Biol. Cybern. 86, 59–71 (2002)

Biological Cybernetics

Invariant recognition of feature combinations in the visual system

M. C. M. Elliffe, E. T. Rolls, S. M. Stringer

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

Received: 4 August 1999 / Accepted in revised form: 11 October 2000

Abstract. The operation of a hierarchical competitive network model (VisNet) of invariance learning in the visual system is investigated to determine how this class of architecture can solve problems that require the spatial binding of features. First, we show that VisNet neurons can be trained to provide transform-invariant discriminative responses to stimuli which are composed of the same basic alphabet of features, where no single stimulus contains a unique feature not shared by any other stimulus. The investigation shows that the network can discriminate stimuli consisting of sets of features which are subsets or supersets of each other. Second, a key feature-binding issue we address is how invariant representations of low-order combinations of features in the early layers of the visual system are able to uniquely specify the correct spatial arrangement of features in the overall stimulus and ensure correct stimulus identification in the output layer. We show that output layer neurons can learn new stimuli if the lower layers are trained solely through exposure to simpler feature combinations from which the new stimuli are composed. Moreover, we show that after training on the low-order feature combinations which are common to many objects, this architecture can – after training with a whole stimulus in some locations – generalise correctly to the same stimulus when it is shown in a new location. We conclude that this type of hierarchical model can solve feature-binding problems to produce correct invariant identification of whole stimuli.

1 Introduction

1.1 Background

This paper explores how the visual system may discriminate between stimuli that are composed of combinations of shared features through the formation of transform-invariant neurons. There is now considerable evidence to support the hypothesis that over successive stages the visual system develops neurons that can respond with view, size and position invariance to objects or faces (Desimone 1991; Tanaka et al. 1991; Rolls 1992, 2000; Rolls and Tovee 1995). Rolls (1992, 1994, 1995) proposed a biologically plausible mechanism for transform-invariant object recognition based on the following: (1) a series of hierarchical competitive networks with local graded inhibition; (2) convergent connections to each neuron from a topologically corresponding region of the preceding layer, leading to an increase in the receptive field size of cells through the visual processing areas; and (3) synaptic plasticity based on a modified Hebb-like learning rule with a temporal trace of each cell's previous activity. Wallis and Rolls (1997) implemented a four-layer neural network model (VisNet) to demonstrate that such an architecture can indeed produce view-invariant neurons that respond to some but not other stimuli.

In this paper we investigate two key issues that arise in such hierarchical layered network architectures, other examples of which have been described and analysed by Fukushima (1980), Ackley et al. (1985) and Rosenblatt (1961). One issue is whether the network can discriminate between stimuli that are composed of the same basic alphabet of features. The second issue is whether such network architectures can find solutions to the spatial-binding problem. These issues are described in the next two paragraphs.

The first issue investigated is whether the VisNet type of hierarchical layered network architecture can discriminate stimuli that are composed of a limited set of features, and where the different stimuli include cases where the feature sets are subsets and supersets of those in the other stimuli. In previous investigations with VisNet, we used stimuli (such as faces, or shapes such as T, L and +) where each stimulus might contain unique features not present in the other stimuli. In Sect. 2.1, the stimuli are derived from a set of four features which are designed so that each feature is spatially separate from the other features, and no unique combination of firing

Correspondence to: E. T. Rolls
(E-mail: edmund.rolls@psy.ox.ac.uk, Web: www.cns.ox.ac.uk
Tel.: +44-1865-271348, Fax: +44-1865-310447)

caused, for example, by overlap of horizontal and vertical filter outputs in the input representation distinguishes any one stimulus from the others. The results described in Sect. 2.1 show that VisNet can indeed learn correct invariant representations of stimuli which do consist of feature sets where individual features do not overlap spatially with each other, and where the stimuli can be composed of sets of features which are supersets or subsets of those in other stimuli. Fukushima and Miyake (1982) did not address this crucial issue where different stimuli might be composed of subsets or supersets of the same set of features, although they did show that stimuli with partly overlapping features could be discriminated by the so-called neocognitron.

In Sect. 2.2 we go on to address the spatial-binding problem in architectures such as VisNet. This computational problem, which needs to be addressed in hierarchical networks such as the primate visual system and VisNet, is how representations of features can be (e.g. translation) invariant, yet can specify stimuli or objects in which the features must be specified in the correct spatial arrangement. This is the feature-binding problem, discussed for example by von der Malsburg (1990), and arising in the context of hierarchical layered systems (Rosenblatt 1961; Fukushima 1980; Ackley et al. 1985). The issue is whether or not features are bound into the correct combinations, or if alternative combinations of known features would elicit the same responses. Von der Malsburg suggested that one potential solution is the addition of a temporal dimension to the neuronal response, so that features that should be bound together would be linked by temporal binding. There has been considerable neurophysiological investigation of this possibility (Singer et al. 1990; Abeles 1991; Hummel and Biederman 1992; Singer and Gray 1995). We note that a problem with this approach is that temporal binding might enable, say, features 1, 2 and 3, which might define one stimulus, to be bound together and kept separate from, for example, another stimulus consisting of features 2, 3 and 4, but would require a further temporal binding (leading in the end potentially to a combinatorial explosion) to indicate the relative spatial positions of the 1, 2 and 3 in the 123 stimulus, so that it can be discriminated from, for example, 312. Another approach to a binding mechanism is to group spatial features based on local mechanisms that might operate for closely adjacent synapses on a dendrite (Finkel and Edelman 1987; Mel et al. 1998). A problem for such architectures is how to force one particular neuron to respond to the same feature combination invariantly with respect to all the ways in which that feature combination might occur in a scene.

The approach to the spatial-binding problem that is proposed for VisNet is that individual neurons at an early stage of processing are set up (by learning) to respond to low-order combinations of input features occurring in a given relative spatial arrangement and position on the retina (Rolls 1992, 1994, 1995; Wallis and Rolls 1997, Rolls and Treves 1998; cf. Feldman 1985). Then invariant representations are developed in the next layer from these feature combination neurons which already contain evidence on the local spatial arrangement of features. Finally, in later layers, only one stimulus would be specified by the particular set of low-order feature combination neurons present, even though each feature combination neuron would itself be somewhat invariant.

1.2 An overview of the VisNet model

The simulations in this paper were performed using the latest version of the VisNet model (VisNet2) which is described more fully by Wallis and Rolls (1997) and Rolls and Milward (2000). The 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 Fig. 1. 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. Typical values are given in Table 1.

