, 72 neurons provided maximal information of 1 bit. Conversely, in the case of dependent rotation, the amount of information is much lower. Similar results are presented for the other two pairings, although in the case of the Cube-Dodecahedron pairing (Figure 8), fewer cells provide maximal information in the independent rotation condition because there were fewer completely invariant responses for the Cube. This can be observed in Figure 5.

In summary, the single cell information analysis plots confirm that independent rotation is enough to cause cells to develop object selective responses that are also invariant. Crucially, in the dependent rotation paradigm, cells are unable to build object selective representations, and fail to discriminate between the objects due to the statistical coupling between them present in the training input.

For all three Figures 7–9, the right-hand plot shows multiple cell information analysis measures. In the independent rotation paradigm, a very small number of cells are required to reach the maximum information of 1 bit, confirming that the

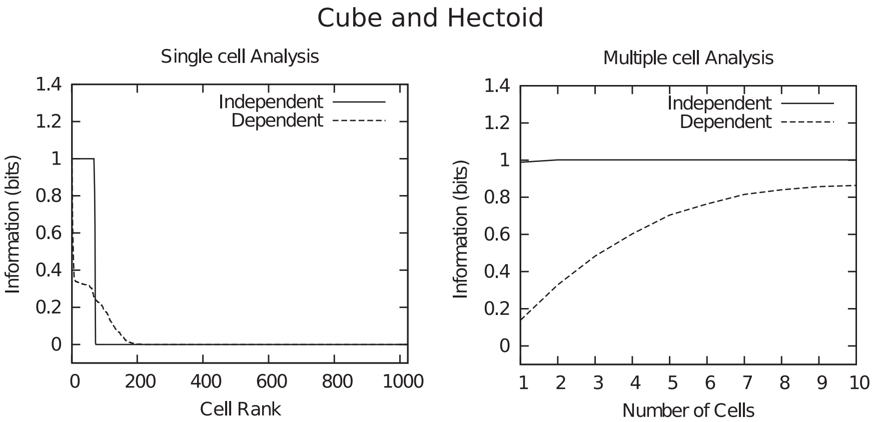

Figure 7. Cube-Hectoid pairing information analysis with the SOM network architecture. Single (left-hand plot) and multiple (right-hand plot) cell information results obtained when the trained SOM network was tested with the Cube and Hectoid rotating separately. Results are presented for both the independent (unbroken line) and dependent rotation (dashed line) paradigms. The single cell information measure for all 4th layer neurons ranked in order of the amount of information they convey about the objects is shown. In the independent rotation paradigm, the network develops many object selective and invariant neurons attaining the maximum level of single cell information of 1 bit. The dependent rotation condition produces less information and no cells reached the maximal information. The multiple cell information measure reaches the maximum level of 1 bit in the independent condition. This confirms that the network represents both trained objects and maximal information is obtained. The dependent condition provides a comparison, where the network has failed to separate the two objects presented during training.

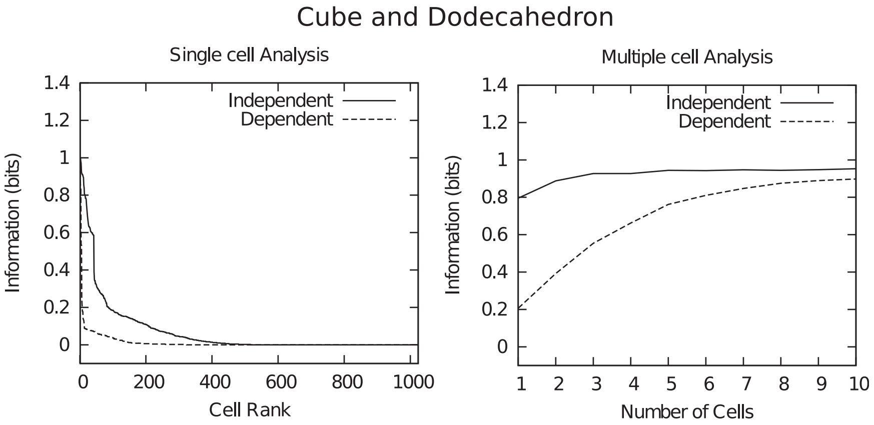

Figure 8. Cube-Dodecahedron pairing information analysis with the SOM network architecture. It may be observed that after independent rotation during training, the single cell information is relatively low. This is because cells did not learn to respond to all 360 views of the Cube and were less invariant.

network learns to represent both objects used during training. In the dependent rotation condition more cells are required and maximum information is never achieved because cells in the network have failed to separate the two objects used during training.

