# A Model of Invariant Object Recognition in the Visual System

Guy Wallis Edmund T. Rolls

February 15, 1996

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

**Abstract**

Neurons in the ventral stream of the primate visual system exhibit responses to the images of objects which are invariant with respect to natural transformations such as translation, size, and view. Anatomical and neurophysiological evidence suggests that this is achieved through a series of hierarchical processing areas. In an attempt to elucidate the manner in which such representations are established, we have constructed a model of cortical visual processing which seeks to parallel many features of this system, specifically the multi-stage hierarchy with its topologically constrained convergent connectivity. Each stage is constructed as a competitive network utilising a modified Hebb-like learning rule, called the trace rule, which incorporates previous as well as current neuronal activity. The trace rule enables neurons to learn about whatever is invariant over short time periods (e.g. 0.5 s) in the representation of objects as the objects transform in the real world. The trace rule enables neurons to learn the statistical invariances about objects during their transformations, by associating together representations which occur close together in time. We show that by using the trace rule training algorithm the model can indeed learn to produce transformation invariant responses to natural stimuli such as faces.

## 1 Introduction

### Background

There is evidence that over a series of cortical processing stages, the visual system of primates produces a representation of objects which shows invariance with respect to, for example, translation, size, and view, as shown by recordings from single neurons in the temporal lobe (Desimone, 1991; Rolls, 1992; Tanaka *et al.*, 1991). In his 1992 paper, Rolls reviews much of this work with specific regard to those cells responsive to faces, and goes on to advance a theory for how these neurons could acquire their transform independent selectivity based upon the known physiology of visual cortex and simple self-organising principles. It is this analysis which forms the basis for the network described here.

Fundamental elements of Rolls' (1992) hypothesis are:

- A series of competitive networks, organised in hierarchical layers, exhibiting mutual inhibition over a short range within each layer.
- A convergent series of connections from a localised population of cells in preceding layers to each cell of the following layer, thus allowing the receptive field size of cells to increase through the visual processing areas or layers.
- A modified Hebb-like learning rule incorporating a temporal trace of each cell's previous activity, which, it is suggested, will enable the neurons to learn transform invariances.

The first two elements of his hypothesis are used to constrain the general architecture of a network model intended to learn invariant representations of objects. As later simulation results using this model will reveal, invariant representations can indeed be learnt. However, the success crucially depends upon the inclusion of the modified version of the Hebb rule.

## The trace rule

The basic learning rule implemented in the simulations utilises 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) and Rolls (1992). 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 (Rolls, 1992; Rolls, 1995).

To clarify the reasoning behind this point, consider the situation in which a single neuron is strongly activated by a stimulus forming part of a real world object. The trace of this neuron's activation will then gradually decay over a time period in the order of 0.5s, say. If, during this limited time window, the net is presented with a transformed version of the original stimulus then not only will the initially active afferent synapses modify onto the neuron, but so also will the synapses activated by the transformed version of this stimulus. In this way the cell will learn to respond to either appearance of the original stimulus. The cell will not, however, tend to make spurious links across stimuli that are part of different objects because of the unlikelihood of one object consistently following another.

Various biological bases for this temporal trace have been advanced[1] :

[1]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 extended cell firing should be reflected in representing the trace as an external firing rate, rather than an internal signal.

- The persistent firing of neurons for as long as 100-400ms 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 cortical areas (O'Reilly & Johnson, 1994; Rolls, 1995)[2].
- The binding period of glutamate in the NMDA channels, which may last for 100 or more ms, may implement a trace rule by producing a narrow time window over which the *average* activity at each presynaptic site affects learning (Rolls, 1992; Rhodes, 1992; Földiák, 1992).
- Chemicals such as nitric oxide may be released during high neural activity and gradually decay in concentration over a short time window during which learning could be enhanced(Földiák, 1992; Montague *et al.*, 1991).

The trace update rule used in these simulations is equivalent to both Földiák's and to the earlier rule of Sutton & Barto (1981), and can be summarised as follows:

$$\Delta w_{ijk}{}^{(t)} = \alpha \overline{y}_{ij}{}^{(t)} . x_k$$

where

$$\overline{y}_{ij}{}^{(t)} = (1 - \eta) y_{ij}{}^{(t)} + \eta \overline{y}_{ij}{}^{(t-1)}$$

| | | | |
|---|---|---|---|
| $x_k$: | $k^{th}$ input to the neuron. | $y_{ij}$: | Output from the neuron in the $i^{th}$ column, $j^{th}$ row of the output layer. |
| $\overline{y}_{ij}^{(t)}$: | Trace value of neuron $i, j$ at time step $t$. | $\alpha$: | Learning rate. Annealed between unity and zero. |
| $w_{ijk}$: | Synaptic weight between $k^{th}$ input and neuron in column $i$, row $j$. | $\eta$: | Trace value. The optimal value varies with presentation sequence length. |

To bound the growth of each cell's dendritic weight vector, $w_{ij}$ is explicitly normalised, a method similarly employed by von der Malsburg (1973) amongst others. An alternative, more biologically relevant implementation, using a local weight bounding operation which utilises a form of heterosynaptic long-term depression (Brown *et al.*, 1990, for review), has in part been explored using a version of the Oja rule (Oja, 1982).