Before stimuli are presented to VisNet's first layer, they are pre-processed by a set of input filters which accord with the general tuning profiles of simple cells in V1 (Hawken and Parker 1987). The input filters used are computed by weighting the difference of two Gaussians

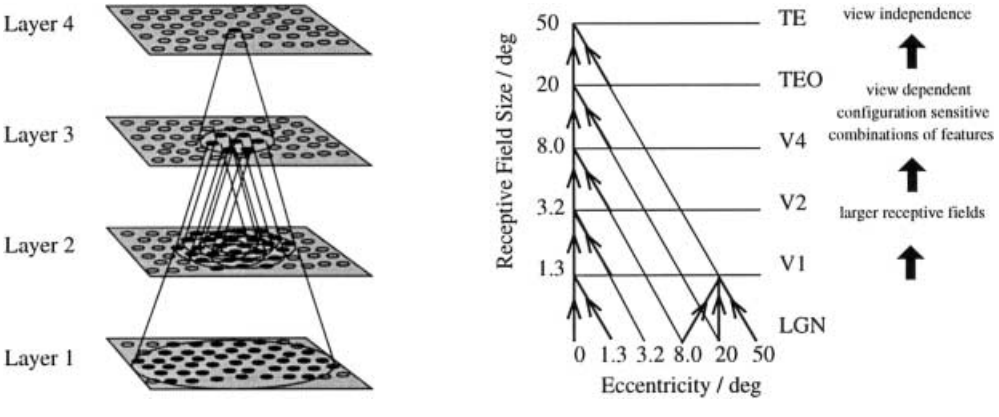

Fig. 1. *Left*: Stylised image of the VisNet four-layer network. Convergence through the network is designed to provide fourth-layer neurons with information from across the entire input retina. In this diagram, the first layer of VisNet corresponds to V1 of the primate visual system shown on the *right*. *Right*: Convergence in the visual system (adapted from Rolls 1992). *V1*, primary (striate) visual cortex area; *TEO*, posterior inferior temporal cortex; *TE*, inferior temporal cortex

Table 1. VisNet dimensions

	Dimensions	No. 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
Input Layer	128 × 128 × 32	–	–

Table 2. VisNet layer-1 connectivity. The frequency is in cycles per pixel

Frequency	0.0625	0.125	0.25	0.5
No. of connections	8	13	50	201

by a third orthogonal Gaussian according to the following (see Wallis and Rolls 1997):

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

where f is the filter spatial frequency, θ is the filter orientation and ρ is the sign of the filter, i.e. ± 1. Individual filters are tuned to spatial frequency (0.0625–0.5 cycles/pixel over 4 octaves), orientation (0–135° in steps of 45°) and sign (± 1). Only even-symmetric (bar) filters were used. The filter outputs were thresholded at zero, and the negative results used to form separate antiphase inputs by other neurons in the network. (This is to allow for the fact that neurons cannot have negative firing rates). The filter outputs also are normalised across scales to compensate for the low-frequency bias in the images of natural objects. The number of layer 1 connections to each spatial frequency filter group is given in Table 2.

Within each layer competition is graded rather than winner-take-all, and is implemented in two stages. First, to implement lateral inhibition the activations 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:

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

Typical lateral inhibition parameters are given in Table 3.

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

$$y = f^{\text{sigmoid}}(r) = \frac{1}{1+e^{-2\beta(r-\alpha)}} \quad (3)$$

where r is the activation (or firing rate) after lateral inhibition, y is the firing rate after contrast enhancement, and α and β are the sigmoid threshold and slope, respectively. The parameters α and β are constant within each layer, although α is adjusted to control the sparseness of the firing rates. For example, to set the sparseness to, say, 5%, the threshold is set to the value of the 95th percentile point of the activations within the layer. Typical parameters for the sigmoid activation function are shown in Table 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

Table 4. Sigmoid parameters

Layer	1	2	3	4
Percentile	99.2	98	88	91
Slope, β	190	40	75	26

The 'trace' learning rules implemented in this paper rely on the natural statistics of real-world objects (i.e. that the position of an object with respect to the eye is likely to change more rapidly than the identity of the object viewed), to be able to learn about their transformations. The premise set out by Földiák (1991) and Rolls (1992) is that individual neurons may learn to respond to different transformations of an object by learning to respond to 'temporal classes' of the views that would tend to occur close together in time. The idea is that because objects have continuous properties in space and time in the world, an object at one place on the retina might activate feature analysers at the next stage of cortical processing, and when the object is translated to a nearby position – because this would occur in a short period (e.g. 0.5 s) – the membrane of the postsynaptic neuron would still be in its associatively modifiable state, and the presynaptic afferents activated with the object in its new position would thus become strengthened on the still-activated postsynaptic neuron. The neuronal mechanisms that might implement this short-term temporal averaging in the modifiability are of interest, and include lasting effects of calcium entry into the postsynaptic neuron as a result of the voltage-dependent activation of NMDA receptors; and continuing firing of the postsynaptic neuron implemented by recurrent collateral connections forming a short-term memory (see Rolls 1992, 2000; Wallis and Rolls 1997). The original trace learning rule used by Wallis and Rolls (1997) took the form

$$\Delta w_j = \alpha \bar{y}^\tau x_j^\tau \quad (4)$$

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

$$\bar{y}^\tau = (1-\eta)y^\tau + \eta \bar{y}^{\tau-1} \quad (5)$$

where x_j is the jth input to the neuron, y is the output from the neuron, $\bar{y}^\tau$ is the trace value of the output of the neuron at time step τ, α is the learning rate (annealed between unity and zero), w_j is the synaptic weight

between jth input and the neuron, and η is the trace value (the optimal value varies with presentation sequence length). The parameter η may be set in the interval $[0, 1]$, and in our simulations with trace learning η is in fact set to 0.8. However, for $\eta = 0$, (4) becomes the standard Hebb rule

$$\Delta w_j = \alpha y^{\tau} x_j^{\tau} \ . \tag{6}$$

However, further learning rules for view-invariant object recognition with enhanced performance are presented by Rolls and Milward (2000) and Rolls and Stringer (2001). First, Rolls and Milward (2000) show that performance is improved by incorporating a trace of activity from only the preceding time step. A basic example of such a rule is

$$\Delta w_j = \alpha \bar{y}^{\tau-1} x_j^{\tau} \ . \tag{7}$$

One way to understand the operation of this version of a trace rule is to note that it is trying to set up the synaptic weight at time τ based on whether the neuron, based on its previous history, is responding to that stimulus (in other positions). Use of the trace accumulated up to $\tau - 1$ as in (7) does this, that is it takes into account the firing of the neuron on previous trials, with no contribution from the firing being produced by the stimulus on the current trial. On the other hand, use of the trace at time τ in the update takes into account the current firing of the neuron to the stimulus in that particular position, which is not a good estimate of whether that neuron should be allocated to invariantly represent that stimulus. Effectively, using the trace at time τ introduces a Hebbian element into the update, which tends to build position-encoded analysers, rather than stimulus-encoded analysers. (The argument has been phrased for a system learning translation invariance, but applies to the learning of all types of invariance.) In the VisNet simulations discussed later in this paper, the learning rule in (7) is used to develop transform invariant neurons, but the results are generic, and similar invariance learning is obtained with rules of the form shown in the learning rule in (4). For a further description of a number of different trace learning rules, the reader is referred to Rolls and Milward (2000) and Rolls and Stringer (2001).

VisNet is compared with other models for achieving invariant object recognition by Wallis and Rolls (1997) and Parga and Rolls (1998) (see also Stone 1996; Bartlett and Sejnowski 1997; Salinas and Abbott 1997).

1.3 Training and test procedure

The stimuli used for training and testing VisNet in this paper are specially constructed to investigate the performance of VisNet on the feature-binding problems described in Sect. 1.1, and are described in Sect. 2. To train the network, a stimulus is presented in a randomised sequence of nine locations in a square grid across the 128×128 input retina. The central location of the square grid is in the centre of the 'retina', and the eight other locations are offset 8 pixels horizontally and/or vertically from this. At each presentation the activation of individual neurons is calculated, then their firing rates are calculated, and then the synaptic weights are updated. After a stimulus has been presented in all the training locations, a new stimulus is chosen at random and the process repeated. The presentation of all the stimuli across all locations constitutes one epoch of training. In this manner the network is trained one layer at a time starting with layer one and finishing with layer 4. In the investigations described here, the numbers of training epochs for layers 1–4 were 50, 100, 100 and 75 respectively.