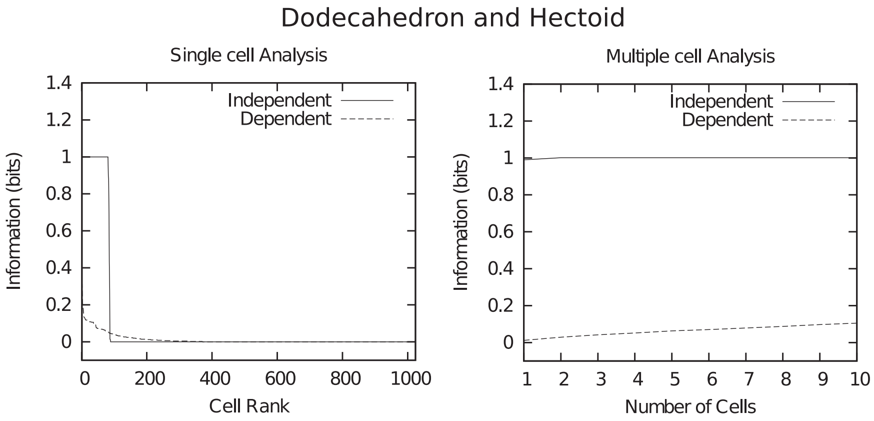

Figure 9. Dodecahedron-Hectoid pairing information analysis with the SOM network architecture. In the independent condition, maximal information is obtained in both the single cell and multiple cell analyses. In the dependent rotation condition, the network performs badly and is unable to build object selective representations. This results in very little information in both the single cell and multiple cell analyses.

In summary, the network produces markedly different responses in the independent rotation paradigm when compared to the dependent paradigm. In the dependent rotation condition, the network learns to build combined representations for both objects after it had been trained. That is, cells learn to respond to both objects, such as the Cube and the Hectoid, with a high degree of invariance, but with no object selectivity. Crucially, in the independent rotation condition, the network forms invariant and object selective representations for both objects. This novel result is also explored in the context of a competitive network next.

Competitive Network trained with Cube and Hectoid

Figure 10 shows four plots each consisting of an 8 × 8 sample of cells in the model's output layer after training using a competitive network architecture. Each set of polar plots show the responses of the output layer neurons when tested with the Cube and the Hectoid rotating individually, for both the dependent and independent training paradigms. Similar to Figure 4, the left hand column shows results when tested with the Cube and the right hand column shows results when tested with the Hectoid. The top row presents results after training with the dependent rotation paradigm and the bottom row presents results after training with the independent rotation paradigm.

In the case of the dependent rotation condition, the network was trained with both objects rotating together in lock-step and tested with each object separately. The results show that the cells learn to respond invariantly to both objects but fail to build object selective representations. That is, cells respond to both the Cube and the Hectoid, regardless of viewing angle. There are no cells that fire exclusively to just one of the objects.