[2]The prolonged firing of inferior temporal cortex neurons during memory delay periods of several seconds, and associative links reported to develop between stimuli presented several seconds apart, (Miyashita & Chang, 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.

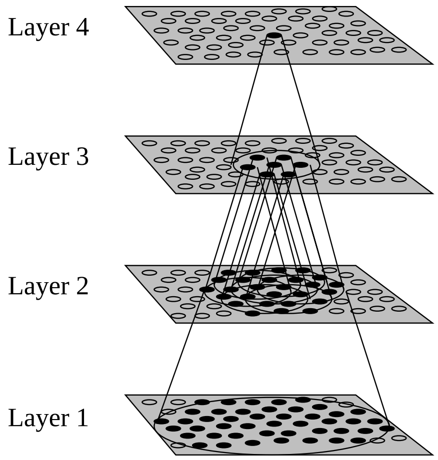



Figure 1: Stylised image of the VisNet four layer network. Convergence through the network is designed to provide fourth layer neurons with information from across the entire input retina.

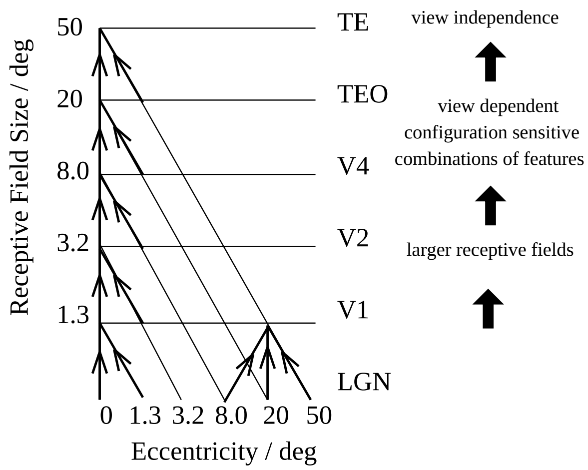


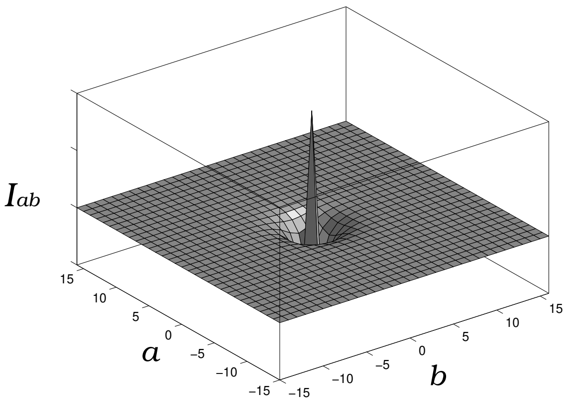

Figure 2: Convergence in the visual system - adapted from Rolls (1992). V1: visual cortex area V1; TEO: posterior inferior temporal cortex; TE: inferior temporal cortex (IT).

## The network

The network itself is designed as a series of hierarchical, convergent, competitive networks, in accordance with the hypothesis advanced above - and in part described in earlier work (Wallis *et al.*, 1993). 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 figure 1. This corresponds to the scheme described by many researchers (Van Essen *et al.*, 1992; Rolls, 1992, for example) as present in the primate visual system - see figure 2.

The forward connections to a cell in one layer are derived from a topologically related and confined region of the preceding layer. The choice of whether a connection between layers exists or not, is based upon a gaussian distribution of connection probabilities which roll off radially from the focal point of connections for each neuron. In practice a minor extra constraint precluding the repeated connection of any cells, is also applied. Each cell receives 100 connections from a 32x32 grid of cells in the preceding layer, initially with a 67% probability that a connection comes from within 4 cells of the distribution centre - although the effective radius of convergence increases slightly through the layers. Figure 1 shows the general convergent network architecture used. Localisation 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 in the binding problem, as described in the discussion section.

