Non-accidental properties, metric invariance, and encoding by neurons in a model of ventral stream visual object recognition, VisNet

Edmund T. Rolls (1,2) and W. Patrick C. Mills(1)
(1) Oxford Centre for Computational Neuroscience, Oxford, UK
www.oxcns.org
Edmund.Rolls@oxcns.org
and (2) University of Warwick, Department of Computer Science, Coventry, UK

Neurobiology of Learning and Memory (2018)

Key words: visual object recognition; non-accidental properties; visual coding; unsupervised learning; invariant representations; inferior temporal visual cortex; VisNet; trace learning rule; slow learning

Running title: Invariant visual object recognition

Corresponding author: Professor E T Rolls, Oxford Centre for Computational Neuroscience, Oxford, UK. Edmund.Rolls@oxcns.org Url: www.oxcns.org.

Abstract

When objects transform into different views, some properties are maintained, such as whether the edges are convex or concave, and these non-accidental properties are likely to be important in view-invariant object recognition. The metric properties, such as the degree of curvature, may change with different views, and are less likely to be useful in object recognition. It is shown that in a model of invariant visual object recognition in the ventral visual stream, VisNet, non-accidental properties are encoded much more than metric properties by neurons. Moreover, it is shown how with the temporal trace rule training in VisNet, non-accidental properties of objects become encoded by neurons, and how metric properties are treated invariantly. We also show how VisNet can generalize between different objects if they have the same non-accidental property, because the metric properties are likely to overlap. VisNet is a 4-layer unsupervised model of visual object recognition trained by competitive learning that utilizes a temporal trace learning rule to implement the learning of invariance using views that occur close together in time. A second crucial property of this model of object recognition is, when neurons in the level corresponding to the inferior temporal visual cortex respond selectively to objects, whether neurons in the intermediate layers can respond to combinations of features that may be parts of two or more objects. In an investigation using the four sides of a square presented in every possible combination, it was shown that even though different layer 4 neurons are tuned to encode each feature or feature combination orthogonally, neurons in the intermediate layers can respond to features or feature combinations present is several objects. This property is an important part of the way in which high capacity can be achieved in the four-layer ventral visual cortical pathway. These findings concerning non-accidental properties and the use of neurons in intermediate layers of the hierarchy help to emphasise fundamental underlying principles of the computations that may be implemented in the ventral cortical visual stream used in object recognition.

1 Introduction

It has been proposed that non-accidental properties of objects as they transform into different views are useful for view-invariant object recognition processes, and that metric properties are much less useful (Biederman 1987, Biederman & Gerhardstein 1993, Kayaert, Biederman & Vogels 2005, Amir, Biederman & Hayworth 2011, Kayaert, Biederman, Op de Beeck & Vogels 2005). Non-accidental properties (NAPs) are relatively invariant over rotations in depth (Biederman 1987). A NAP is an image property, such as the linearity of a contour or the cotermination of a pair of contours, that is unaffected by rotation in depth, as long the surfaces manifesting that property are still present in the image (Lowe 1985). NAPs can be distinguished from metric properties (MPs), such as the aspect ratio of a part or the degree of curvature of a contour, which do vary continuously with rotation in depth. There is psychophysical and functional neuroimaging evidence that supports this distinction (Kayaert, Biederman & Vogels 2005, Amir et al. 2011, Kayaert, Biederman, Op de Beeck & Vogels 2005). In addition, there is evidence that neurons in the lateral occipital cortex and inferior temporal visual cortex of the macaque are tuned more to non-accidental than to metric properties (Kim & Biederman 2012, Vogels, Biederman, Bar & Lorincz 2001). So far, the use of non-accidental properties of objects in models of how the primate ventral visual system implements invariant visual object recognition has been relatively little studied. Also relatively uninvestigated is how single neurons in these models of invariant object recognition may represent non-accidental properties and generalize across metric changes in them, and how these representations may be formed.

In the present investigation, we analyzed how non-accidental properties vs metric properties are encoded by neurons in a 4-layer unsupervised model, VisNet, of visual object recognition trained by competitive learning that utilizes a temporal trace learning rule to implement the learning of invariance using views that occur close together in time (Rolls 2012, Rolls 2016, Wallis & Rolls 1997). This model captures many aspects of how the ventral visual system learns and represents view-invariant representations of objects, including being unsupervised (unlike deep learning (LeCun, Kavukcuoglu & Farabet 2010, LeCun, Bengio & Hinton 2015, Bengio, Goodfellow & Courville 2017)), utilizing only 4 layers in the hierarchy after the primary visual cortex (V1) (corresponding approximately to V2, V4, posterior inferotemporal cortex, and anterior inferotemporal cortex); forming sparse distributed representations; and forming transform-invariant representations by layer 4 (Rolls 2012, Rolls 2016, Wallis & Rolls 1997). The hypothesis being tested is that the VisNet architecture would enable representations to be formed of non-accidental properties of visual stimuli; be relatively insensitive to metric properties; and to generalize across metric properties (for a given non-accidental property) present in different objects even without explicit training that the two different objects had some property in common. This latter hypothesis is described in more detail in Section 2.7.2. A previous investigation with a version of HMAX fundamentally modified to include a temporal trace rule did show performance that reflected NAPs better than metric properties, but the concepts involved were not clearly illustrated, and generalization to objects with metric properties other than those trained for an object was not demonstrated (Parker &

Serre 2015), both of which are addressed here with other issues relating to feature binding and transform invariance considered elsewhere (Robinson & Rolls 2015). Further, although nodes in a deep learning architecture trained by backpropagation of error may respond preferentially to NAP compared to metric properties (Kubilius, Bracci & de Beeck 2016), this may reflect just that NAPs better distinguish objects seen from different views, rather than providing any insight into the mechanisms that implement NAP selectivity and metric property invariance in the brain, as the mechanisms involved in deep learning are so different from those involved in neuronal computation in the brain (Rolls 2016, Marcus 2018).

We also investigated a second key property of a feature hierarchy object recognition system (illustrated in Fig. 1), namely that it may be possible at intermediate layers of the network to form feature combination neurons that can be useful for a number of different objects in each of which a particular feature combination may be present. If this is a property of encoding in the ventral visual system, this would help the capacity in terms of the number of objects that can be encoded to be high, because feature combination neurons in the intermediate layers could be used for many different objects, with orthogonal representations of whole objects only made explicit in neuronal firing at a late or the last stage of processing. We tested the hypothesis that this could be implemented by the relatively simple and biologically plausible architecture of VisNet by training VisNet to identify separately every single feature and feature combinations of an object by its final layer (4); and then examining whether in VisNet's intermediate layers (2 and 3) single neurons typically responded to features or low-order feature combinations that were components of several whole objects represented orthogonally in layer 4. The underlying overall hypothesis here is that at any layer of VisNet neurons need only learn low-order feature combinations; and that if this is repeated over several layers, then by the last layer neurons will become object selective (Rolls 1992, Rolls 2012, Rolls 2016). This overcomes any need at all for neurons to respond to high order feature combinations at any one layer, which would be biologically implausible as it would require so many neurons. At the same time, this feature-hierarchy approach would enable neurons in the final layer or layers to respond as if they were sensitive to a high order combination of features, that is, to be quite object selective (though using a sparse distributed representation), as has been found for neurons in the inferior temporal visual cortex (Rolls & Tovee 1995, Rolls, Treves, Tovee & Panzeri 1997, Booth & Rolls 1998, Franco, Rolls, Aggelopoulos & Jerez 2007, Rolls, Treves & Tovee 1997, Rolls 2012, Rolls 2016).

2 Methods

2.1 Overview of the architecture of the ventral visual stream model, VisNet

The architecture of VisNet is summarized briefly next, with a full description provided after this. The fundamental hypotheses for the design of VisNet were described by Rolls

(1992), and the design was implemented by Wallis, Rolls & Földiák (1993), and Wallis & Rolls (1997), with further additions by Rolls & Milward (2000), of Gabor filters for the preprocessing of images by Deco & Rolls (2004), and exploration of different learning rules by Rolls & Stringer (2001). The architecture and properties of VisNet and the investigations performed with it are described by Rolls (2012) and Rolls (2016).

Fundamental elements of Rolls' (1992) theory for how cortical networks might implement invariant object recognition provide the basis for the design of VisNet, which can be summarized as:

- A series of competitive networks, organized in hierarchical layers, exhibiting mutual inhibition over a short range within each layer. These networks allow combinations of features or inputs occurring in a given spatial arrangement to be learned by neurons using competitive learning (Rolls 2016), ensuring that higher order spatial properties of the input stimuli are represented in the network. In VisNet, layer 1 corresponds to V2, layer 2 to V4, layer 3 to posterior inferior temporal visual cortex, and layer 4 to anterior inferior temporal cortex. Layer one is preceded by a simulation of the Gabor-like receptive fields of V1 neurons produced by each image presented to VisNet (Rolls 2012).

- A convergent series of connections from a localized population of neurons in the preceding layer to each neuron of the following layer, thus allowing the receptive field size of neurons to increase through the visual processing areas or layers, as illustrated in Fig. 2.