The network's performance is assessed using two information-theoretic measures: single and multiple cell information about which stimulus was shown. Full details on the application of these measures to VisNet are given by Rolls and Milward (2000). These measures reflect the extent to which cells respond invariantly to a stimulus over a number of retinal locations, but respond differently to different stimuli. 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 which stimulus was shown. In general, the more information that is obtained about a stimulus, the better the invariant representation. The information a single cell conveys about a stimulus s from the set S is computed using the formula below, with details available in Rolls et al. (1997a) and Rolls and Milward (2000). The stimulus-specific information (on surprise) $I(s, R)$ is the amount of information the set of responses R of a single cell has about a specific stimulus s, and is given by

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

The set of responses R consisted of the firing rate y of a cell to every stimulus presented in every location. The calculation procedure was identical to that described by Rolls et al. (1997a) with the following exceptions. First, no correction was made for the limited number of trials, because in VisNet2 (as in VisNet), each measurement of a response is exact, with no variation due to sampling on different trials. Second, the binning procedure was altered in such a way that the firing rates were binned into equispaced rather than equipopulated bins. This small modification was useful because the data provided by VisNet2 can produce perfectly discriminating responses with little trial-to-trial variability. Because the cells in VisNet2 can have bimodally distributed responses, equipopulated bins could fail to perfectly separate the two modes. (This is because one of the equipopulated bins might contain responses from both of the modes.) The number of bins used was equal to or less than the number of trials per stimulus, which for VisNet is the number of positions used on the retina (Rolls et al. 1997a).

Because VisNet operates as a form of competitive net to perform categorization of the inputs received, good performance of a neuron will be characterized by large

responses to one or a few stimuli regardless of their position on the retina (or other transform), and small responses to the other stimuli. We are thus interested in the maximum amount of information that a neuron provides about any of the stimuli, rather than the average amount of information it conveys about the whole set S of stimuli (known as the mutual information). Thus for each cell the performance measure was the maximum amount of information a cell conveyed about any one stimulus (with a check – which in practice was always satisfied – that the cell had a large response to that stimulus, as a large response is what a correctly operating competitive net should produce to an identified category). Some of the graphs in this paper show the amount of information that each of a number of the most informative cells had about any stimulus.

If all the output cells of VisNet learned to respond to the same stimulus, then the information about the set of stimuli S would be very poor, and would not reach its maximal value of $\log_2$ of the number of stimuli (in bits). A measure that is useful here is the information provided by a set of cells about the stimulus set. If the cells provide different information because they have become tuned to different stimuli or subsets of stimuli, then the amount of this multiple-cell information should increase with the number of different cells used, up to the total amount of information needed to specify which of the N_S stimuli have been shown, i.e. $\log_2 N_S$ bits. Procedures for calculating the multiple-cell information have been developed for multiple neuron data by Rolls et al. (1997b) (see also Rolls and Treves 1998), and the same procedures were used for the responses of VisNet. In brief, what was calculated was the mutual information $I(S, \mathbf{R})$, that is, the average amount of information that is obtained from a single presentation of a stimulus from the responses of all the cells. For multiple cell analysis, the set of responses, $\mathbf{R}$, consists of response vectors composed of the responses from each cell. Ideally, we would like to calculate

$$I(S, \mathbf{R}) = \sum_{s \in S} P(s) I(s, \mathbf{R}) \tag{9}$$

However, the information cannot be measured directly from the probability table $P(\mathbf{r}, s)$ embodying the relationship between a stimulus s and the response rate vector $\mathbf{r}$ provided by the firing of the set of neurons to a presentation of that stimulus. (Note, as is made clear at the start of this paper, 'stimulus' refers to an individual object that can occur with different transforms (e.g. as translation here), but elsewhere view and size transforms; see Wallis and Rolls (1997).) This is because the dimensionality of the response vectors is too large to be adequately sampled by trials. Therefore, a decoding procedure is used, in which the stimulus s' that gave rise to the particular firing rate response vector on each trial is estimated. This involves, for example, maximum likelihood estimation or dot-product decoding. (For example, given a response vector $\mathbf{r}$ to a single presentation of a stimulus, its similarity to the average response vector of each neuron to each stimulus is used to estimate – using a dot-product comparison – which stimulus was shown. The probabilities of it being each of the stimuli can be estimated in this way. Details are provided by Rolls et al. (1997b) and by Panzeri et al. (1999).) A probability table is then constructed of the real stimuli s and the decoded stimuli s'. From this probability table, the mutual information between the set of actual stimuli S and the decoded estimates S' is calculated as

$$I(S, S') = \sum_{s,s'} P(s, s') \log_2 \frac{P(s, s')}{P(s) P(s')} \quad . \tag{10}$$

This was calculated for the subset of cells which had as single cells the most information about which stimulus was shown. Often five cells for each stimulus with high information values for that stimulus were used for this.

2 VisNet simulations

2.1 Discrimination between stimuli with super- and subset feature combinations

Previous investigations with VisNet (Wallis and Rolls 1997) have involved groups of stimuli that might be identified by some unique feature common to all transformations of a particular stimulus. This would allow VisNet to solve the problem of transform invariance by simply learning to respond to the unique feature present in each stimulus. For example, even in the case where VisNet was trained on invariant discrimination of T, L and +, the representation of the T stimulus at the spatial-filter-level inputs to VisNet might contain unique patterns of filter outputs where the horizontal and vertical parts of the T join. The unique filter outputs thus formed might distinguish the T from, for example, the L. In this section we test whether VisNet is able to form transform-invariant cells with stimuli that are specially composed from a common alphabet of features, with no stimulus containing any firing in the spatial filter inputs to VisNet not present in at least one of the other stimuli. The limited alphabet enables the set of stimuli to consist of feature sets which are subsets or supersets of those in the other stimuli. In this section we examine the performance of VisNet on a set of such stimuli. Such a task is easily solved by the human visual system, but it remains to be established whether or not it can be solved by VisNet.

For these experiments the common pool of stimulus features chosen was a set of two horizontal and two vertical 8×1 bars, each aligned with the sides of a 32×32 square. The stimuli can be constructed by arbitrary combinations of these base features. We note that effectively the stimulus set consists of four features, a top bar (T), a bottom bar (B), a left bar (L) and a right bar (R). Figure 2 shows the complete set used, containing every possible image feature combination. (Note that, in the interests of retaining symmetry and equal interobject overlap within each feature-combination level, the two double-feature combinations where the features are parallel to each other are not included.)