Modelling topological constraints in connectivity leads to an issue concerning neurons




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

Figure 3: Contrast enhancing filter, which has the effect of local lateral inhibition. The parameters $\delta$ and $\sigma$ are variables used to modify the amount and extent of inhibition respectively.

at the edges of the network layers. In principle these neurons may either receive no input from beyond the edge of the preceding layer, or have their connections repeatedly sample neurons at the edge of the previous layer. In practice either solution is liable to introduce artificial weighting on the few active inputs and hence cause the edge to have unwanted influence over the development of the network as a whole. In the real brain such edge-effects would be naturally smoothed by the transition of the locus of cellular input from the fovea to the low acuity periphery of the visual field. However, it poses a problem here because we are in effect only simulating the small high acuity foveal portion of the visual field in our simulations. As an alternative to the former solutions we have elected to form the connections into a toroid, such that connections wrap back onto the network from opposite sides. This solution has the advantage of making all of the boundaries effectively invisible to the network. However, one consequence of this form of connectivity is that all cells in the final layer will have access to information from the entire retina - which would not otherwise be the case. To confirm that this does not of itself result in neurons with translation invariant responses it will be important to judge as a baseline of performance, what level of invariance neurons in an untrained net exhibit.

In order to act as a competitive network some form of lateral inhibition is required within each layer, which should help to ensure that all stimuli presented are evenly represented by the neurons in each layer. In these simulations, a local inhibitory function is applied to each neuron and its neighbouring cells, similar to the work of von der Malsburg (1973). This inhibition is not winner-take-all (with only one neuron left active after the competition), but instead graded, to produce a soft competitive network. This soft competition can be advantageous in the way neurons are allocated to stimuli (Bennett, 1990), but in particular has the important advantage of leading to distributed representations, which allow more stimuli to be represented in the network. This inhibition is essential if the net is to act as a competitive net and should help to ensure that all stimuli presented are evenly represented by the neurons in each layer.

Inhibition is simulated via a linear local contrast enhancing filter, consisting of a positive central spike surrounded by a negative gaussian field, the general shape and

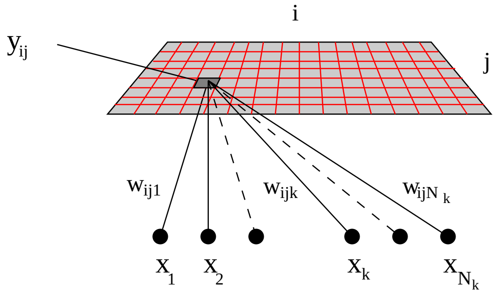



Figure 4: Pictorial representation of the neural elements referred to in calculating neural activity.

formula for which is given in figure 3. As with the network connectivity, inhibition acts toroidally, and as before the effects of this artificial connection scheme will have to be tested. In practice, the choice of parameters describing the mask $\delta = 1$ and $\sigma = 1$ mean that inhibition is largely restricted to the nearest neuronal neighbours, thus reducing the spread of the effects of inhibition at the edges. In addition to this local inhibitory mechanism there is also a global check on the average cellular activity within a layer, which is maintained at a constant value by normalising the firing rates across a layer. The intention in doing this is once again, to ensure that learning for each stimulus in each location is approximately constant. The firing rates themselves are a nonlinear function of the original neural activation, realised by raising the activity of the neuron to a fixed power - the optimal value of which is described later. The biological rationale for this is the greater than linear increase in neuronal firing as a function of the activation of the neuron which characterises the first steeply rising portion close to threshold of the sigmoid-like activation function of real neurons. More sophisticated models of the neural transfer function using non-zero or variable thresholds, or saturation are not considered here, but would be obvious ideas to try in subsequent new testing of the model.

The general calculation of the response of a neuron $y$, in row $i$ column $j$ of any particular layer - shown pictorially in figure 4 - is then:

$$y_{ij} = \frac{\left[\sum_{a,b}^{N_a,N_b}\left(I_{(a-i)(b-j)}\sum_k^{N_k} x_k w_{iab}\right)\right]^n}{\sum_{i,j}^{N_i,N_j}\left[\sum_{a,b}^{N_a,N_b}\left(I_{(a-i)(b-j)}\sum_k^{N_k} x_k w_{iab}\right)\right]^n}$$

| | | | |
|---|---|---|---|
| $x_k$: | $k^{th}$ input to the neuron. | $y_{ij}$: | Output from the neuron in the $i^{th}$ column, $j^{th}$ row of the output layer. |
| $N_i, N_j$: | Dimensions of the output layer. $N_i = N_j = 32$. | $N_a, N_b$: | Dimensions of the inhibitory field. $N_a = N_b = 32$ |
| $n$: | Transfer function nonlinearity. $n = 6$ in layer one, $n = 2$ in all subsequent layers. | $N_k$: | Number of inputs to a neuron. $N_k = 272$ in layer one, $N_k = 100$ otherwise. |
| $w_{ijk}$: | Synaptic weight between $k^{th}$ input and neuron in column $i$, row $j$. | $I_{ab}$ | Inhibitory mask value. See figure 3. |

## Network input

Several unsupervised neural models have been successful in learning to produce cells with the centre surround response properties of cells in the lateral geniculate nucleus, and the oriented edge and bar sensitive simple cells of V1 (von der Malsburg, 1973; Nass & Cooper, 1975; Linsker, 1986). In this work we make no attempt to learn the response properties of simple cells, but instead start with a fixed feature extraction level, as have other researchers in the field (Hummel & Biederman, 1992; Buhmann *et al.*, 1991; Fukushima, 1980, e.g.), with the intention of simulating the more complicated response properties of cells between V1 and the inferior temporal cortex (IT).