- A modified associative (Hebb-like) learning rule incorporating a temporal trace of each neuron's previous activity, which, it has been shown (Földiák 1991, Rolls 1992, Wallis et al. 1993, Wallis & Rolls 1997, Rolls & Milward 2000, Rolls & Stringer 2001, Rolls 2012, Rolls 2016), enables the neurons to learn transform invariances.

The learning rates for each of the four layers were 20, 0.4, 0.1, and 0.1, as these rates were shown to produce convergence of the synaptic weights after 10–50 training epochs. 10 training epochs were run.

2.2 The VisNet trace learning rule

The learning rule implemented in the VisNet simulations utilizes the spatio-temporal constraints placed upon the behaviour of 'real-world' objects to learn about natural object transformations. By presenting consistent sequences of transforming objects the cells in the network can learn to respond to the same object through all of its naturally transformed states, as described by Földiák (1991), Rolls (1992), Wallis, Rolls & Földiák (1993), Wallis & Rolls (1997), and Rolls (2012). The learning rule incorporates a decaying trace of previous cell activity and is henceforth referred to simply as the 'trace' learning rule. The learning paradigm we describe here is intended in principle to enable learning of any of the transforms tolerated by inferior temporal cortex neurons, including position, size, view, lighting, and spatial frequency (Rolls 1992, Rolls 2000, Rolls & Deco 2002, Rolls 2016, Rolls 2012).

Various biological bases for this temporal trace have been advanced as follows: The precise mechanisms involved may alter the precise form of the trace rule which should be used. Földiák (1992) describes an alternative trace rule which models individual NMDA channels. Equally, a trace implemented by temporally extended cell firing in a local cortical attractor could implement a short-term memory of previous neuronal firing (Rolls 2016).

- The persistent firing of neurons for as long as 100–400 ms observed after presentations of stimuli for 16 ms (Rolls & Tovee 1994) could provide a time window within which to associate subsequent images. Maintained activity may potentially be implemented by recurrent connections between as well as within cortical areas (Rolls 2016). The prolonged firing of anterior ventral temporal / perirhinal cortex neurons during memory delay periods of several seconds, and associative links reported to develop between stimuli presented several seconds apart (Miyashita 1988) are on too long a time scale to be immediately relevant to the present theory. In fact, associations between visual events occurring several seconds apart would, under *normal* environmental conditions, be detrimental to the operation of a network of the type described here, because they would probably arise from different objects. In contrast, the system described benefits from associations between visual events which occur close in time (typically within 1 s), as they are likely to be from the same object.

- The binding period of glutamate in the NMDA channels, which may last for 100 ms or more (Spruston, Jonas & Sakmann 1995, Hestrin, Sah & Nicoll 1990), may implement a trace rule by producing a relatively long time window over which the *average* activity at each presynaptic site affects learning (Rolls 1992, Rhodes 1992, Földiák 1992).

- In the neocortex in vivo, covariance of pre- and post-synaptic activity over time periods appears to be important in synaptic plasticity (Fregnac, Pananceau, Rene, Huguet, Marre, Levy & Shulz 2010).

The trace update rule used in the baseline simulations of VisNet (Wallis & Rolls 1997) is equivalent to both Földiák's used in the context of translation invariance (Wallis, Rolls & Földiák 1993) and to the earlier rule of Sutton & Barto (1981) explored in the context of modelling the temporal properties of classical conditioning, and can be summarized as follows:

$$\delta w_j = \alpha \overline{y}^\tau x_j \tag{1}$$

where

$$\overline{y}^\tau = (1-\eta) y^\tau + \eta \overline{y}^{\tau-1} \tag{2}$$

and

x_j:	j^{th} input to the neuron.	y:	Output from the neuron.
$\overline{y}^{\tau}$:	Trace value of the output of the neuron at time step τ.	α:	Learning rate.
w_j:	Synaptic weight between j^{th} input and the neuron.	η:	Trace value. The optimal value varies with presentation sequence length.

At the start of a series of investigations of different forms of the trace learning rule, Rolls & Milward (2000) demonstrated that VisNet's performance could be greatly enhanced with a modified Hebbian trace learning rule (equation 3) that incorporated a trace of activity from the preceding time steps, with no contribution from the activity being produced by the stimulus at the current time step. This rule took the form

$$\delta w_j = \alpha \overline{y}^{\tau-1} x_j^{\tau}. \tag{3}$$

The trace shown in equation 3 is in the postsynaptic term. The crucial difference from the earlier rule (see equation 1) was that the trace should be calculated up to only the preceding timestep. This has the effect of updating the weights based on the preceding activity of the neuron, which is likely given the spatio-temporal statistics of the visual world to be from previous transforms of the same object (Rolls & Milward 2000, Rolls & Stringer 2001). This is biologically not at all implausible, as considered in more detail elsewhere (Rolls 2016, Rolls 2012), and this version of the trace rule was used in this investigation.

The optimal value of η in the trace rule is likely to be different for different layers of VisNet. For early layers with small receptive fields, few successive transforms are likely to contain similar information within the receptive field, so the value for η might be low to produce a short trace. In later layers of VisNet, successive transforms may be in the receptive field for longer, and invariance may be developing in earlier layers, so a longer trace may be beneficial. In practice, after exploration we used η values of 0.6 for layer 2, and 0.8 for layers 3 and 4. In addition, it is important to form feature combinations with high spatial precision before invariance learning supported by a temporal trace starts, in order that the feature combinations and not the individual features have invariant representations. For this reason, purely associative learning with no temporal trace was used in layer 1 of VisNet (Rolls & Milward 2000).

The following principled method was introduced to choose the value of the learning rate α for each layer. The mean weight change from all the neurons in that layer for each epoch of training was measured, and was set so that with slow learning over 10–20 trials, the weight changes per epoch would gradually decrease and asymptote with that number of epochs, reflecting convergence. Slow learning rates are useful in competitive nets, for if the learning rates are too high, previous learning in the synaptic weights will be overwritten by large weight changes later within the same epoch produced if a neuron starts to respond to another stimulus (Rolls 2016). If the learning rates are too low, then no useful learning or convergence will occur. It was found that the following learning rates enabled good

operation with the 8 transforms of each of 7 objects used in each epoch in the present investigation: Layer 1 α=0.4; Layer 2 α=0.4 (this is relatively high to allow for the sparse representations in layer 1); Layer 3 α=0.1; Layer 4 α=0.1.

To bound the growth of each neuron's synaptic weight vector, $\mathbf{w}_i$ for the ith neuron, its length is explicitly normalized (a method similarly employed by von der Malsburg (1973) which is commonly used in competitive networks (Rolls 2016)). An alternative, more biologically relevant implementation, using a local weight bounding operation which utilizes a form of heterosynaptic long-term depression (Rolls 2016), has in part been explored using a version of the Oja (1982) rule (see Wallis & Rolls (1997)).

2.3 The network implemented in VisNet

The network itself is designed as a series of hierarchical, convergent, competitive networks, in accordance with the hypotheses advanced above. The actual network consists of a series of four layers, constructed such that the convergence of information from the most disparate parts of the network's input layer can potentially influence firing in a single neuron in the final layer – see Fig. 2. This corresponds to the scheme described by many researchers (Van Essen, Anderson & Felleman 1992, Rolls 1992, Olshausen, Anderson & Van Essen 1993, Rolls 2016, for example) as present in the primate visual system – see Fig. 2. The forward connections to a neuron in one layer are derived from a topologically related and confined region of the preceding layer. The choice of whether a connection between neurons in adjacent layers exists or not is based upon a Gaussian distribution of connection probabilities that roll off radially from the focal point of connections for each neuron. (A minor extra constraint precludes the repeated connection of any pair of cells.) In particular, the forward connections to a neuron in one layer come from a small region of the preceding layer defined by the radius in Table 1 which will contain approximately 67% of the connections from the preceding layer. Table 1 shows the dimensions for the research described here, which utilized 32x32 neurons per layer.

Table 1: VisNet dimensions

	Dimensions	# Connections	Radius
Layer 4	32x32	100	12
Layer 3	32x32	100	9
Layer 2	32x32	100	6
Layer 1	32x32	272	6
Input layer	256x256x16	–	–

Figure 2 shows the general convergent network architecture used. Localization and limitation of connectivity in the network is intended to mimic cortical connectivity, partially because of the clear retention of retinal topology through regions of visual cortex. This architecture also encourages the gradual combination of features from layer to layer which has relevance to the binding problem, as described elsewhere (Rolls 2016, Rolls 2012).

2.4 Competition and lateral inhibition in VisNet

In order to act as a competitive network some form of mutual inhibition is required within each layer, which should help to ensure that all stimuli presented are evenly represented by the neurons in each layer. This is implemented in VisNet by a form of lateral inhibition. The idea behind the lateral inhibition, apart from this being a property of cortical architecture in the brain, was to prevent too many neurons that received inputs from a similar part of the preceding layer responding to the same activity patterns. The purpose of the lateral inhibition was to ensure that different receiving neurons coded for different inputs. This is important in reducing redundancy (Rolls 2016). The lateral inhibition is conceived as operating within a radius that was similar to that of the region within which a neuron received converging inputs from the preceding layer (because activity in one zone of topologically organized processing within a layer should not inhibit processing in another zone in the same layer, concerned perhaps with another part of the image). The lateral inhibition used in this investigation used the parameters for σ shown in Table 3.