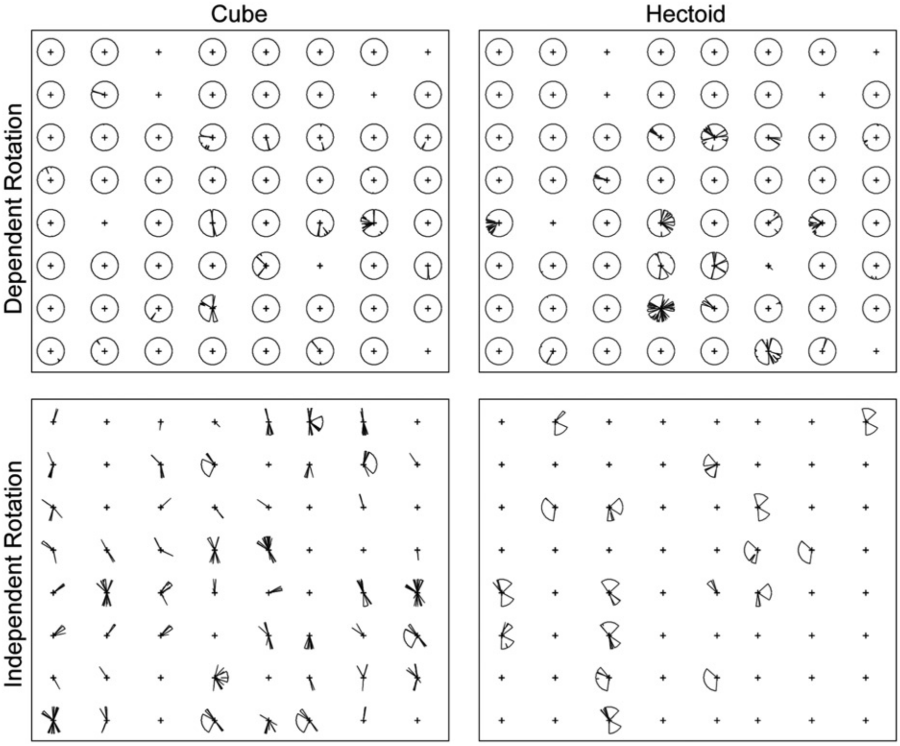

Figure 10. Results with competitive network architecture. The figure shows the responses of a 8 × 8 subset of output cells after training with the dependent and independent rotation paradigms and testing with the Cube and Hectoid individually. Conventions as for Figure 4. In the dependent rotation paradigm, cells learn to respond to both objects as if they were one, with complete view invariance to all 360 views, and fail to build object specific responses. In the independent rotation paradigm, object selective responses form. However, cells do not develop complete view invariance, primarily responding to contiguous views of their preferred object.

Results from the independent rotation condition show a markedly different profile to that of the dependent rotation condition. In the independent rotation condition, the network has developed cells that exclusively respond to contiguous portions of the view-space of just their preferred object. That is, cells do not respond to more than one object. These cells do not respond completely invariantly to their preferred object and instead represent a portion of the total view-space. All views of the preferred object are encoded in this manner, such that a group of cells together are able to exclusively identify a specific object from any viewing angle. Both the Cube and the Hectoid are fully represented in this way.

We calculated information analysis measures on the competitive network architecture to help quantify the network's performance after training. Figure 11 shows the single and multiple cell information analysis plots when the network was trained with object pairs rotating independently and rotating dependently. Because the competitive network did not build invariant representations, the information analysis measures provide low levels of information, even in the case of the

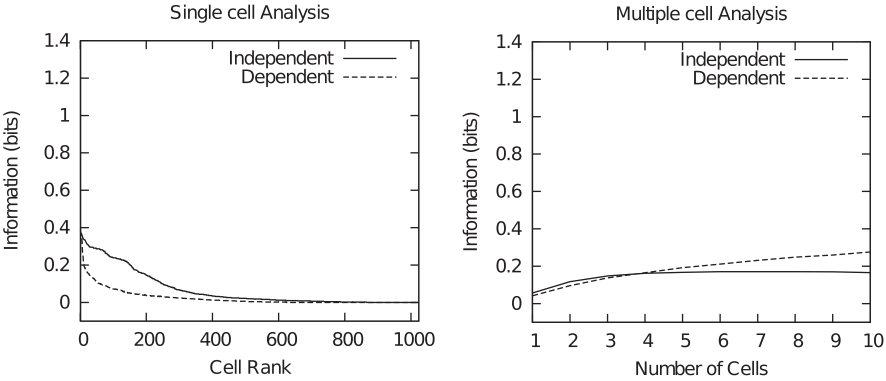

Figure 11. Information analysis with the competitive network architecture. Single (left-hand plot) and multiple (right-hand plot) cell information results obtained when the trained competitive network was tested with the Cube and Hectoid rotating separately. Results are presented for both the independent (unbroken line) and dependent rotation (dashed line) paradigms. The single cell information measure for all 4th layer neurons ranked in order of the amount of information they convey about the objects is shown. Although in the independent rotation paradigm the network was able to form object selective responses, both training conditions produced lower levels of information for both the single and multiple cell information analysis. This can be attributed to the lack of object selectivity in the dependent rotation paradigm, and the lack of invariance achieved in the independent rotation paradigm.

independent rotation paradigm. The left-hand plot shows the single cell information analysis and it may be observed that although the independent rotation paradigm provides more information than the dependent condition, it does not reach the maximum information of 1 bit because cells do not respond invariantly to their preferred object.