The response characteristics of the input neurons to the network are provided by a series of fixed, spatially tuned filters with image contrast sensitivities chosen to accord with the general tuning profiles observed in the simple cells of V1. In the current model only even symmetric - bar detecting - filter shapes are used, which take the form of a gaussian shape along the axis of orientation tuning for the filter, and a difference of gaussians along the perpendicular axis.

This filter is referred to henceforth as an oriented difference of gaussians, or DOG filter[3]. It was chosen in preference to the often used Gabor filter on the grounds of its better fit to available neurophysiological data including its zero D.C. response (Hawken & Parker, 1987; Wallis, 1994). Any zero D.C. filter can of course produce a negative as well as positive output, which would mean that this simulation of a simple cell would permit negative as well as positive firing. In contrast to some other models 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 normalised across scales to compensate for the low frequency bias in the images of natural objects.

Cells of layer one receive a topologically consistent, localised, 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. Figure 5 shows pictorially the general filter sampling paradigm, and figure 6 the typical connectivity to a layer 1 cell from the filters of the input layer. The blank squares indicate that no connection exists

[3] We warmly thank Professor R. Watt, of Stirling University, for assistance with the implementation of this filter scheme.

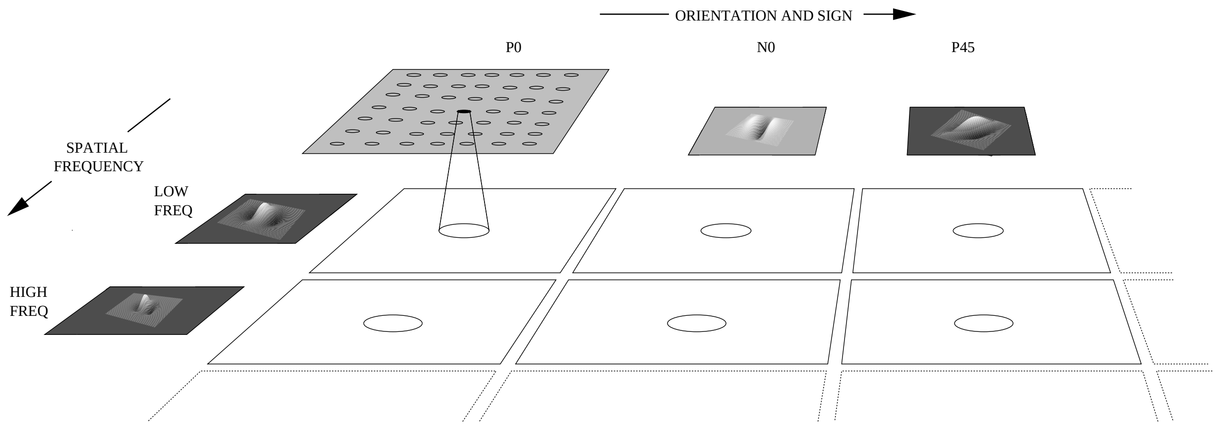



Figure 5: The filter sampling paradigm. Here each square represents the retinal image presented to the network after being filtered by a DOG filter of the appropriate orientation sign and frequency. The circles represent the consistent retinotopic coordinates used to provide input to a layer one cell. The filters double in spatial frequency towards the reader. Left to right the orientation tuning increases from $0^o$ in steps of $45^o$, with segregated pairs of positive (P) and negative (N) filter responses.

between the layer one cell chosen and the filters of that particular orientation, sign, and spatial frequency.

## Measures for network performance

Before describing the results of the experimentation we describe a metric for how well the network has solved the classification problem, and how invariant its responses are. For our purposes a neuron can be said to have learnt an invariant representation if it discriminates one set of stimuli from another set, across all transformations. For example, a neuron's response is translation invariant if its response to one set of stimuli is consistently higher than for all other stimuli irrespective of presentation location. Note that we state 'set of stimuli' since neurons in IT cortex are not generally selective for a single stimulus but rather a sub-population of stimuli (Baylis *et al.*, 1985; Abbott *et al.*, 1996).