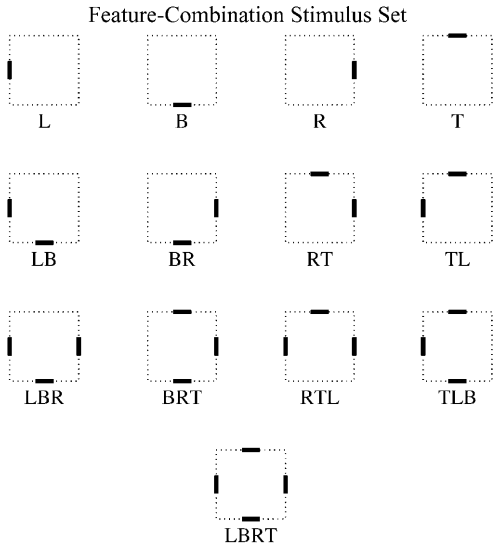

Fig. 2. Merged feature objects. All members of the full object set are shown, using a *dotted line* to represent the central 32 × 32 square on which the individual features are positioned, with the features themselves shown as *dark line segments*. Nomenclature is by acronym of the features present: *T*, top bar; *B*, bottom bar; *L*, left bar; *R*, right bar

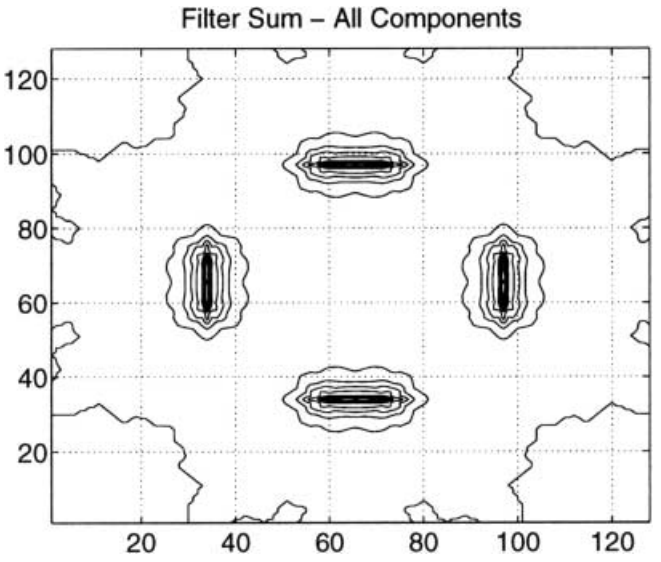

Fig. 3. Merged image feature components. The graph is a contour diagram of the simple sum of all pre-processed image features. The *outer* (lowest) *contour* is the lowest non-zero value, with the remaining nine contours linearly spaced to the maximum

Subsequent discussion will group these objects by the number of features each contain: single-, double-, triple- and quadruple-feature objects correspond to the respective rows of Fig. 2. Stimuli are referred to by the list of features they contain; e.g. 'LBR' contains the left, bottom and right features, while 'TL' contains top and left features only.

A potential pitfall in the construction of the stimuli is that all images presented to VisNet are pre-processed by a filtering operation described in Wallis and Rolls (1997), which produces firing in a set of cells that is intended to emulate some of the processing performed by V1. Unfortunately, this can introduce uniquely identifying pixels for individual feature combinations due to interaction between individual feature elements during the filtering process. Such a unique pixel present in one stimulus but not in any other stimulus could provide a way for a network to operate as a look-up table, in which a particular stimulus could be identified by the presence of a particular pixel. We wished to eliminate this possibility for this particular investigation, as the hypotheses we are investigating are the extent to which networks such as (the neural network part of) VisNet can learn about stimuli composed of subsets and supersets of a set of features. An approach to stimulus preparation which avoids the problem is construction from pre-processed component features. (Of course, interactions between features will occur at later stages, and indeed are part of the way in which such networks operate, but the aim was to prevent the network from using any unique identifying pixel or pixels in the stimuli themselves.) That is, base stimuli are first constructed consisting of a single image feature only, and are then pre-processed in the normal way. More complex stimuli are then constructed by merging these pre-processed representations, with the maximum taken wherever component pixel values differ. This prevents formation of some high pixel values that might indicate, regardless of the transform, that a particular feature combination was present. Other steps to minimise feature interaction were to place individual features far apart, and in some simulations even to remove the firing of the lowest-spatial-frequency filters to ensure no overlap of any of the activity produced by features in different spatial positions. Figure 3 shows the sum of all filter outputs representing a single stimulus built from merging all pre-processed image features. (The ten contours on the diagram as shown begin with the lowest non-zero value, and are then linearly spaced to the maximum.) We note that when VisNet operates normally without this special stimulus preparation, then some interaction between features in the visual scene may be introduced by the pre-processing filtering stage of VisNet, and that such interactions are also known to occur in primate V1.

In the following experiments, the training procedure was carried out as described in Sect. 1.3 with all of the individual stimuli presented in sequences of nine locations across the input. As described in that section, to train the network a stimulus is presented in a randomised sequence of nine locations in a square grid across the 128 × 128 input retina. The central location of the square grid is in the centre of the retina, and the eight other locations are offset 8 pixels horizontally and/or vertically from this. In these experiments we tested performance using two different learning rules: 'Hebbian' (6) and 'trace' (7), and also an untrained condition with random weights. As in earlier work (Wallis and Rolls 1997; Rolls and Milward 2000), only the trace rule led to any cells with invariant responses, and the results shown here are for networks trained with the trace rule.

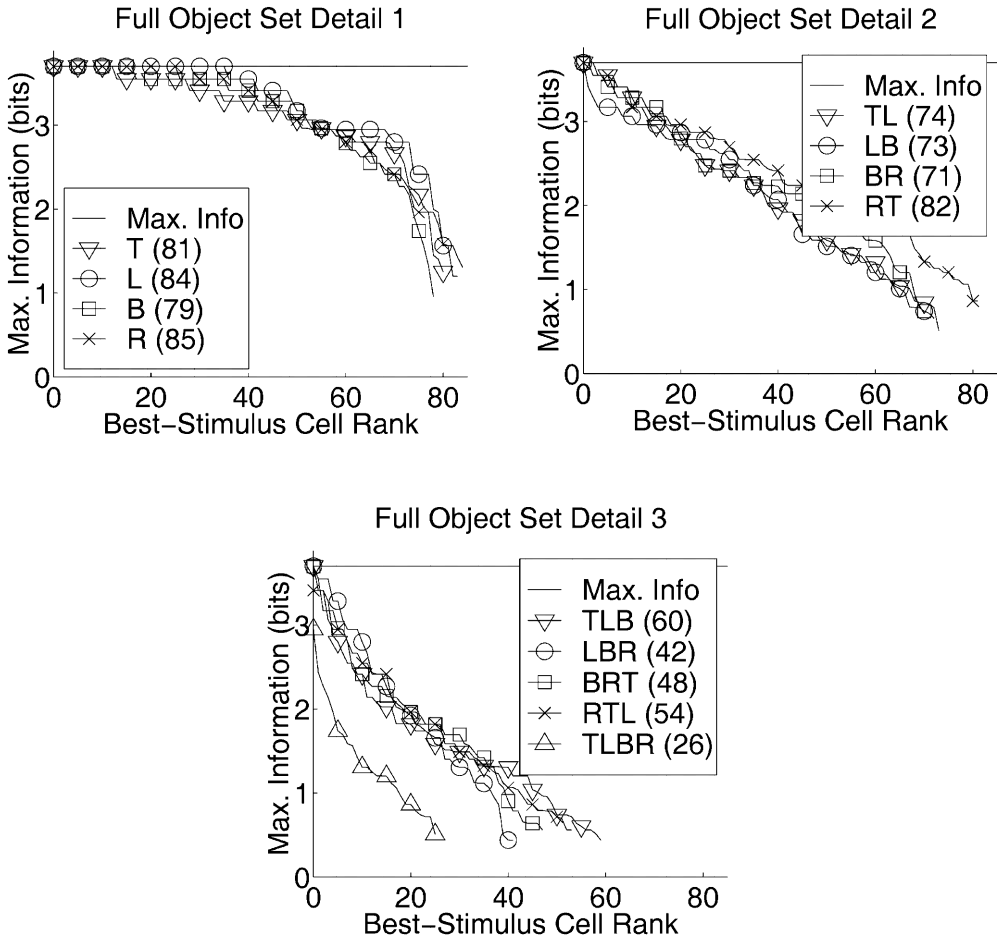

Fig. 4. Performance of VisNet2 on the full set of stimuli shown in Fig. 2. Separate graphs showing the information available about the stimulus for cells tuned to respond best to each of the stimuli are shown. The number of cells responding best to each of the stimuli is indicated in parentheses. The information values are shown for the different cells ranked according to how much information about that stimulus they encode. Separate graphs are shown for cells tuned to stimuli consisting of single features, pairs of features, and triples of features, as well as the quadruple feature stimulus TLBR