The lateral inhibition and contrast enhancement just described are actually implemented in VisNet (Rolls & Milward 2000, Perry, Rolls & Stringer 2010) in two stages, to produce filtering of the type illustrated elsewhere (Rolls 2016, Rolls 2012). The lateral inhibition was implemented by convolving the activation of the neurons in a layer 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

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

The second stage involves contrast enhancement. A sigmoid activation function was used in the way described previously (Rolls & Milward 2000):

$$y = \mathrm{f}^{\mathrm{sigmoid}}(r) = \frac{1}{1 + e^{-2\beta(r-\alpha)}} \tag{5}$$

where r is the activation (or firing rate) of the neuron after the lateral inhibition, y is the firing rate after the contrast enhancement produced by the activation function, and β is the slope or gain and α is the threshold or bias of the activation function. The sigmoid bounds the firing rate between 0 and 1 so global normalization is not required. The slope and threshold are held constant within each layer. The slope is constant throughout training, whereas the threshold is used to control the sparseness of firing rates within each layer. The (population) sparseness of the firing within a layer is defined (Rolls & Treves 1998, Franco, Rolls, Aggelopoulos & Jerez 2007, Rolls 2016, Rolls & Treves 2011) as:

$$a = \frac{(\sum_i y_i/n)^2}{\sum_i y_i^2/n} \tag{6}$$

where n is the number of neurons in the layer. The sparseness was set to the values shown in Table 2 unless otherwise stated, and this parameter was used to set the threshold parameter in the sigmoid activation function.

Table 2: Sigmoid parameters

Layer	1	2	3	4
Sparseness	0.02	0.01	0.01	0.02
Slope β	0.5	4	8	4

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

The sigmoid activation function was used with parameters (selected after a number of optimization runs) as shown in Table 2.

In addition, the lateral inhibition parameters are as shown in Table 3.

2.5 The input to VisNet

VisNet is provided with a set of input filters which can be applied to an image to produce inputs to the network which correspond to those provided by simple cells in visual cortical area 1 (V1). The purpose of this is to enable within VisNet the more complicated response properties of cells between V1 and the inferior temporal cortex (IT) to be investigated, using as inputs natural stimuli such as those that could be applied to the retina of the real visual system. This is to facilitate comparisons between the activity of neurons in VisNet and those in the real visual system, to the same stimuli. In VisNet no attempt is made to train the response properties of simple cells, but instead we start with a defined series of filters to perform fixed feature extraction to a level equivalent to that of simple cells in V1, as have other researchers in the field (Hummel & Biederman 1992, Buhmann, Lange, von der Malsburg, Vorbrüggen & Würtz 1991, Fukushima 1980), because we wish to simulate the more complicated response properties of cells between V1 and the inferior temporal cortex (IT). The elongated orientation-tuned input filters used accord with the general tuning profiles of simple cells in V1 (Hawken & Parker 1987) and were computed by Gabor filters. Each individual filter is tuned to spatial frequency (0.0626 to 0.5 cycles / pixel over four octaves); orientation (0^o to 135^o in steps of 45^o); and sign (± 1). Of the 272 layer 1 connections, the number to each group in VisNet is as shown in Table 4. Any zero D.C. filter can of course produce a negative as well as positive output, which would mean

Table 4: VisNet layer 1 connectivity. The frequency is in cycles per pixel.

Frequency	0.5	0.25	0.125	0.0625
# Connections	201	50	13	8

that this simulation of a simple cell would permit negative as well as positive firing. The response of each filter is zero thresholded and the negative results used to form a separate anti-phase input to the network. The filter outputs are also normalized across scales to compensate for the low frequency bias in the images of natural objects.

The Gabor filters used were similar to those used previously (Deco & Rolls 2004, Rolls 2012, Rolls & Webb 2014, Webb & Rolls 2014). Following Daugman (1988) the receptive fields of the simple cell-like input neurons are modelled by 2D-Gabor functions. The Gabor receptive fields have five degrees of freedom given essentially by the product of an elliptical Gaussian and a complex plane wave. The first two degrees of freedom are the 2D-locations of the receptive field's centre; the third is the size of the receptive field; the fourth is the orientation of the boundaries separating excitatory and inhibitory regions; and the fifth is the symmetry. This fifth degree of freedom is given in the standard Gabor transform by the real and imaginary part, i.e by the phase of the complex function representing it, whereas in a biological context this can be done by combining pairs of neurons with even and odd receptive fields. This design is supported by the experimental work of Pollen & Ronner (1981), who found simple cells in quadrature-phase pairs. Even more, Daugman (1988) proposed that an ensemble of simple cells is best modelled as a family of 2D-Gabor wavelets sampling the frequency domain in a log-polar manner as a function of eccentricity. Experimental neurophysiological evidence constrains the relation between the free parameters that define a 2D-Gabor receptive field (De Valois & De Valois 1988). There are three constraints fixing the relation between the width, height, orientation, and spatial frequency (Lee 1996). The first constraint posits that the aspect ratio of the elliptical Gaussian envelope is 2:1. The second constraint postulates that the plane wave tends to have its propagating direction along the short axis of the elliptical Gaussian. The third constraint assumes that the half-amplitude bandwidth of the frequency response is about 1 to 1.5 octaves along the optimal orientation. Further, we assume that the mean is zero in order to have an admissible wavelet basis (Lee 1996). Cells of layer 1 receive a topologically consistent, localized, random selection of the filter responses in the input layer, under the constraint that each cell samples every filter spatial frequency and receives a constant number of inputs. The mathematical details of the Gabor filtering are described elsewhere (Rolls 2012, Rolls & Webb 2014, Webb & Rolls 2014).

2.6 Measures for network performance

The performance of VisNet was measured by Shannon information-theoretic measures that are identical to those used to quantify the specificity and selectiveness of the representations provided by neurons in the brain (Rolls & Milward 2000, Rolls 2012, Rolls & Treves 2011, Rolls 2016). A single cell information measure indicated how much information was conveyed by the firing rates of a single neuron about the most effective stimulus. A multiple cell information measure indicated how much information about every stimulus was conveyed by the firing rates of small populations of neurons, and was used to ensure that all stimuli had some neurons conveying information about them.

A neuron can be said to have learnt an invariant representation if it discriminates one

set of stimuli from another set, across all transforms. For example, a neuron's response is view invariant if its response to one set of stimuli irrespective of presentation is consistently higher than for all other stimuli irrespective of presentation view. Note that we state 'set of stimuli' since neurons in the inferior temporal cortex are not generally selective for a single stimulus but rather a subpopulation of stimuli (Baylis, Rolls & Leonard 1985, Abbott, Rolls & Tovee 1996, Rolls, Treves & Tovee 1997, Rolls 2007, Franco, Rolls, Aggelopoulos & Jerez 2007, Rolls 2016, Rolls & Treves 2011). The measures of network performance based on information theory and similar to those used in the analysis of the firing of real neurons in the brain (Rolls 2016, Rolls & Treves 2011) are described in detail elsewhere (Rolls & Milward 2000, Rolls & Stringer 2007).

In this paper, after assessing the results with these information theoretic methods, we found it useful to show how the different layers of a network categorised the different objects by computing a correlation matrix between all the objects (with all of their transforms) based on the firing rates of all the neurons in a layer for every transform of every object. It was also useful to show how different single neurons in a layer categorised the stimuli by showing their firing rates to the set of stimuli.

2.7 Coding of Non-Accidental vs Metric properties of Objects in VisNet

2.7.1 Coding of Non-Accidental vs Metric Properties

The hypothesis to be tested is that the VisNet architecture would enable representations to be formed of non-accidental properties of visual stimuli; be relatively insensitive to metric properties; and to generalize across metric properties even without explicit training to generalize across metric properties.

The stimuli shown in Fig. 3 were generated using Blender for this investigation. Objects 1–3 have the non-accidental property of concave edges. Object 4 has the non-accidental property of parallel edges. Objects 4–7 have the non-accidental property of convex edges. The vertical view of each object was 0 deg. Different amounts of tilt of the top towards or away from the viewer are shown at the tilt angles indicated. Each object was thin, and was cut off at the top and bottom to ensure that any view of the top or bottom of the object did not appear, so that the type of curvature of the edges (concave, straight, or convex) was the main cue available. (Each object might be considered as cut out from a piece of cardboard, which is then held vertical in front of the viewer, and then tilted towards or away from the viewer.)

If we take object 1 as an example (Fig. 3), then we can see that it is approximately cylindrical but with concave sides. Its appearance when vertical is at 0 deg (close to the view illustrated of -6 and 6 deg). As the object is tilted towards or away from the viewer, the degree of curvature appears larger. The degree of curvature is thus a metric property (MP) of the object which is not very useful in transform-invariant object recognition. On the other hand, the fact that object 1 has concave sides is apparent across all these transforms, and therefore is a non-accidental property (NAP) used in identifying an object.

This is made evident by a comparison with object 7, which has convex edges, which are always convex through these transforms, while the degree of curvature is a metric property and again varies as a function of the tilt transform. Further, object 4, a cylinder, always has straight sides whatever its tilt, so having straight sides is in this case also a NAP.