In the independent rotation paradigm, even though cell response plots presented in Figure 10 reveal cells respond exclusively to their preferred object and not to the alternative object, the multiple cell information analysis produces low levels of information. This is also the case for the dependent motion condition. Maximum information is achieved when cells respond invariantly and exclusively to their preferred object. As such, low levels of information can be attributed to the lack of object selectivity in the dependent rotation paradigm, and the lack of invariance achieved in the independent rotation paradigm.

The effect of the SOM width

We explored the effect of the SOM width by varying the radii of the excitatory and inhibitory components of the SOM filter. To investigate the overall effect the SOM filter width has on the learning and self-organisation of the network, we maintained a constant ratio between the excitatory and inhibitory components. A range of SOM widths were explored. Figure 12 presents the firing responses of output layer cells, after training the network with the cube and Hectoid rotating independently during training. In particular, the figure shows results with narrow and wide SOM filters alongside the original, intermediate, SOM width results, initially presented in Figure 4. The values for the intermediate SOM width are presented in Table 4 and

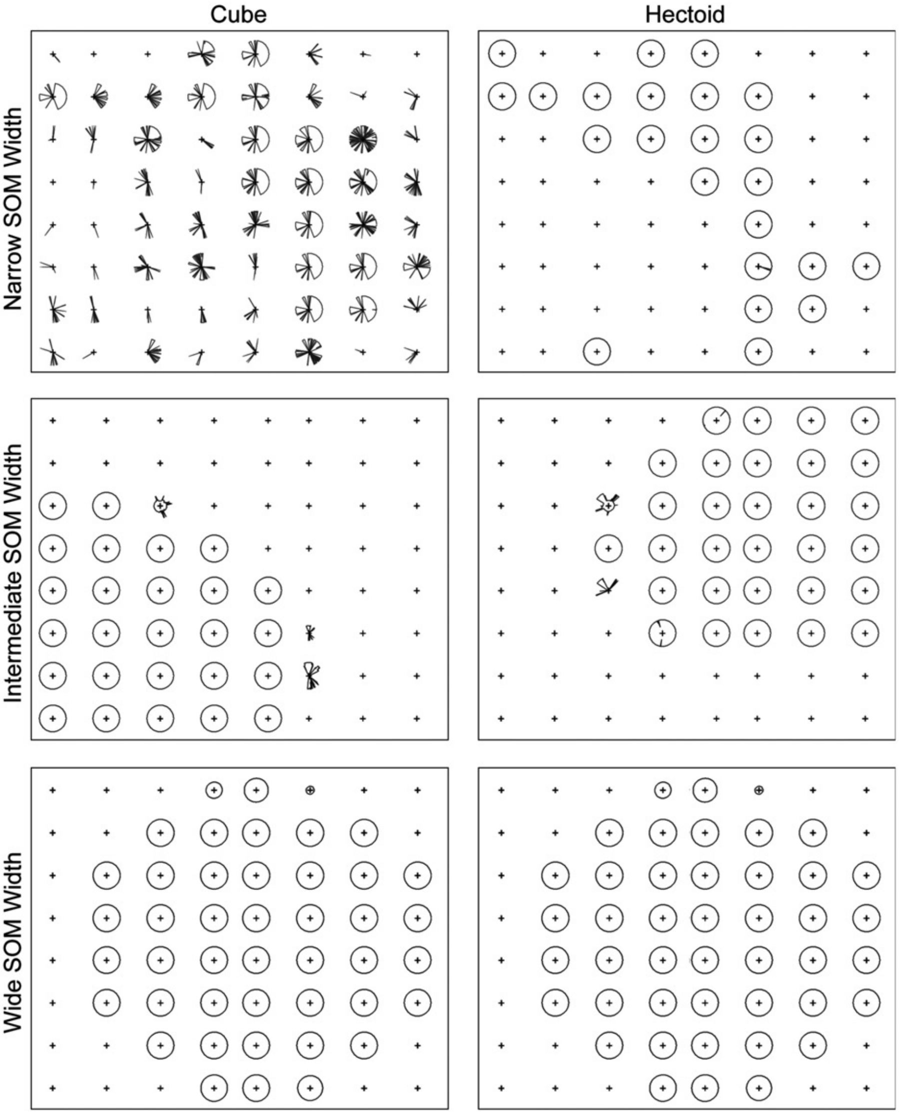

Figure 12. Cube-Hectoid pairing results after independent rotation during training with three different SOM widths: narrow, intermediate and wide. Conventions as in Figure 4. When the SOM width is relatively narrow, fully invariant cells develop for the Hectoid but not for the Cube. However, as the SOM width is increased, the number of invariant cells also increases. With an intermediate SOM width, object selective and fully invariant cells form for both the Cube and the Hectoid. When the SOM width is relatively wide, the output cells learn to represent both objects invariantly but fail to produce object selective responses. The same set of cells are shown within each training paradigm.

are three times as wide as the 'narrow' SOM width. The 'wide' SOM width is five times as wide as the narrow SOM width. Corresponding information results are shown in Figure 13.