The results with VisNet trained on the set of stimuli shown in Fig. 2 with the trace rule are as follows. Firstly, it was found that single neurons in the top layer learned to differentiate between the stimuli, in that the responses of individual neurons were maximal for one of the stimuli and had no response to any of the other stimuli invariantly with respect to location. Secondly, to assess how well every stimulus was encoded for in this way, Fig. 4 shows the information available about each of the stimuli consisting of feature singles, feature pairs, feature triples and the quadruple feature stimulus 'TLBR'. The single-cell information available from the 26–85 cells with best tuning to each of the stimuli is shown. The cells in general conveyed translation-invariant information about the stimulus to which they responded, with indeed some cells which perfectly discriminated one of the stimuli from all others over every testing position for all stimuli except 'RTL' and 'TLBR'. The results presented show clearly that the VisNet paradigm can accommodate networks which can perform invariant discrimination of objects which have a subset-superset relationship. The result has important consequences for feature binding and for discriminating stimuli from other stimuli which may be supersets of the first stimulus. For example, a VisNet cell which responds invariantly to feature combination TL can genuinely signal the presence of exactly that combination, and will not necessarily be activated by T alone, or by TLB. The basis for this separation by competitive networks of stimuli which are subsets and supersets of each other is described by Rolls and Treves (1998, Sect. 4.3.6).

2.2 Feature binding in a hierarchical network with invariant representations of local feature combinations

2.2.1 Feature binding. In this section we investigate the ability of output-layer neurons to learn new stimuli if the lower layers are trained solely through exposure to simpler feature combinations from which the new stimuli are composed. A key question we address is how invariant representations of low-order feature combinations in the early layers of the visual system are able to uniquely specify the correct spatial arrangement of features in the overall stimulus, and contribute to preventing false recognition errors in the output layer.

The problem, and its proposed solution, can be considered as follows. Consider an object 1234 made from the features 1, 2, 3 and 4. The invariant low-order feature combinations might represent 12, 23 and 34. Then, if neurons at the next layer respond to combinations of these neurons, the only next-layer neurons that would respond would be those tuned to 1234, and not those tuned to, for example, 3412, which is distinguished from 1234 by the input of a pair neuron responding to 41 rather than to 23. The argument (Rolls 1992) is that low-order spatial-feature-combination neurons in the early stage of visual processing contain sufficient spatial information so that a particular combination of those low-order feature-combination neurons specifies a unique object, even if the relative positions of the low-order feature-combination neurons are not known, because they are somewhat invariant.

The architecture of VisNet is intended to solve this problem partly by allowing high spatial precision combinations of input features to be formed in layer 1. The actual input features in VisNet are, as described above, the output of oriented spatial-frequency-tuned filters, and the combinations of these formed in layer 1 might thus be thought of in a simple way as, for example, a T or an L or for that matter a Y. Then, in layer 2, application of the trace rule might enable neurons to respond to a T with limited spatial invariance (limited to the size of the region of layer 1 from which layer 2 cells receive their input). Then an 'object' such as H might be formed at a higher layer because of a conjunction of two Ts in the same small region.

To show that VisNet can actually solve this problem, we performed the experiments described below. In particular, we trained the first two layers of VisNet with feature-pair combinations, forming representations of feature pairs with some translation invariance in layer 2. Then we used feature triples as input stimuli, allowed no more learning in layers 1 and 2, and then investigated whether layers 3 and 4 could be trained to produce invariant representations of the triples, where the triples could only be distinguished if the local spatial arrangement of the features within the triple had effectively to be encoded in order to distinguish the different triples. For this experiment, we needed stimuli that could be specified in terms of a set of different features (we chose vertical, diagonal and horizontal bars) each capable of being shown at a set of different relative spatial positions (designated A, B and C), as shown in Fig. 5. The stimuli are thus defined in terms of what features are present and their precise spatial arrangement with respect to each other. The length of the horizontal and vertical feature bars shown in Fig. 5 is 8 pixels. To train the network, a stimulus (i.e. a two- or three-feature combination) is presented in a randomised sequence of nine locations in a square grid across the 128×128 input retina. The central location of the square grid is in the centre of the retina, and the eight other locations are offset 8 pixels horizontally and/or vertically from this. We refer to the two and three-feature stimuli as 'pairs' and 'triples', respectively. Individual stimuli are denoted by three numbers which refer to the individual features

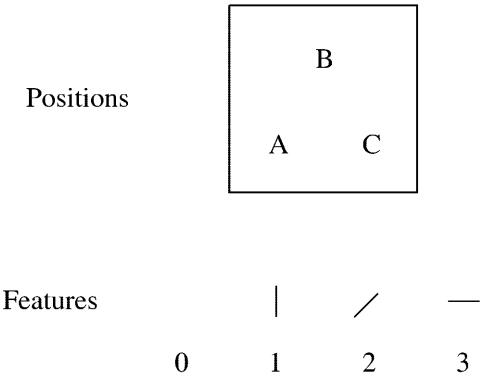

Fig. 5. Feature combinations for experiments of Sect. 2.2: there are three features denoted by 1, 2 and 3 (including a blank space 0) that can be placed in any of three positions: A, B, and C. Individual stimuli are denoted by three consecutive numbers which refer to the individual features present in positions A, B and C, respectively. In the experiments in Sect. 2.2, layers 1 and 2 were trained on stimuli consisting of pairs of the features, and layers 3 and 4 were trained on stimuli consisting of triples. Then the network was tested to show whether layer-4 neurons would distinguish between triples, even though the first two layers had only been trained on pairs. In addition, the network was tested to show whether individual cells in layer 4 could distinguish between triples even in locations where the triples were not presented during training

present in positions A, B and C, respectively. For example, a stimulus with positions A and C containing a vertical and diagonal bar, respectively, would be referred to as stimulus 102, where the 0 denotes that no feature present in position B. In total there are 18 pairs (120, 130, 210, 230, 310, 320, 012, 013, 021, 023, 031, 032, 102, 103, 201, 203, 301 and 302) and 6 triples (123, 132, 213, 231, 312 and 321). This nomenclature not only defines which features are present within objects, but also the spatial relationships of their component features. The computational problem can be illustrated by considering the triple 123. If invariant representations are formed of single features, then there would be no way that neurons higher in the hierarchy could distinguish the object 123 from 213 or any other arrangement of the three features. An approach to this problem (see e.g. Rolls 1992) is to form, early on in the processing, neurons that respond to overlapping combinations of features in the correct spatial arrangement, and then to develop invariant representations in the next layer from these neurons which already contain evidence on the local spatial arrangement of features. An example might be that with the object 123, the invariant feature pairs would represent 120, 023 and 103. Then if neurons at the next layer correspond to combinations of these neurons, the only next layer neurons that would respond would be those tuned to 123, not to, for example, 213. The argument is that the low-order spatial-feature-combination neurons in the early stage contain sufficient spatial information so that a particular combination of those low-order feature-combination neurons specifies a unique object, even if the relative positions of the low-order feature-combination neurons are not known because these neurons are somewhat translation invariant (cf. Fukushima 1988).