Essentially any measure should ensure low variance in neural response across the transformation (transform invariance) and high variance across stimuli (stimulus selectivity). The immediately intuitive way to measure this would be to run a two-way classification ANOVA, using location and stimulus as repeated measures. However, these measures are not truly random, and it is therefore unclear whether the underlying assumptions of the analysis are violated. With this in mind we have instead chosen to use an information measure. This measure relates how much information one gains about the neural response if the stimulus is known (should be high) verses how much information one gains by knowing the location at which the presentation was made (should be low). The details of how to calculate this measure referred to now as the 'discrimination factor', are given in the appendix.

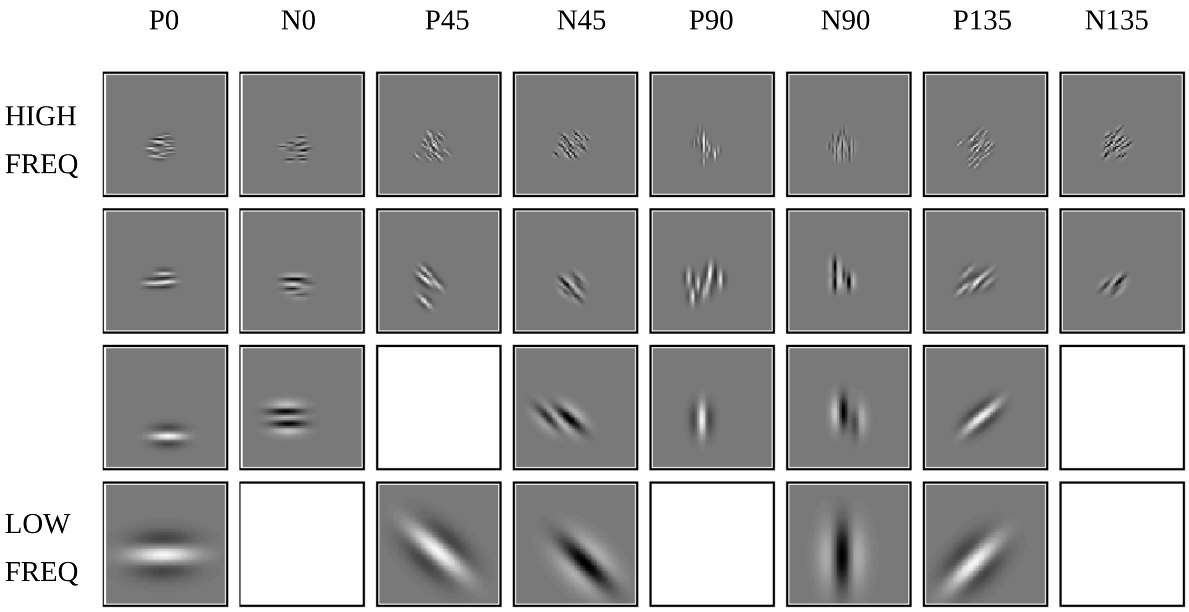



Figure 6: Typical connectivity between a single cell in the first layer of the network and the input layer, represented by plotting the receptive fields of every input layer cell connected to the particular layer one cell. Here the overall receptive field of the layer one cell is centred just below the centre-point of the retina. The connection scheme allows for relatively fewer connections to lower frequency cells than to high frequency cells in order to cover a similar region of the input at each frequency. A blank square indicates that there is no connection to the layer one neuron from an input neuron with that particular filter type.

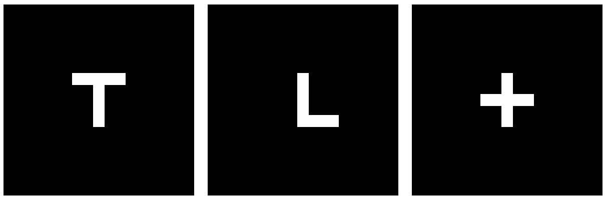



Figure 7: The three stimuli used in the first two experiments.

# 2 Experimentation

## Introduction

Having established a network model we now describe four experiments in which the theory is implemented for a variety of stimuli undergoing a variety of natural transformations. In each case the network should exhibit neurons whose response is largely invariant across a transformation and highly discriminating between stimuli or sets of stimuli.

## 'T','L' and '+' as stimuli part I: Learning invariance

One of the classical properties of IT face cells is their invariant response to facial stimuli translated across the visual field (Tovee *et al.*, 1994). In this first experiment, the learning of translation invariant representations by VisNet was investigated.