VisNet was trained with its trace learning rule on each of these objects. Each object, chosen in random permuted sequence, was presented during training with its views shown successively (in random permuted sequence). The concept is that the different views of that object will be associated together by the short-term memory trace-based synaptic learning rule. Then another object was chosen, and the training for that object was performed. One training epoch consisted of training each object in this way, selecting the objects in random permuted sequence. Ten training epochs were performed.

2.7.2 Generalization to a different object because of the overlap of the metric values of a Non-Accidental Property

Consideration of what is illustrated in Fig. 3 leads in fact to a new theory not present in the literature about how Non-Accidental Properties are learned, and how they are relatively invariant with respect to metric changes. If we look at object 1, we see that a range of curvatures are shown as the object transforms. These different curvatures are all learned as properties of this class of object by the temporal trace learning rule. But if we then look at object 2, we see that some of the degrees of curvature in some of its views are similar to what has already been learned for object 1. So without further training of any association between object 1 and object 2, there will be some generalization across these two objects with different degrees of curvature, because some of the degrees of curvature are similar for some of the views of these two objects. *It is in this way, we propose now, that non-accidental properties are learned, and generalize across a range of objects with the same non-accidental property (e.g. concave sides), even though the degree of curvature, the metric property, may be quite different.* This is tested in the simulations that were performed.

2.8 Intermediate layer neurons of VisNet can respond to feature combinations that are parts of different objects, yet layer 4 neurons are object-selective

The hypothesis to be tested was that VisNet could be trained to identify separately every single feature and feature combinations of an object by its final layer (4); and that then in VisNet's intermediate layers (2 and 3) single neurons might respond to features or low-order feature combinations that were components of several whole objects represented orthogonally in layer 4. The underlying overall hypothesis here is that at any layer of VisNet neurons need only learn low-order feature combinations; and that if this is repeated across several layers, then by the last layer neurons will become object-selective (Rolls 1992, Rolls 2012, Rolls 2016). By coding parts of several objects, this would enable intermediate

layer neurons to contribute to the identification of several or many objects, and this would facilitate a high capacity of VisNet in terms of the number of different objects that could be correctly categorised by layer 4.

This hypothesis was tested by training VisNet on the set of 13 objects shown in Fig. 6. One of the objects is a square, and the other objects are combinations of the adjoining line edges that are parts of the square, or the line elements themselves. If the representations of this set of objects is orthogonal in layer 4 of VisNet, but not in layers 2–3, this indicates that neurons in layers 2 and 3 can be used for more than one object. This was further tested by examining how many objects neurons in layers 2–4 typically responded to, with the prediction that in layer 4 some neurons would respond primarily to one of the 13 objects, and that in layers 2 and 3 neurons would typically respond to several objects.

3 Results

3.1 The representation of non-accidental properties but not metric properties of objects in VisNet

After training for 10 epochs on the set of stimuli illustrated in Fig. 3, some neurons in layer 4 of VisNet became tuned to respond to any transform of objects 1–3 (the objects with concave sides), but to no transform of any other object. Other neurons became tuned to respond best to any transform of object 4, the cylinder with straight sides. Other neurons became tuned to respond mainly to any transform of objects 5–7, the object with convex sides.

The representation that was formed can be appreciated quantitatively by the matrix of correlations between the stimuli formed from the responses for all the neurons in a layer to each stimulus, which is what is shown in Fig. 4. Each object has 8 transforms, so this is a 7-object by 8-transform (56x56) correlation matrix. This shows that all the transforms of objects 1–3 (the objects with concave sides) produced similar responses in layer 4, that were uncorrelated with the responses to any other objects. Similarly, all transforms of all the concave-sided objects (4–7) produced similar responses in layer 4 neurons, with no correlations with the responses to the concave-sided objects (1–3), and only a small correlation with the responses to the straight-sided object, 4. Similarly, object 4 produced responses in layer 4 neurons that were correlated with the transforms of itself, not at all with the transforms of the concave objects (1–3), and were only weakly correlated with the responses to any of the transforms of the convex-sided objects, 1–3.

Fig. 5 shows the firing rates to the different stimuli and tilt views of a single neuron in layer 4 selected to be responsive to object 1. The neuron had responses to all views of all the concave objects 1–3, and to no other object. This shows that some single neurons have learned non-accidental properties of the set of stimuli, and that these single neurons have invariance for the metric properties. This neuron also makes the point that although the population of neurons as a whole in layer 4 that responded to objects 1–3 had a small response to the cylinder object 4 (Fig. 4), some of the neurons distinguished

perfectly between the non-accidental property of concave sides (objects 1–3) vs straight sides (object 4), as shown in Fig. 5.

These results confirm the hypothesis that non-accidental properties arise naturally in VisNet, and support the theory that non-accidental transforms can be learned with metric property invariance because the different views of any one object share metric properties with other objects of the same non-accidental property type when such objects transform across views, as illustrated in Fig. 3.

3.2 Intermediate layer neurons of VisNet can respond to feature combinations that are parts of different objects, yet layer 4 neurons are object-selective

VisNet was trained on the set of 13 objects shown in Fig. 6. The results of the training are shown in Fig. 7. Measuring the correlations between the neuronal representations for each of the stimuli across all the neurons in layer 4 showed orthogonal representations: each object was represented by neurons specialised just for one object. This was confirmed by the finding that single neurons in layer 4 can respond to just one of the objects (Fig. 8(a)).

In layer 3, the neuronal population had somewhat less orthogonal representations (Fig. 7), produced by the fact that some neurons in layer 3 might respond best to one object, but might also have good responses to one or more other objects. This was confirmed by examining the responses of individual neurons in layer 3.

In layers 2 and 1 of Visnet, the neuronal populations had very non-orthogonal representations (Fig. 7), produced by the fact that some neurons in these layers might respond best to one object, but might also have good responses to one or several other objects. This was confirmed by examining the responses of individual neurons in layers 2 and 1, as illustrated in Fig. 8. Fig. 8(a) confirms that a typical single neuron in layer 4 has orthogonal representations of the set of stimuli. This layer 4 neuron, selected to have responses to object 1, the whole square, responded only to that object, and not to any of the components (objects 2–13). This exemplifies the orthogonal object selectivity of single neurons in layer 4. Fig. 8(b) shows that a layer 2 neuron selected to have responses to object 1 also in fact responded to objects 2 and 4 (with the objects illustrated in Fig. 6). This shows that layer 2 neurons can respond to several feature combinations that are part of an object. Fig. 8(c) shows that another layer 2 neuron selected to have responses to object 1 also in fact responded to objects 2, 3 and 7. Comparison of Fig. 8(b) and (c) shows that different neurons in layer 2 can respond to different combinations of features. Fig. 8(d) shows that a layer 2 neuron selected to have responses to object 13 (the edge at the top of the square) also had responses graded differently to objects 8, 9, 3, 4, 5, and 1 (which also contained as a feature an edge at the top). This shows how intermediate layer neurons can respond to features (in this case a top edge) that are a component of several different objects.

This thus demonstrates the important principle in VisNet as a hierarchical unsupervised network approach to computations in the ventral visual system, that representations in intermediate layers can be used for several different objects represented as different objects

in the top layer of the hierarchy (layer 4 in VisNet). The distributed tuning of neurons in intermediate layers enables the capacity, the number of objects that can be represented, to be high, as considered further in the Discussion.

4 Discussion

The investigation on the representation of non-accidental properties by neurons in the ventral visual stream (VisNet) showed not only that they can arise in a relatively simple network modelling many aspects of processing in the ventral visual cortical stream, but also showed how non-accidental properties could arise, and could show insensitivity, that is, in fact, invariance, with respect to metric properties. The mechanism underlying the encoding of non-accidental properties in VisNet, and we propose in the ventral cortical visual stream, is that as an object transforms into different views over short times, slow learning implemented by for example the temporal trace synaptic learning rule in VisNet results in different metric properties such as the degree of curvature described above to be associated together. This invariance learning of metric properties then enables the neurons to generalize to other objects with the same non-accidental property (e.g. concave curvature), but different metric properties (e.g. degree of curvature), because some of the metric properties of the two objects overlap. This is illustrated in Fig. 4, in which there was no associative learning between the concave-sided objects 5–7, yet the individual neurons, and the population of selective neurons as a whole, responded to all of the concave-sided objects 5–7. This investigation thus shows how invariance can be learned for metric properties of objects, and at the same time how non-accidental properties such as concave vs convex edges which are present for many views of an object become what neurons are tuned to.

The hypothesis confirmed in the second investigation was that VisNet could be trained to identify separately every single feature and feature combination of an object by its final layer (4); and that then in VisNet's intermediate layers (2 and 3) single neurons did respond to features or low-order feature combinations that were components of several whole objects represented orthogonally in layer 4. This was shown by the population encodings shown in the different layers of VisNet (Fig. 7), and by the responses of the single neurons in the intermediate layers (Fig. 8). This has been little investigated previously. The underlying overall hypothesis here is that at any layer of VisNet neurons need only learn low-order feature combinations; and that if this is repeated through several layers, then by the last layer neurons will become object selective (Rolls 1992, Rolls 2012, Rolls 2016). By coding parts of several objects, this enables intermediate layer neurons to contribute to the identification of several or many objects, and this facilitates a high capacity of VisNet in terms of the number of different objects that could be correctly categorised by layer 4. This is the new finding reported here, and this finding helps to provide an account of how a network such as VisNet can have a high capacity for the number of objects that it can recognise. In previous investigations with VisNet, the emphasis was usually on measuring the performance in the final layer, rather than on the encoding in the intermediate layers

that is important for what can be produced by the final layer. The new finding was made clear by the use of the simple but well defined stimulus set.