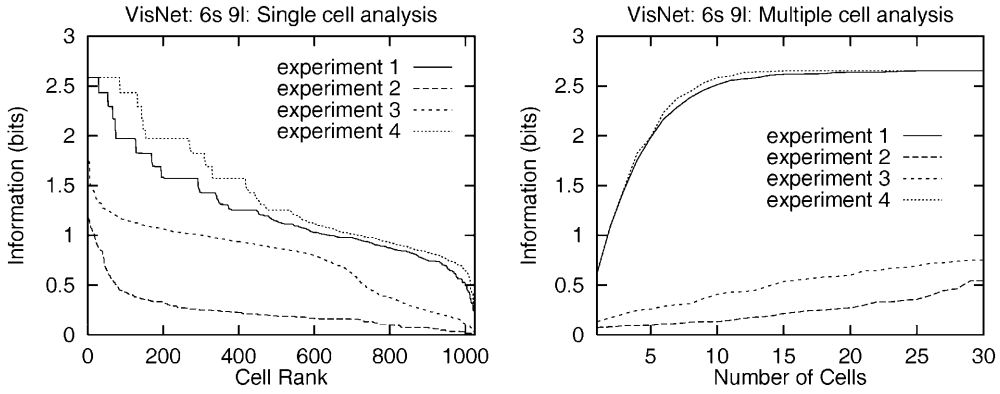

Fig. 6. Numerical results for experiments 1–4 as described in Table 5, with the trace learning rule (7). On the *left* are single-cell information measures, and on the *right* are multiple-cell information measures

Table 5. Alternative training regimes used in VisNet experiments 1–4. In the no training condition, the synaptic weights were left at their initial untrained random values

	Layers 1, 2	Layers 3, 4
Experiment 1	Trained on pairs	Trained on triples
Experiment 2	No training	No training
Experiment 3	No training	Trained on triples
Experiment 4	Trained on triples	Trained on triples

In these experiments the stimuli are constructed from pre-processed component features as discussed in Sect. 2.1. That is, base stimuli containing a single feature are constructed and filtered, and then the pairs and triples are constructed by merging these pre-processed single-feature images. In the first experiment, layers 1 and 2 of VisNet were trained with the 18 feature pairs, each stimulus being presented in sequences of nine locations across the input as described in Sect. 1.3. This led to the formation of neurons which responded to the feature pairs with some translation invariance in layer 2. Then we trained layers 3 and 4 on the six feature triples in the same nine locations, while allowing no more learning in layers 1 and 2, and examined whether the output layer of VisNet had developed transform-invariant neurons to the six triples. The idea was to test whether layers 3 and 4 could be trained to produce invariant representations of the triples, where the triples could only be distinguished if the local spatial arrangement of the features within the triple had effectively to be encoded in order to distinguish the different triples. The results from this experiment were compared and contrasted with results from three other experiments which involved different training regimes for layers 1, 2 and layers 3, 4. All four experiments are summarised in Table 5. Experiment 2 involves no training in layers 1, 2 and 3, 4, with the synaptic weights left unchanged from their initial random values. These results are included as a baseline performance with which to compare results from the other experiments, 1, 3 and 4. The model parameters used in these experiments were as described in Sect. 1.2, and as used in Rolls and Milward (2000) and Rolls and Stringer (2001).

In Fig. 6 we present numerical results for the four experiments listed in Table 5. On the left are the single cell information measures for all top (4th) layer neurons ranked in order of their invariance to the triples, while on the right are multiple cell information measures. To help to interpret these results, we can compute the maximum single cell information measure according to

$$\text{Maximum single-cell information} = \log_2(\text{number of triples}) \tag{11}$$

where the number of triples is 6. This gives a maximum single-cell information measure of 2.6 for these test cases. First, comparing the results for experiment 1 with the baseline performance of experiment 2 (no training) demonstrates that even with the first two layers trained to form invariant responses to the pairs, and then only layers 3 and 4 trained on feature triples, layer 4 is indeed capable of developing translation-invariant neurons that can discriminate effectively between the six different feature triples. Indeed, from the single cell information measures it can be seen that a number of cells have reached the maximum level of performance in experiment 1. In addition, the multiple-cell information analysis presented in Fig. 6 shows that all the stimuli could be discriminated from each other by the firing of a number of cells. Analysis of the response profiles of individual cells showed that a fourth-layer cell could respond to one of the triple feature stimuli and have no response to any other of the triple-feature stimuli invariantly with respect to location.

A comparison of the results from experiment 1 with those from experiment 3 (see Table 5 and Fig. 6) reveals that training the first two layers to develop neurons that respond invariantly to the pairs (performed in experiment 1) actually leads to improved invariance of fourth-layer neurons to the triples, as compared with when the first two layers are left untrained (experiment 3).

Two conclusions follow from these results. First, a hierarchical network which seeks to produce invariant representations in the way used by VisNet can solve the feature-binding problem. In particular, when feature pairs in layer 2 with some translation invariance are used as the input to later layers, these later layers can nevertheless build invariant representations of objects, where all the individual features in the stimulus must occur in the correct spatial position relative to each other. This is possible because the feature-combination

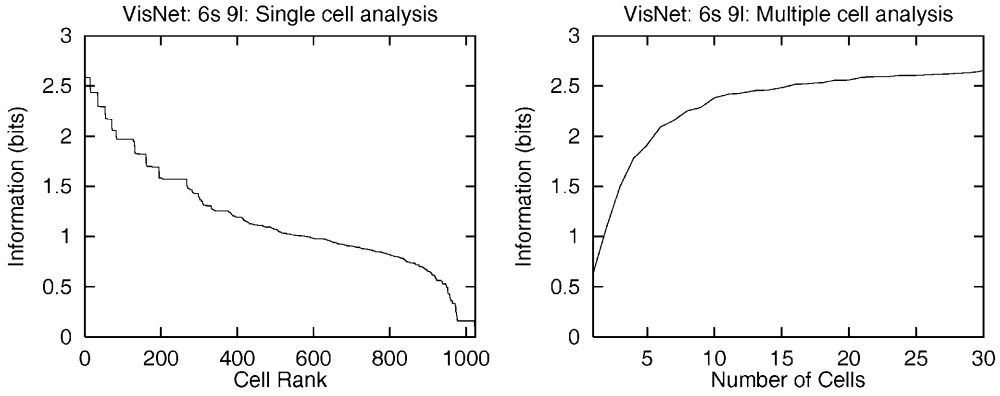

Fig. 7. Numerical results for a repeat of experiment 1 with the triples presented at only seven of the original nine locations during training, and with the trace learning rule (7). On the *left* are single-cell information measures, and on the right are multiple-cell information measures

neurons formed in the first layer (which could be trained just with a Hebbian rule) do respond to combinations of input features in the correct spatial configuration, partly because of the limited size of their receptive fields. The second conclusion is that even though early layers can in this case only respond to small feature subsets, these provide, with no further training of layers 1 and 2, an adequate basis for learning to discriminate in layers 3 and 4 stimuli consisting of combinations of larger numbers of features. Indeed, comparing results from experiment 1 with experiment 4 (in which all layers were trained on triples, see Table 5) demonstrates (see Fig. 6) that training the lower-layer neurons to develop invariant responses to the pairs offers almost as good performance as training all layers on the triples.