These results show that when the SOM width is relatively narrow, the network builds invariant responses to one object, but not the other. However, as the width of

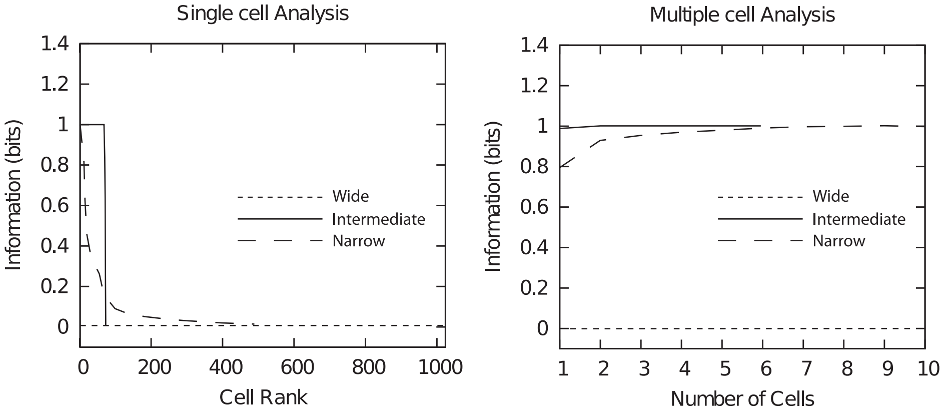

Figure 13. Information analysis of Cube-Hectoid pairing results after independent rotation during training, with the SOM architecture with three different widths. Single (left-hand plot) and multiple (right-hand plot) cell information results are shown for the narrow (dashed line), intermediate (unbroken line) and wide (dotted line) SOM widths when tested with the Cube and Hectoid rotating separately after training. A small amount of information is available in the narrow SOM width condition, while maximal information is achieved in the intermediate SOM width condition. Zero information is obtained when the SOM is wide because cells learn to respond in an identical manner to both the Cube and the Hectoid.

the SOM increases so does the number of object selective invariant neurons for both objects. When the SOM width becomes too wide, output cells respond invariantly to all views of both objects at the same time. That is, they are not object selective. Possible causes for these results are mentioned in the Discussion.

Discussion

Previous results from VisNet have shown that separate object representations can develop in the output layer when there is a sufficient level of decoupling between a large number of individual objects presented as random pairs during training. Crucially, there must be a large enough superset of objects to pick from when creating the different object pairings (Stringer et al. 2007; Stringer and Rolls 2008). The aim of this research was to establish whether independent rotation might play an important role for helping the primate visual system to build separate representations of just two objects always presented together. We also investigated the differences in performance between a self-organizing map architecture and competitive network.

In this paper, we confirm that the network is able to successfully form object selective representations when trained with objects that rotate independently. This applied to both a SOM architecture and a competitive network architecture.