In order to test the network a set of three stimuli, based upon probable 3-D edge cues - consisting of a 'T', 'L' and '+' shape - was constructed[4]. The actual stimuli used are shown in figure 7. These stimuli were chosen partly because of their significance as form cues, but on a more practical note because they each contain the same fundamental features - namely a horizontal bar conjoined with a vertical bar. In practice this means that the oriented simple cell filters of the input layer can not distinguish these stimuli on the basis of which features are present. As a consequence of this, the representation of the stimuli received by the network is non-orthogonal and hence considerably more difficult to classify than was the case in earlier experiments involving the trace rule - see Földiák (1991). The expectation is that layer one neurons will learn to respond to spatially selective combinations of the basic features thereby helping to distinguish these non-orthogonal stimuli. The trajectory followed by each stimulus consists of sweeping left to right horizontally across three locations in the top row, and then sweeping back, right to left across the middle row, before returning to the right hand side across the bottom row - tracing out a 'Z' shape path across the retina. Unless stated otherwise this pattern of nine presentation locations was adopted in all image translation experiments.

Training was carried out by permutatively presenting all stimuli in each location a total of 800 times. The sequence described above was followed for each stimulus, with the

[4] Chakravarty (1979) describes the application of these shapes as cues for the 3D interpretation of edge junctions, and Tanaka et. al. (1991) have demonstrated the existence of cells responsive to such stimuli in IT.

sequence start point and direction of sweep being chosen at random.

Figure 8 shows the response of a first layer neuron selective for the 'T' stimulus. The weighted sum of all filter inputs reveals the combination of horizontally and vertically tuned filters in identifying the stimulus. In this case many connections to the lower frequency filters have been reduced to zero by the learning process, except at the relevant orientations. This contrasts strongly with the random wiring present before training, as seen previously in figure 6. Likewise, Figure 9 depicts two neuronal responses, but now from the two intermediate layers of the network, taken from the top 30 most highly invariant cells, not merely the top two or three! The gradual increase in the discrimination indicates that the tolerance to shifts of the preferred stimulus gradually builds up through the layers.

The results for layer four neurons are illustrated in figure 10. By this stage translation invariant, stimulus identifying cells have emerged. The response profiles confirm the high level of neural selectivity for a particular stimulus irrespective of location.

Figure 11 contrasts the measure of invariance, or discrimination factor, achieved by cells in the four layers, averaged over five separate runs of the network. Translation invariance clearly increases through the layers, with a considerable increase in translation invariance between layers three and four. This sudden increase may well be a result of the geometry of the network, which enables cells in layer 4 to receive inputs from any part of the input layer.

Having established that invariant cells have emerged in the final layer, we now move on to perhaps the most important issue. This concerns the role of the trace rule in the overall performance of the network. Convergence of the network connections and the toroidal connection scheme employed at the edges of each layer almost certainly provide a number of cells responsive to stimuli in more than one location which may therefore exhibit some degree of translation invariance. In order to assess the importance of these factors relative to the importance of training the network using the trace rule, the network was tested under two new conditions. Firstly, the performance of the network was measured before learning occurs, that is with its initially random connection weights. Secondly, the network was trained with $\eta$ in the trace rule set to 0, which causes learning to proceed in a memoryless, standard Hebbian fashion.

Figure 12 shows the results under the three training conditions. The results show that the trace rule is the decisive factor in establishing the invariant responses in the layer four neurons. It is interesting to note that the Hebbian learning results are actually *worse* than those achieved by chance in the untrained net. In general, with Hebbian learning, the most highly discriminating cells barely rate higher than 1. This value of discrimination corresponds to the case in which a cell responds to only one stimulus and in only one location. The poor performance with the Hebb rule comes as a direct consequence of the presentation paradigm being employed. If we consider an image as representing a vector in multidimensional space, a particular image in the top left hand corner of the input retina will tend to look more like any other image in that same location than the same image presented elsewhere. A simple competitive network using just Hebbian learning will thus tend to categorise images by *where* they are rather than what they are - the exact opposite of what the net was intended to learn. This result is probably the most important thus far, since it indicates that a small memory trace acting in the standard Hebbian learning