2.2.2 Stimulus generalisation to new locations. Another important aspect of the architecture of VisNet is that it need not be trained with every stimulus in every possible location. Indeed, part of the hypothesis (Rolls 1992) is that training early layers (e.g. 1–3) with a wide range of visual stimuli will set up feature analysers in these early layers which are appropriate later on with no further training of early layers for new objects. For example, presentation of a new object might result in large numbers of low-order feature-combination neurons in early layers of VisNet being active, but the particular set of feature-combination neurons active would be different for the new object. The later layers of the network (layer 4 in VisNet) would then learn this new set of active layer-3 neurons as the new object. However, if the new object was then shown in a new location, the same set of layer-3 neurons would be active because they respond with spatial invariance to feature combinations, and given that the layer 3–4 connections had already been set up by the new object, the correct layer-4 neurons would be activated by the new object in its new untrained location, and without any further training.

To test this hypothesis we repeated the general procedure of experiment 1 (training layers 1 and 2 with feature pairs) but then instead we trained layers 3 and 4 on the triples in only seven of the original nine locations. The crucial test was to determine whether VisNet could form top-layer neurons that responded invariantly to the six triples when presented over all nine locations, not just the seven on which the triples had been presented during training. The results are presented in Fig. 7, with single-cell information measures on the left and multiple-cell information measures on the right. VisNet is still able to develop some fourth-layer neurons with perfect invariance, that is, which have invariant responses over all nine location, as shown by the single-cell information analysis. The response profiles of individual fourth-layer cells showed that they can continue to discriminate between the triples even in the two locations where the triples were not presented during training. In addition, the multiple-cell analysis shown in Fig. 7 shows that a small population of cells was able to discriminate between all of the stimuli irrespective of location, even though for two of the test locations the triples had not been trained at those particular locations during the training of layers 3 and 4.

3 Discussion

In this paper, we first showed (in Sect. 2.1) that hierarchical feature-detecting neural networks can learn to respond differently to stimuli which consist of unique combinations of non-unique input features, and that this extends to stimuli that are direct subsets or supersets of the features present in other stimuli.

Second, we investigated (in Sect. 2.2) the hypothesis that hierarchical layered networks can produce identification of unique stimuli even when the feature-combination neurons used to define the stimuli are themselves partly translation invariant. The stimulus identification should work correctly because feature-combination neurons in which the spatial features are bound together with high spatial precision are formed in the first layer. Then at later layers, when neurons with some translation invariance are formed, the neurons nevertheless contain information about the relative spatial position of the original features. There is only then one object which will be consistent with the set of active neurons at earlier layers, which though somewhat translation-invariant as combination neurons, reflect in the activity of each neuron information about the original spatial position of the features. We note that the trace-rule training used in early layers (1 and 2) in experiments 1 and 4 would set up partly invariant feature-combination neurons, and yet the late layers (3 and 4) were able to produce, during

training, neurons in layer 4 that responded to stimuli that consisted of unique spatial arrangements of lower-order feature combinations. Moreover, and very interestingly, we were able to demonstrate in Sect. 2.2.2 that VisNet layer-4 neurons would respond correctly to visual stimuli at untrained locations, provided that the feature subsets had been trained in early layers of the network at all locations, and that the whole stimulus had been trained at some locations in the later layers of the network.

The computational problem that needs to be addressed in hierarchical networks such as the primate visual system and VisNet is how representations of features can be (e.g. translation) invariant, yet can specify stimuli or objects in which the features must be specified in the correct spatial arrangement. This is the feature-binding problem, discussed for example by von der Malsburg (1990), and arising in the context of hierarchical layered systems (Ackley et al. 1985; Fukushima 1980, 1988; Rosenblatt 1961). The results described in this paper provide one solution to the feature-binding problem. The solution which has been shown to work in the model is that in a multilayer competitive network, feature-combination neurons which encode the spatial arrangement of the bound features are formed at intermediate layers of the network. Then neurons at later layers of the network which respond to combinations of active intermediate layer neurons do contain sufficient evidence about the local spatial arrangement of the features to identify stimuli, because the local spatial arrangement is encoded by the intermediate-layer neurons. The information required to solve the visual-feature-binding problem thus becomes encoded by self-organisation into what become hard-wired properties of the network. In this sense, feature binding is not solved at run-time by the necessity to instantiate arbitrary syntactic links between sets of co-active neurons. The computational solution proposed to the superset/subset aspect of the binding problem will apply in principle to other multilayer competitive networks, although the issues considered here have not been explicitly addressed in architectures such as the neocognitron (Fukushima and Miyake 1982).

Consistent with these hypotheses about how VisNet operates to achieve, by layer 4, position-invariant responses to stimuli defined by combinations of features in the correct spatial arrangement, the effective stimuli for neurons in intermediate layers of VisNet were as follows. In layer 1, cells responded to the presence of individual features, or to low-order combinations of features (e.g. a pair of features) in the correct spatial arrangement at a small number of nearby locations. In layers 2 and 3, neurons responded to single features or to higher-order combinations of features (e.g. stimuli composed of feature triples) in more locations. These findings provide direct evidence that VisNet does operate as described above to solve the feature-binding problem.

The type of solution investigated here is thus different to the proposal of von der Malsburg (1990), that feature-selective neurons might be linked by temporal binding. There has been considerable neurophysiological investigation of this possibility (Singer et al. 1990; Abeles 1991; Hummel and Biederman 1992; Singer and Gray 1995). We note that a problem with this approach is that temporal binding might enable, say, features 1, 2 and 3 which might define one stimulus to be bound together and kept separate from, for example, another stimulus consisting of features 2, 3 and 4, but would require a further temporal binding (leading in the end potentially to a combinatorial explosion) to indicate the relative spatial positions of the 1, 2 and 3 in the 123 stimulus, so that it can be discriminated from, for example, 312. Another approach to a binding mechanism is to group spatial features based on local mechanisms that might operate for closely adjacent synapses on a dendrite (Finkel and Edelman 1987; Mel et al. 1998).

A further issue with hierarchical multilayer architectures such as VisNet is that false binding errors might occur in the following way (Mozer 1991; Mel and Fiser 2000). Consider the output of one layer in such a network in which there is information only about which pairs are present. How then could a neuron in the next layer discriminate between the whole stimulus (such as the triple 123 in the above experiment) and what could be considered a more distributed stimulus or multiple different stimuli composed of the separated subparts of that stimulus (e.g. the pairs 120, 023 and 103 occurring in three of the nine training locations in the above experiment)? The problem here is to distinguish a single object from multiple other objects containing the same component combinations (e.g. pairs). We propose that part of the solution to this general problem in real visual systems is implemented through lateral inhibition between neurons in individual layers, and that this mechanism, implemented in VisNet, acts to reduce the possibility of false recognition errors in the following two ways.