In the independent rotation paradigm, different views of one object, say the Cube, were presented with different views of the Hectoid. Because the features that make up a specific view of an object are always presented together 100% of the time, they are maximally correlated. On the other hand, these features are not well correlated to any of the features of the alternate object. That is, during training, a particular view of an object was presented with ten different views of the alternate object. Particular pairs of views are seen together relatively infrequently. The decoupling between the different views of the two objects is sufficient for a particular output cell to strengthen its connection with a specific view of the Cube, without necessarily strengthening connections associated with a particular view of the Hectoid.

When the objects rotate together in lock-step, both network architectures failed to build representations specific to either object. Instead, VisNet built a single representation, as if both objects were in fact one. This result is a consequence of the statistical coupling between the objects that occurs due to the dependent motion. With dependent motion, each view of the Cube is always presented with the same view of the Hectoid. If a particular view of the Cube happens to strongly activate a given output neuron due to the initially random feedforward synaptic weights, then not only will the synaptic weights from the Cube view strengthen, but the weights from the Hectoid view will too. After multiple training epochs, the output neuron will learn to respond equally well to views of both objects as if they are one.

The CT learning effect is able to bind together views that share a high degree of spatial overlap and helps form view invariant cells. We only observed view invariant cells in the case of the SOM architecture. The competitive network architecture produced contiguous blocks of activation over a limited range of object views spanning about 45°. A network exhibiting these types of cell responses requires a population of cells to encode the whole view-space. In the case of the SOM paradigm, there were mutually excitatory lateral connections that cause cells with similar response properties to form close together. For example, assume that Cell A responds to a contiguous range of views of an object, from 90° to 135°. Assume that Cell B responds preferentially to a range of views of the same object spanning 100° to 145°. These two cells will form close together and as training progresses, cell A will cause cell B to become active when view 90 is presented during training, and similarly, cell B will cause cell A to become active when view 145 is present during training. The synaptic connections will strengthen according to Hebbian learning and this may have an effect of broadening the tuning profiles of both cells until the network reaches a trained state where both cells respond invariantly over the whole range of views.

The sparseness of the network defined in Equation 7 was identified as one of the key parameters for developing object selective invariant representations with the SOM network paradigm. When the sparseness value was raised to high levels it promoted excessive broadening of the cell response profiles. Cells began to respond to both objects with increasing invariance. It was observed that a sparseness value of 0.04 creates broad response profiles that remain object selective. At this sparseness value, the SOM filter is able to distribute neurons successfully, therefore developing functional maps where a single cell response is similar to the average response of its neighbours. The network was robust to different learning rates.

We also explored the effect of different SOM filter widths. The primary effect of the SOM is to cause cells with similar response properties to develop close together in the output layer and cells with different responses to form in separate clusters. Adjacent cells influence one another reciprocally, gradually becoming active to the particular views that their neighbours respond to best. This has the general effect of amplifying the Hebbian learning mechanism and broadens the cells' response profiles so that they learn to respond to an increasing number of similar views. Even with the narrow SOM, this effect still helps build invariant responses in the output layer relative to a competitive network with comparable parameters. As the SOM increases in width, the network learns to build increasingly invariant responses. However, when the SOM is very wide it forces the output cells to respond to all views of both objects due to overwhelming levels of mutual excitation between cells.

In summary, we have presented a mechanism for how the visual system could self-organise when presented with complex natural scenes, using only independent movement as a method to build view invariant, object specific, representations. Early research with VisNet focussed on presenting just one object at a time (Stringer and Rolls 2000). More recent research required a large superset of a number of different objects from which to present random object pairings. For the first time, we have shown that it is possible for the network to develop completely view invariant, object selective, responses when trained with only two objects. Crucially, the objects must move independently from one another. We also implemented a SOM network architecture for the first time within each layer of the VisNet model, and have shown that this causes cells with similar response profiles to form close together and develop completely view invariant responses. This observation is documented in vivo, where neurons in IT cortex responding preferentially to similar features, can be grouped together (Tsunoda et al. 2001; Kiani et al. 2007).