A useful encoding to achieve this high capacity is to build neurons with their best response to a given stimulus, and have the next most effective stimuli for a given neuron selected in a random fashion from the other stimuli. This helps to ensure that individual neurons have relatively uncorrelated responses to the set of stimuli, which is what helps to achieve high capacity. This type of encoding has been found in the inferior temporal visual cortex to objects, in that the information increases approximately linearly with the number of neurons in the sample. This is a signature of independent coding by different neurons, for which the neuronal response profiles to the set of stimuli must be close to uncorrelated (Rolls, Treves & Tovee 1997, Rolls, Franco, Aggelopoulos & Reece 2003, Rolls, Aggelopoulos, Franco & Treves 2004, Franco, Rolls, Aggelopoulos & Treves 2004, Aggelopoulos, Franco & Rolls 2005, Franco et al. 2007, Rolls & Treves 2011, Rolls 2016). What is demonstrated here is that the neurons in the intermediate layers of VisNet have some of these same properties, as shown in Fig. 8, which helps to make VisNet a useful approach and model of ventral stream visual cortical processing.

For comparison, deep learning approaches (LeCun et al. 2015) are very biologically implausible, because they utilise supervised learning, and may utilize hundreds of layers of processing which is completely implausible given the speed of neuronal processing in the brain, which limits hierarchical processing in the brain to 4 or 5 cortical areas (or layers in the terminology used here) of processing (Rolls 2016, Marcus 2018).

In this context, the further developments of VisNet described here, showing how it can account for coding of non-accidental properties and invariance for metric properties, and how VisNet utilizes encoding in intermediate layers that provides the potential for high capacity of the numbers of objects that can be recognised.

References

Abbott, L. F., Rolls, E. T. & Tovee, M. J. (1996). Representational capacity of face coding in monkeys, *Cerebral Cortex* **6**: 498–505.

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

Amir, O., Biederman, I. & Hayworth, K. J. (2011). The neural basis for shape preferences, *Vision Res* **51**(20): 2198–206.

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 Research* **342**: 91–102.

Bengio, Y., Goodfellow, I. J. & Courville, A. (2017). *Deep Learning*, MIT Press, Cambridge, MA.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding, *Psychological Review* **94**(2): 115–147.

Biederman, I. & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence and conditions for 3d viewpoint invariance, *Journal of Experimental Psychology: Human Perception and Performance* **20**(1): 80.

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

Buhmann, J., Lange, J., von der Malsburg, C., Vorbrüggen, J. C. & Würtz, R. P. (1991). Object recognition in the dynamic link architecture: Parallel implementation of a transputer network, *in* B. Kosko (ed.), *Neural Networks for Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, pp. 121–159.

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

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

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

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

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

Franco, L., Rolls, E. T., Aggelopoulos, N. C. & Jerez, J. M. (2007). Neuronal selectivity, population sparseness, and ergodicity in the inferior temporal visual cortex, *Biological Cybernetics* **96**: 547–560.

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

Fregnac, Y., Pananceau, M., Rene, A., Huguet, N., Marre, O., Levy, M. & Shulz, D. E. (2010). A re-examination of Hebbian-covariance rules and spike timing-dependent plasticity in cat visual cortex in vivo, *Frontiers in Synaptic Neuroscience* **2**: 147.

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

Hawken, M. J. & Parker, A. J. (1987). Spatial properties of the monkey striate cortex, *Proceedings of the Royal Society, London B* **231**: 251–288.

Hestrin, S., Sah, P. & Nicoll, R. (1990). Mechanisms generating the time course of dual component excitatory synaptic currents recorded in hippocampal slices, *Neuron* **5**: 247–253.

Hummel, J. E. & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition, *Psychological Review* **99**: 480–517.

Kayaert, G., Biederman, I., Op de Beeck, H. P. & Vogels, R. (2005). Tuning for shape dimensions in macaque inferior temporal cortex, *Eur J Neurosci* **22**(1): 212–24.

Kayaert, G., Biederman, I. & Vogels, R. (2005). Representation of regular and irregular shapes in macaque inferotemporal cortex, *Cerebral Cortex* **15**: 1308–1321.

Kim, J. G. & Biederman, I. (2012). Greater sensitivity to nonaccidental than metric changes in the relations between simple shapes in the lateral occipital cortex, *Neuroimage* **63**(4): 1818–26.

Kubilius, J., Bracci, S. & de Beeck, H. P. O. (2016). Deep neural networks as a computational model for human shape sensitivity, *PLoS computational biology* **12**(4): e1004896.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning, *Nature* **521**: 436–444.

LeCun, Y., Kavukcuoglu, K. & Farabet, C. (2010). Convolutional networks and applications in vision, *2010 IEEE International Symposium on Circuits and Systems* pp. 253–256.

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

Lowe, D. (1985). *Perceptual Organization and Visual Recognition*, Kluwer, Boston.

Malsburg, C. v. d. (1973). Self-organization of orientation-sensitive columns in the striate cortex, *Kybernetik* **14**: 85–100.

Marcus, G. (2018). Deep learning: A critical appraisal, *arXiv preprint arXiv:1801.00631* .

Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex, *Nature* **335**: 817–820.

Oja, E. (1982). A simplified neuron model as a principal component analyzer, *Journal of Mathematical Biology* **15**: 267–273.

Olshausen, B. A., Anderson, C. H. & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information, *Journal of Neuroscience* **13**: 4700–4719.

Parker, S. M. & Serre, T. (2015). Unsupervised invariance learning of transformation sequences in a model of object recognition yields selectivity for non-accidental properties, *Front Comput Neurosci* **9**: 115.

Perry, G., Rolls, E. T. & Stringer, S. M. (2010). Continuous transformation learning of translation invariant representations, *Experimental Brain Research* **204**: 255–270.

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

Rhodes, P. (1992). The open time of the NMDA channel facilitates the self-organisation of invariant object responses in cortex, *Society for Neuroscience Abstracts* **18**: 740.

Robinson, L. & Rolls, E. T. (2015). Invariant visual object recognition: biologically plausibile approaches, *Biological Cybernetics* **109**: 505–535.

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

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

Rolls, E. T. (2007). The representation of information about faces in the temporal and frontal lobes of primates including humans, *Neuropsychologia* **45**: 124–143.

Rolls, E. T. (2012). Invariant visual object and face recognition: neural and computational bases, and a model, VisNet, *Frontiers in Computational Neuroscience* **6**(35): 1–70.

Rolls, E. T. (2016). *Cerebral Cortex: Principles of Operation*, Oxford University Press, Oxford.

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

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

Rolls, E. T., Franco, L., Aggelopoulos, N. C. & Reece, S. (2003). An information theoretic approach to the contributions of the firing rates and the correlations between the firing of neurons, *Journal of Neurophysiology* **89**: 2810–2822.

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

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

Rolls, E. T. & Stringer, S. M. (2007). Invariant global motion recognition in the dorsal visual system: a unifying theory, *Neural Computation* **19**: 139–169.

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

Rolls, E. T. & Tovee, M. J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex, *Journal of Neurophysiology* **73**: 713–726.

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

Rolls, E. T. & Treves, A. (2011). The neuronal encoding of information in the brain, *Progress in Neurobiology* **95**: 448–490.

Rolls, E. T., Treves, A. & Tovee, M. J. (1997). The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex, *Experimental Brain Research* **114**: 149–162.

Rolls, E. T., Treves, A., Tovee, M. & Panzeri, S. (1997). Information in the neuronal representation of individual stimuli in the primate temporal visual cortex, *Journal of Computational Neuroscience* **4**: 309–333.

Rolls, E. T. & Webb, T. J. (2014). Finding and recognising objects in natural scenes: complementary computations in the dorsal and ventral visual systems, *Frontiers in Computational Neuroscience* **8**: 85.

Spruston, N., Jonas, P. & Sakmann, B. (1995). Dendritic glutamate receptor channel in rat hippocampal CA3 and CA1 pyramidal neurons, *Journal of Physiology* **482**: 325–352.

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

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

Vogels, R., Biederman, I., Bar, M. & Lorincz, A. (2001). Inferior temporal neurons show greater sensitivity to nonaccidental than to metric shape differences, *J Cogn Neurosci* **13**(4): 444–53.

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

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

Webb, T. J. & Rolls, E. T. (2014). Deformation-specific and deformation-invariant visual object recognition: pose vs identity recognition of people and deforming objects, *Frontiers in Computational Neuroscience* **8**: 37.