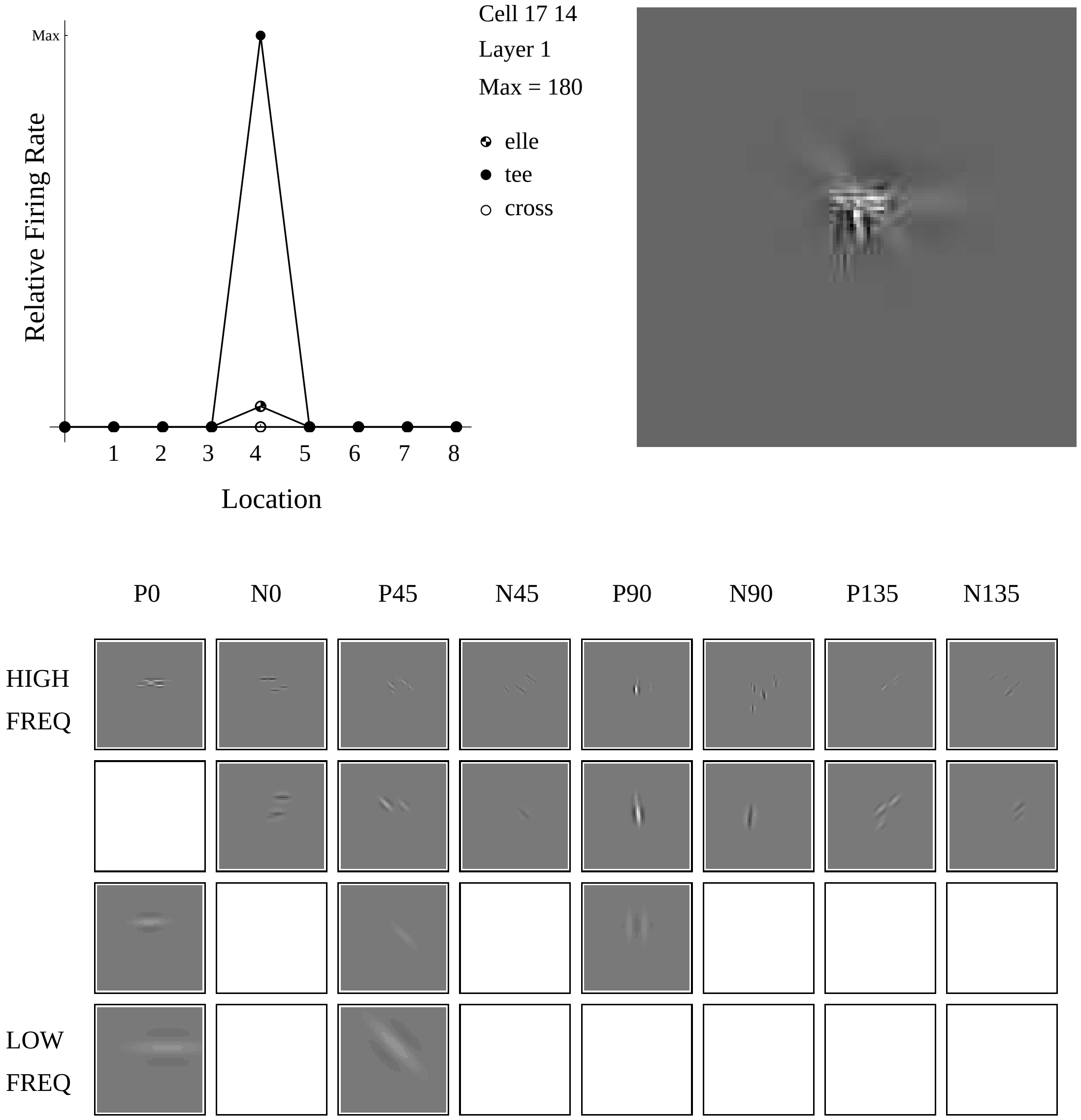



Figure 8: The top left graph shows the response of a layer one neuron to the three training stimuli for the nine training locations. Alongside this are the results of summating all the filter inputs to the neuron. The diagram below shows the same input reconstruction results but separated into the four rows of differing spatial frequency, and eight columns representing the four filter tuning orientations in positive and negative complementary pairs. The discrimination factor for this cell was 1.04.