Acknowledgements

This research was supported by the Economic and Social Research Council via J.M. Tromans.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

References

Booth M, Rolls ET. 1998. View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. Cereb Cortex 8(6):510–523.

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

Gattass R, Sousa A, Gross C. 1988. Visuotopic organization and extent of v3 and v4 of the macaque. J Neurosci. 8(6):1831–1845.

Gegenfurtner KR, Kiper DC, Levitt JB. 1997. Functional properties of neurons in macaque area v3. Journal of Neurophysiology 77(4):1906–1923.

Goodale MA, Milner A. 1992. Separate visual pathways for perception and action. Trends in Neurosciences 15(1):20–25.

Hasselmo ME, Rolls ET, Baylis GC, Nalwa V. 1989. Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey. Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation Cérébrale 75(2):417–429, PMID: 2721619.

Hawken MJ, Parker AJ. 1987. Spatial properties of neurons in the monkey striate cortex. Proceedings of the Royal Society of London. Series B, Biological Sciences 231(1263):251–288.

Hertz JA, Krogh AS, Palmer RG. 1991. Introduction to the Theory of Neural Computation, Volume I. Westview Press.

Ito M, Tamura H, Fujita I, Tanaka K. 1995. Size and position invariance of neuronal responses in monkey inferotemporal cortex. J Neurophysiol 73(1):218–226.

Kiani R, Esteky H, Mirpour K, Tanaka K. 2007. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. J Neurophysiol 97(6):4296–4309.

Kobatake E, Tanaka K. 1994. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. J Neurophysiol 71(3):856–867.

Kohonen T. 1982. Self-organized formation of topologically correct feature maps. Biological Cybernetics 43(1):59–69.

Levitt JB, Kiper DC, Movshon JA. 1994. Receptive fields and functional architecture of macaque v2. Journal of Neurophysiology 71(6):2517–2542.

Op De Beeck H, Vogels R. 2000. Spatial sensitivity of macaque inferior temporal neurons. The Journal of Comparative Neurology 426(4):505–518, PMID: 11027395.

Perrett D, Oram M. 1993. Neurophysiology of shape processing. Image and Vision Computing 11(6):317–333.

Perry G, Rolls ET, Stringer SM. 2006. Spatial vs temporal continuity in view invariant visual object recognition learning. Vision Research 46(23):3994–4006.

Rolls E, Treves A. 1990. The relative advantages of sparse versus distributed encoding for associative neuronal networks in the brain. Network: Computation in Neural Systems 1(4):407–421.

Rolls ET, Cowey A, Bruce V. 1992. Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. Philosophical Transactions: Biological Sciences 335(1273):11–21.

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 Computation 12(11):2547–2572, PMID: 11110127.

Rolls ET, Treves A. 1998. Neural Networks and Brain Function. 1st ed. USA: Oxford University Press.

Rolls ET, Treves A, Tovee MJ, Panzeri S. 1997. Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. 4:309–333.

Royer S, Pare D. 2003. Conservation of total synaptic weight through balanced synaptic depression and potentiation. Nature 422(6931):518–522.

Stringer S, Perry G, Rolls E, Proske J. 2006. Learning invariant object recognition in the visual system with continuous transformations. Biological Cybernetics 94(2):128–142.

Stringer SM, Rolls ET. 2000. Position invariant recognition in the visual system with cluttered environments. Neural Networks 13(3):305–315.

Stringer SM, Rolls ET. 2008. Learning transform invariant object recognition in the visual system with multiple stimuli present during training. Neural Netw. 21(7):888–903, ACM ID: 1411863.

Stringer SM, Rolls ET, Tromans JM. 2007. Invariant object recognition with trace learning and multiple stimuli present during training. Network: Computation in Neural Systems 18(2):161.

Tovee MJ, Rolls ET, 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(3):1049–1060.

Tsunoda K, Yamane Y, Nishizaki M, Tanifuji M. 2001. Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. Nat Neurosci 4(8):832–838.

von der Malsburg C. 1973. Self-organization of orientation sensitive cells in the striate cortex. Kybernetik 14(2):85–100, PMID: 4786750.

Wallis G, Rolls ET. 1997. Invariant Face and Object Recognition in the Visual system. Progress in Neurobiology.