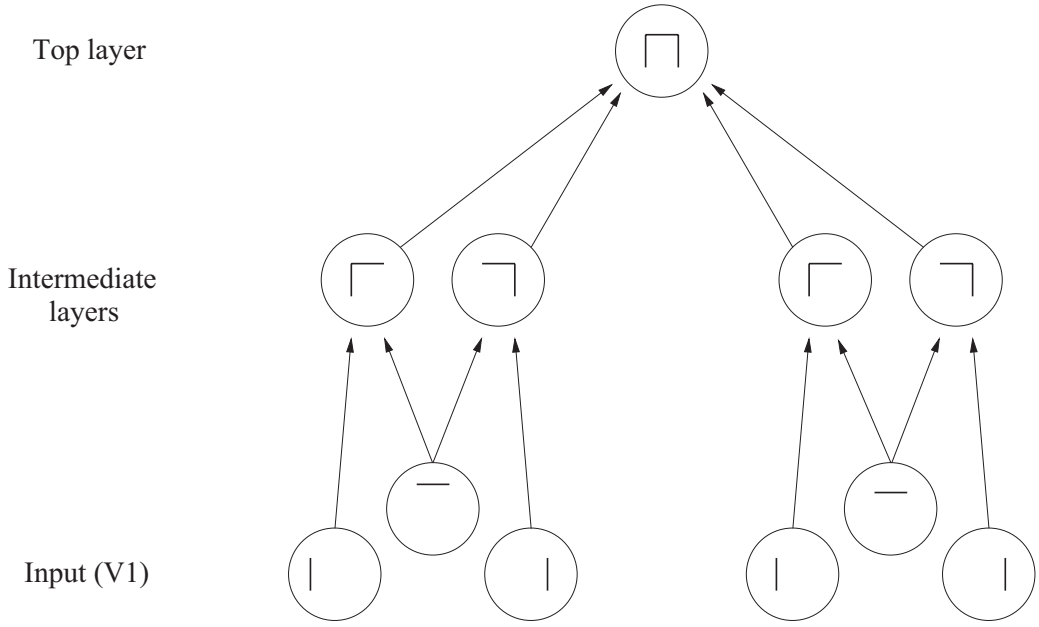

Figure 1: The feature hierarchy approach to object recognition. The inputs may be neurons tuned to oriented straight line segments. In early intermediate levels neurons respond to a combination of these inputs in the correct spatial position with respect to each other. In further intermediate levels, of which there may be several, neurons respond with some invariance to the feature combinations represented early, and form higher order feature combinations. Finally, in the top level, neurons respond to combinations of what is represented in the preceding intermediate level, and thus provide evidence about objects in a position (and scale and even view) invariant way. Convergence through the network is designed to provide top level neurons with information from across the entire input retina, as part of the solution to translation invariance, and other types of invariance are treated similarly.

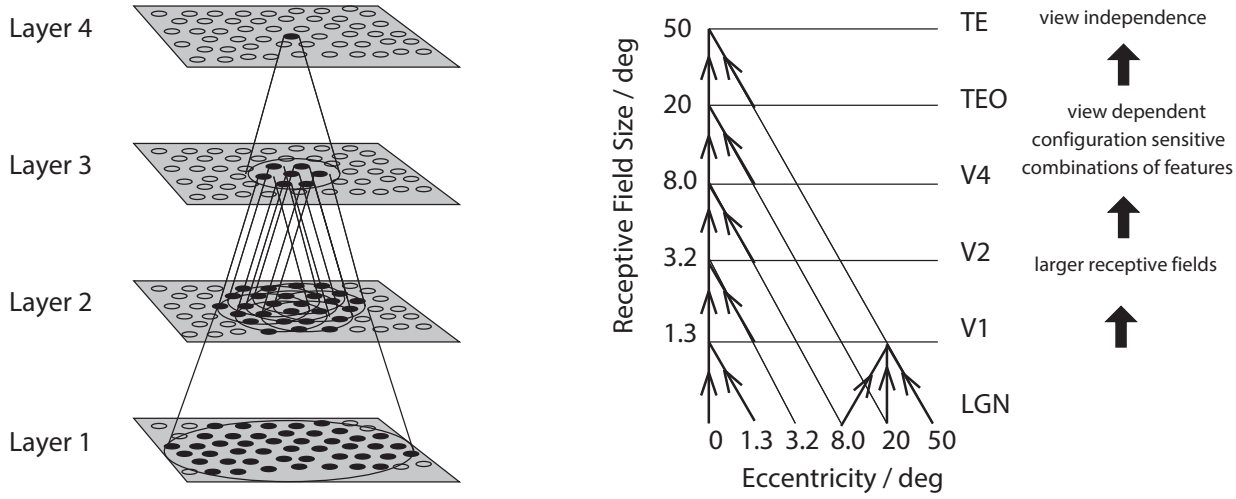

Figure 2: Convergence in the visual system. Right – as it occurs in the brain. V1, visual cortex area V1; TEO, posterior inferior temporal cortex; TE, inferior temporal cortex (IT). Left – as implemented in VisNet. Convergence through the network is designed to provide fourth layer neurons with information from across the entire input retina.

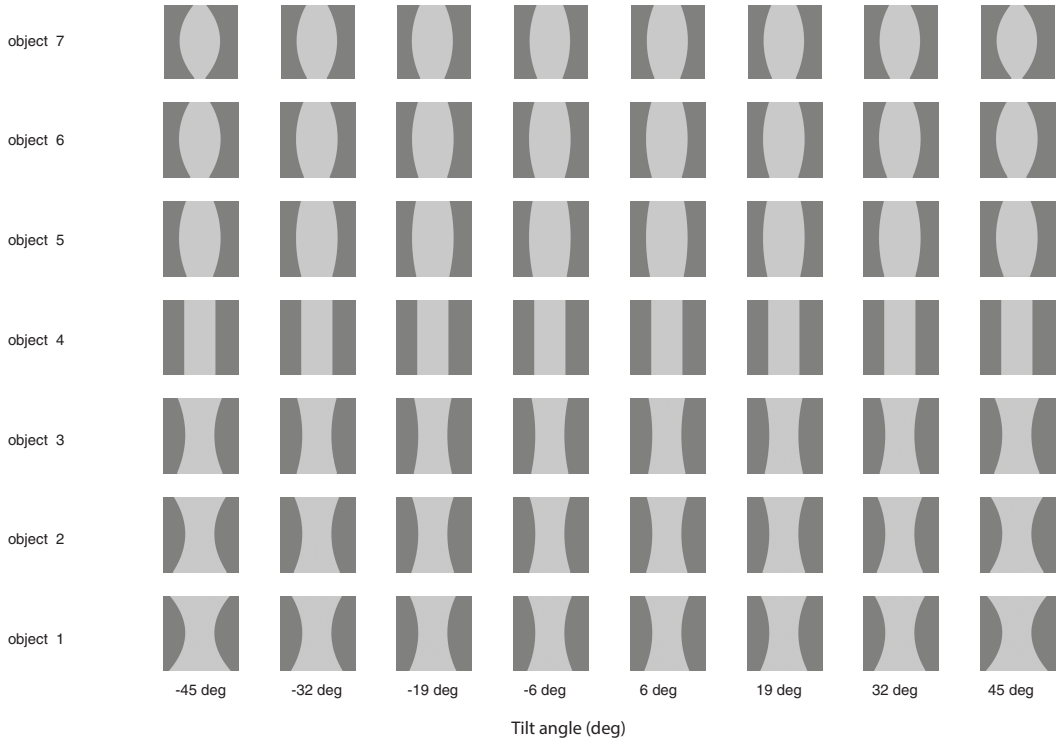

Figure 3: Encoding of non-accidental properties. Stimuli used to investigate non-accidental properties (NAP) vs metric properties (MP). Each object is shown as white on a grey background. Objects 1–3 have the non-accidental property of concave edges. Object 4 has the non-accidental property of parallel edges. Objects 5–7 have the non-accidental property of convex edges. The vertical view of each object was at 0 deg, with the images at -6 and 6 deg of tilt illustrated. Different amounts of tilt of the top towards or away from the viewer are shown at the tilt angles indicated. Each object was thin, and was cut off at the top and bottom to ensure that any view of the top or bottom of the object did not appear, so that the type of curvature of the edges (concave, straight, or convex) was the main cue available.