First, consider the situation in which neurons in layer N have learned to represent small feature combinations with location invariance, and where a neuron n in layer $N+1$ has learned to respond to a particular set Ω of these feature combinations. The problem is that neuron n receives the same input from layer N as long as the same set Ω of feature combinations is present, and cannot distinguish between different spatial arrangements of these feature combinations. The question is how can neuron n respond only to a particular favoured spatial arrangement Ψ of the feature combinations contained within the set Ω. We suggest that as the favoured spatial arrangement Ψ is altered by rearranging the spatial relationships of the component feature combinations, the new feature combinations that are formed in new locations will stimulate additional neurons nearby in layer $N+1$, and these will tend to inhibit the firing of neuron n. Thus, lateral inhibition within a layer will have the effect of making neurons more selective, ensuring that neuron n responds only to a single spatial arrangement Ψ from the set of feature combinations Ω, and hence reducing the possibility of false recognition.

The second way in which lateral inhibition may help to reduce binding errors is through limiting the sparsity of neuronal firing rates within layers. In our discussion above, the spurious stimuli we suggested that might lead to false recognition of triples were obtained from splitting up the component feature combinations (pairs) so that they occurred in separate training locations. However, this would lead to an increase in the number of features present in the complete stimulus; triples contain three features while their spurious counterparts would contain six features (resulting from three separate pairs). For this trivial example, the increase in the number of features is not dramatic, but if we consider, say, stimuli composed of four features where the component feature combinations represented by lower layers might be triples, then to form spurious stimuli we need to use 12 features (resulting from four triples occurring in separate locations). But if the lower layers also represented all possible pairs, then the number of features required in the spurious stimuli would increase further. In fact, as the size of the stimulus increases in terms of the number of features, and as the size of the component feature combinations represented by the lower layers increases, there is a combinatorial explosion in terms of the number of features required as we attempt to construct spurious stimuli to trigger false recognition. And the construction of such spurious stimuli will then be prevented through setting a limit on the sparsity of firing rates within layers, which will in turn set a limit on the number of features that can be represented. Lateral inhibition is likely to contribute in both these ways to the performance of VisNet when the stimuli consist of subsets and supersets of each other, as described in Sect. 2.1.

In conclusion, in this paper we have addressed one of the major issues in multilayer hierarchical networks, that of feature binding. Other issues that arise in this class of architecture and its application to learning-invariant representations are addressed elsewhere, including the effect of increasing the number of locations over which translation-invariant representations must be formed (see Wallis and Rolls (1997) and Rolls and Milward (2000) who trained with, for example, 17 faces at 49 locations), the nature of the learning rule (Rolls and Milward 2000; Rolls and Stringer 2001), the operation of the network in cluttered environments (Stringer and Rolls 2000), and the operation of a related network with the trace synaptic learning rule implemented in the recurrent collateral connections of an attractor network (Parga and Rolls 1998; Elliffe et al. 2000). Another issue that arises in this class of network is whether forming neurons that respond to feature combinations in the way described here leads to a combinatorial explosion in the number of neurons required. The solution that is proposed to this issue is to form only low-order combinations of features at any one stage of the network (Rolls 1992; cf. Feldman 1985). Using low-order combinations limits the number of neurons required, yet enables the type of computation that relies on feature-combination neurons that is analysed in this paper to still be performed. The actual number of neurons required depends also on the redundancies present in the statistics of real-world images. Even given these factors, it is likely that a large number of neurons would be required if the central visual system performs the computation of invariant representations in the manner captured by the hypotheses implemented in VisNet. Consistent with this, a considerable part of the non-human primate brain is devoted to visual information processing. The fact that large numbers of neurons and a multilayer organization are present in the primate ventral visual system is actually thus consistent with the type of model of visual information processing described here and by Rolls and Deco (2002).

References

Abeles M (1991) Corticonics – neural circuits of the cerebral cortex. Cambridge University Press, Cambridge

Ackley DH, Hinton GE, Sejnowski TJ (1985) A learning algorithm for Boltzmann machines. Cogn Sci 9: 147–169

Bartlett MS, Sejnowski TJ (1997) Viewpoint invariant face recognition using independent component analysis and attractor networks. In: Mozer M, Jordan M, Petsche T (eds) Advances in neural information processing systems 9. MIT Press, Cambridge, Mass., pp 817–823

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

Elliffe M, Rolls E, Parga N, Renart A (2000) A recurrent model of transformation invariance by association. Neural Netw 13: 225–237

Feldman JA (1985) Four frames suffice: a provisional model of vision and space. Behav Brain Sci 8: 265–289

Finkel LH, Edelman GM (1987) Population rules for synapses in networks. In: Edelman GM, Gall WE, Cowan WM (eds) Synaptic function. Wiley, New York, pp 711–757

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

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

Fukushima K (1988) A neural network for visual pattern recognition. IEEE Comput 21: 65–75

Fukushima K, Miyake S (1982) Neocognitron: a new algorithm for pattern recognition tolerant of deformations and shifts in position. Pattern Recog 15: 455–469

Hawken MJ, Parker AJ (1987) Spatial properties of the monkey striate cortex. Proc R Soc Lond B 231: 251–288

Hummel J, Biederman I (1992) Dynamic binding in a neural network for shape recognition. Psychol Rev 99: 480–517

Malsburg C von der (1990) A neural architecture for the representation of scenes. In: McGaugh JL, Weinberger NM, Lynch G (eds) Brain organization and memory: cells, systems and circuits. Oxford University Press, New York, pp 356–372

Mel BW, Fiser J (2000) Minimizing binding errors using learned conjunctive features. Neural Comput 12: 731–762

Mel BW, Ruderman DL, Archie KA (1998) Translation-invariant orientation tuning in visual "complex" cells could derive from intradendritic computations. J Neurosci 18: 4325–4334

Mozer M (1991) The perception of multiple objects: a connectionist approach. MIT Press, Cambridge Mass

Panzeri S, Treves A, Schultz S, Rolls ET (1999) On decoding the responses of a population of neurons from short time windows. Neural Comput 11: 1553–1577

Parga N, Rolls ET (1998) Transform-invariant recognition by association in a recurrent network. Neural Comput 10: 1507–1525

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

Rolls ET (1994) Brain mechanisms for invariant visual recognition and learning. Behav Process 33: 113–138

Rolls ET (1995) Learning mechanisms in the temporal lobe visual cortex. Behav Brain Res 66: 177–185

Rolls ET (2000) Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. Neuron 27: 1–20

Rolls ET, Deco G (2002) Computational neuroscience of vision. Oxford University press, Oxford

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

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

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

Rolls ET, Treves A (1998) Neural networks and brain function. Oxford University Press, Oxford

Rolls ET, Treves A, Tovee M, Panzeri S (1997a) Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. J Comput Neurosci 4: 309–333

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

Rosenblatt F (1961) Principles of neurodynamics: perceptrons and the theory of brain mechanisms. Spartan, Washington DC

Salinas E, Abbott LF (1997) Invariant visual responses from attentional gain fields. J Neurophysiol 77: 3267–3272

Singer W, Gray CM (1995) Visual feature integration and the temporal correlation hypothesis. Annu Rev Neurosci 18: 555–586

Singer W, Gray C, Engel A, Konig P, Artola A, Brocher S (1990) Formations of cortical cell assemblies. In: Proceedings of the Cold Spring Harbor Symposium on Quantitative Biology, pp 939–952

Stone JV (1996) A canonical microfunction for learning perceptual invariances. Perception 25: 207–220

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

Tanaka K, Saito H, Fukada Y, Moriya M (1991) Coding visual images of objects in the inferotemporal cortex of the macaque monkey. J Neurophysiol 66: 170–189

Wallis G, Rolls ET (1997) A model of invariant object recognition in the visual system. Prog Neurobiol 51: 167–194