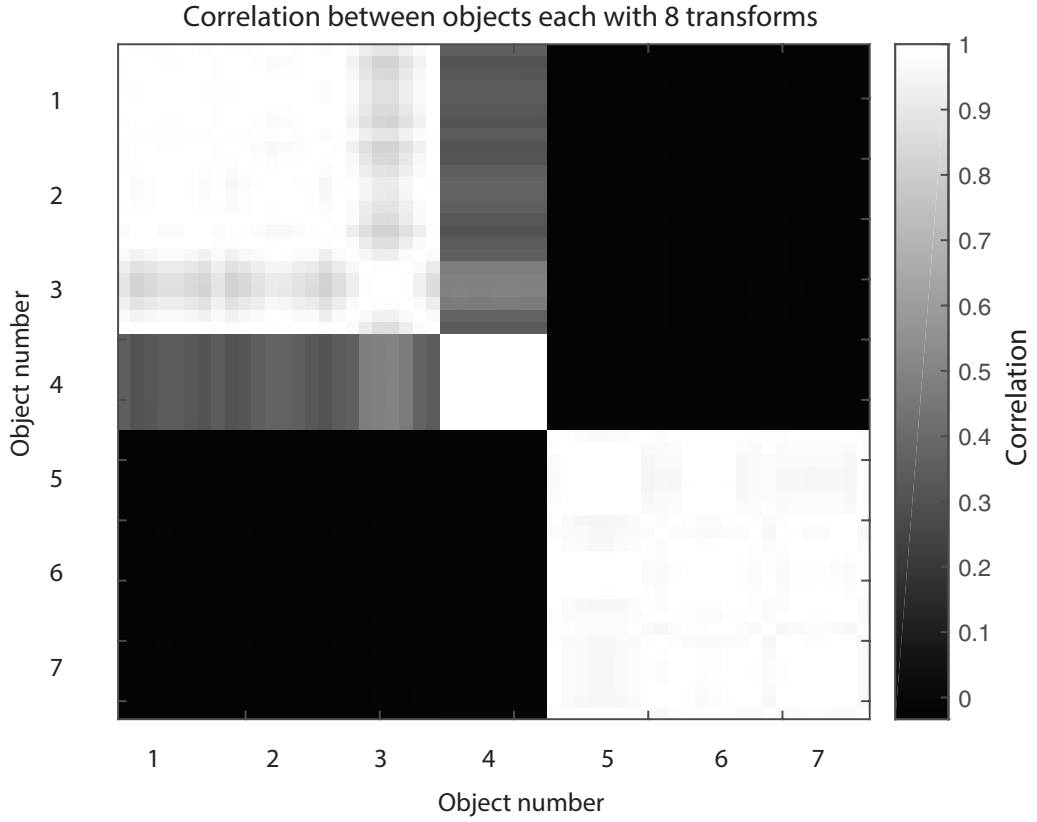

Figure 4: Encoding of non-accidental properties. Correlations between the neuronal representations of the 7 objects used in the non-accidental property (NAP) investigation provided in VisNet layer 4. Each NAP object had 8 transforms, as illustrated in Fig. 3, and so each object in the correlation matrix is represented by each of its 8 transforms (with transform 1 of each object corresponding to a tilt of -40 degrees). Objects 1–3 have the non-accidental property of concave edges. Object 4 has the non-accidental property of parallel edges. Objects 5–7 have the non-accidental property of convex edges.

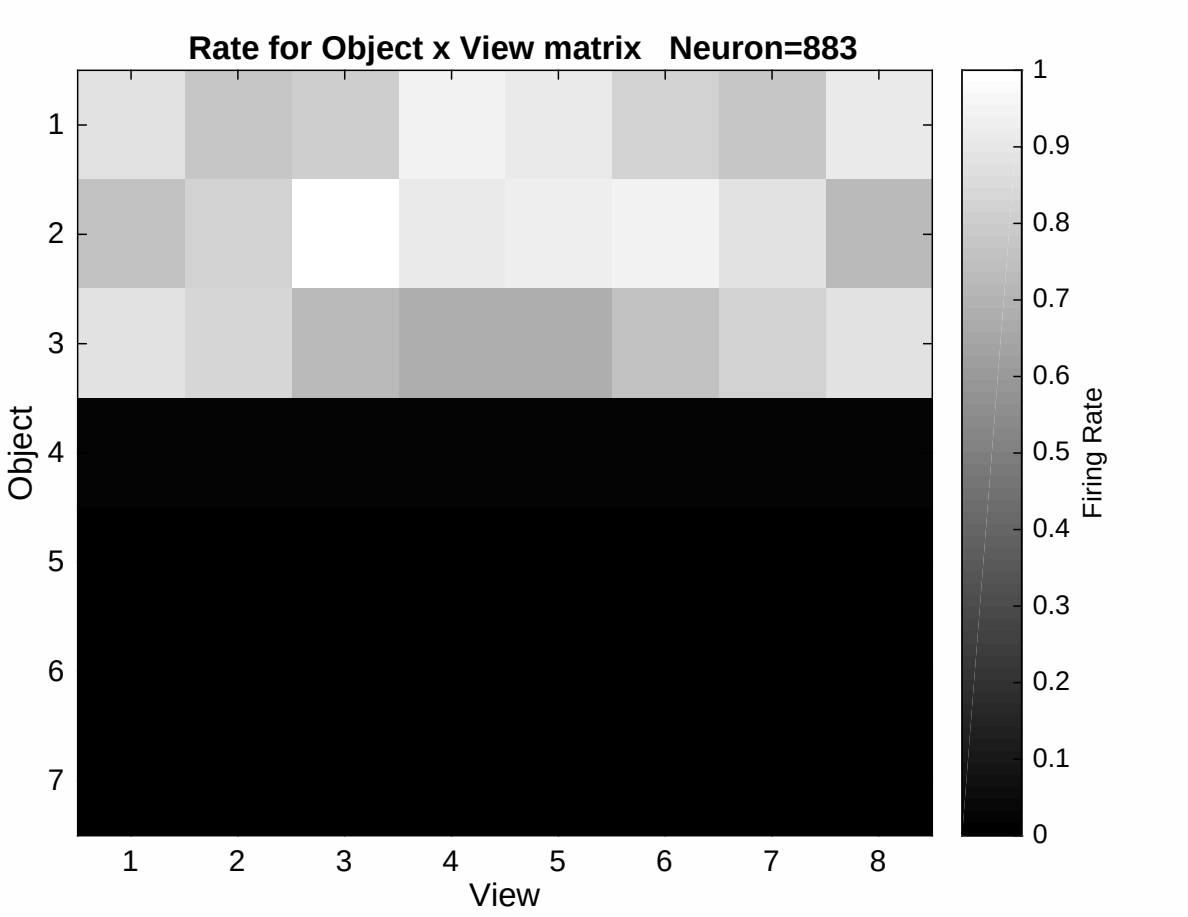

Figure 5: Encoding of non-accidental properties. Responses of a single neuron in layer 4 of VisNet to the set of 7 objects each shown with eight views. View 1 corresponds to a tilt of -40 deg as shown in Fig. 3, and view 8 to + 30 deg. Objects 1–3 have the non-accidental property of concave edges. Object 4 has the non-accidental property of parallel edges. Objects 5–7 have the non-accidental property of convex edges.

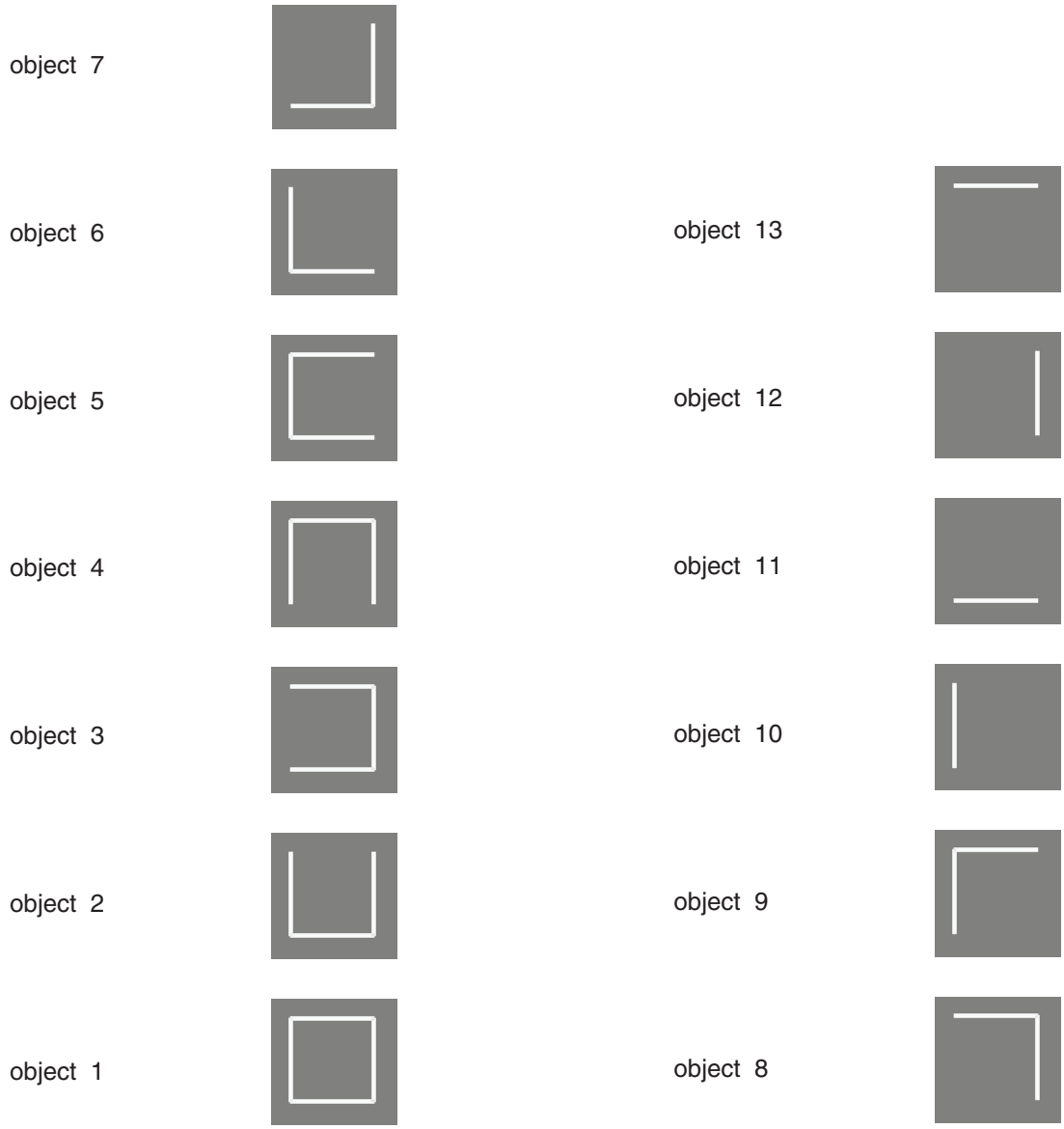

Figure 6: Encoding of information in intermediate layers of VisNet. Stimuli used to investigate coding of feature combinations in intermediate layers of VisNet. Every feature or feature combination with adjacent features was a different object to be learned as a different object by VisNet.

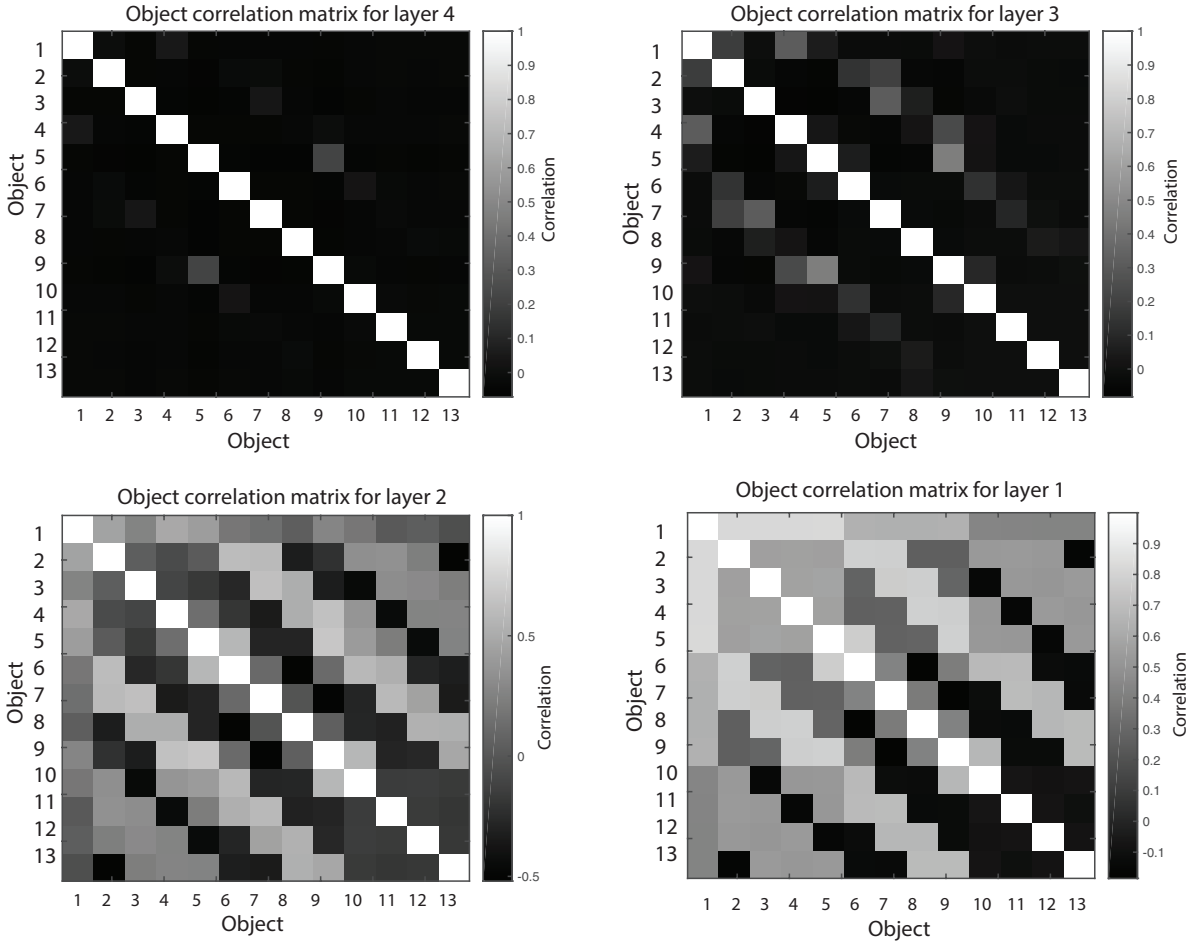

Figure 7: Encoding of information in intermediate layers of VisNet. Correlations between the firing of neurons that represent each of the 13 objects in the Square experiment in different layers of VisNet. Layer 4 is the top layer, and layer 1 is the layer that receives from V1.

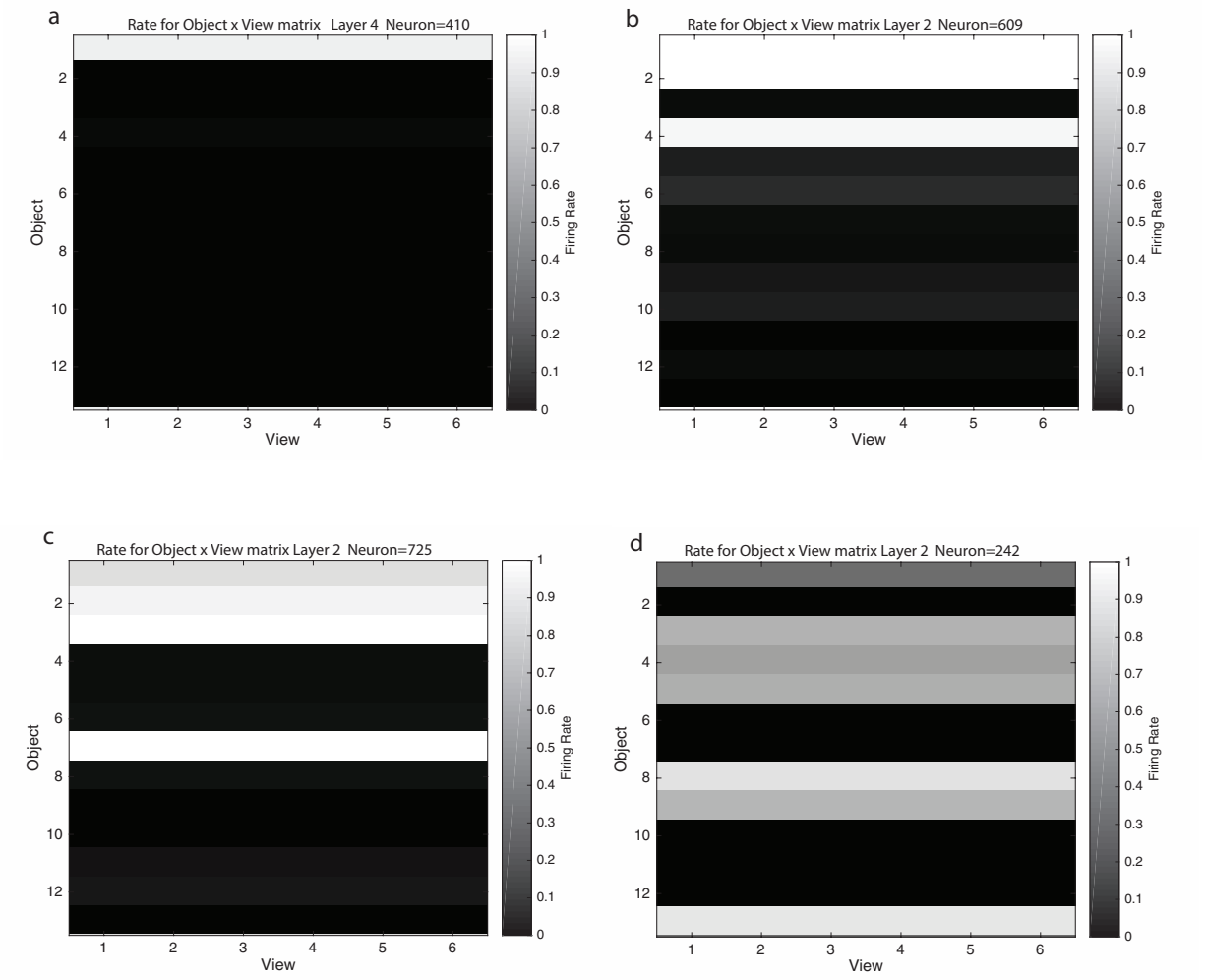

Figure 8: Encoding of information in intermediate layers of VisNet. Examples of the tuning of single neurons in different layers to the set of stimuli. (a). A neuron in layer 4 selected to have responses to object 1, the whole square, responded only to that object, and not to any of the components (objects 2–13). This exemplifies the orthogonal object selectivity of single neurons in layer 4. (b). A layer 2 neuron selected to have responses to object 1 also in fact responded to objects 2 and 4 (with the objects illustrated in Fig. 6). (c). Another layer 2 neuron selected to have responses to object 1 also in fact responded to objects 2, 3 and 7. (d) A layer 2 neuron selected to have responses to object 13 (the edge at the top of the square) also had responses graded differently to objects 8, 9, 3, 4, 5, and 1 (which also contained as a feature an edge at the top). To make the relevant points clear, for this simulation the views for each object were the same, as this allows clear interpretation of the combinatorial encoding investigated in this experiment. Layer 4 is the top layer, and layer 1 is the layer that receives from V1.