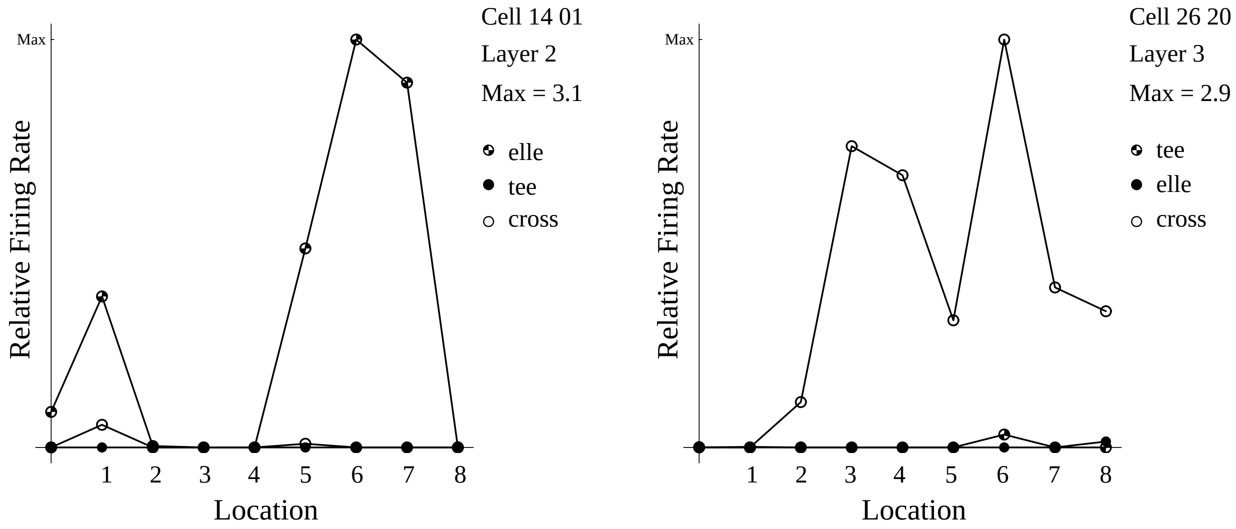



Figure 9: Response profiles for two intermediate layer neurons - discrimination factors 1.34 and 1.64.

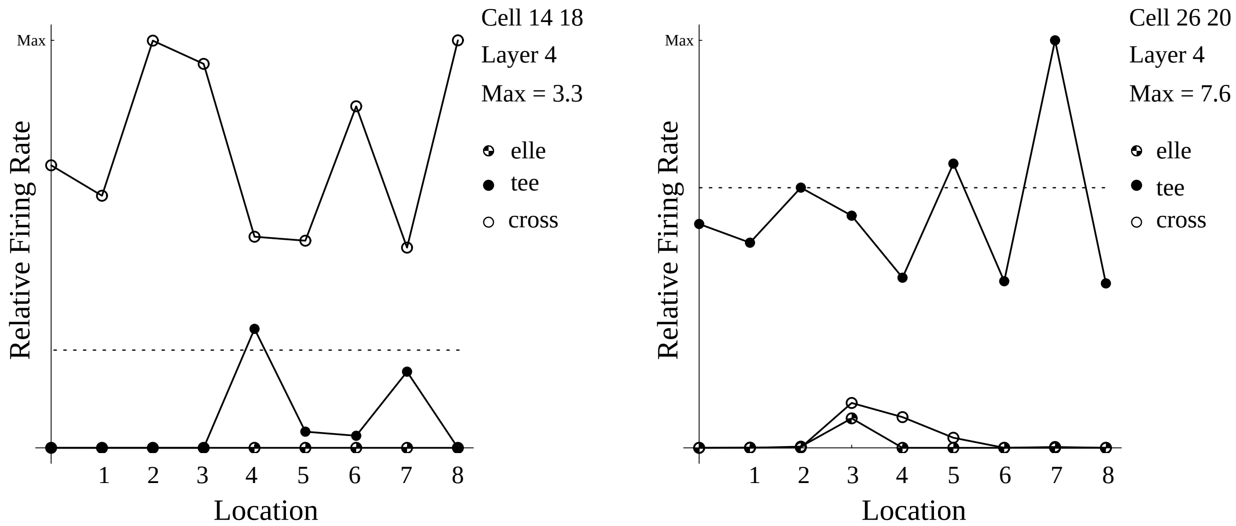



Figure 10: Response profiles for two fourth layer neurons - discrimination factors 4.07 and 3.62.

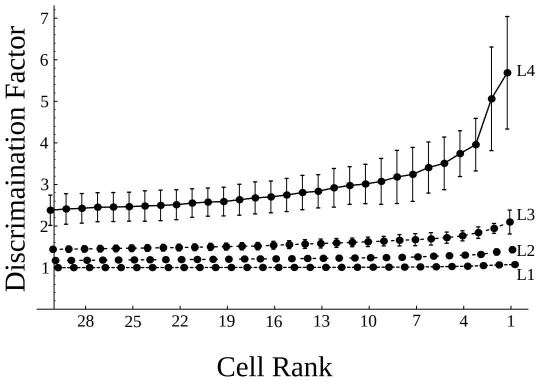



Figure 11: Variation in neural discrimination factors as a measure of performance for the top 30 most highly discriminating cells through the four layers of the network, averaged over five runs of the network.

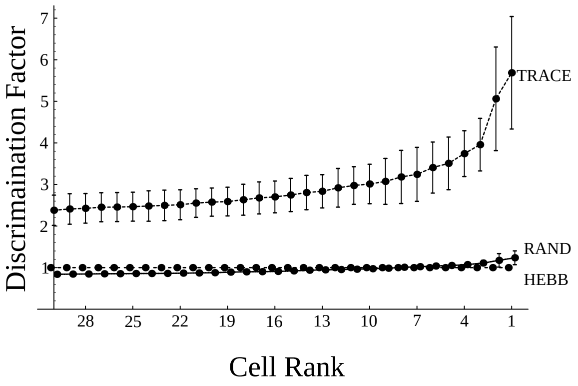



Figure 12: Variation in neural discrimination factors as a measure of performance for the top 30 most highly discriminating cells in the fourth layer for the three training regimes, averaged over five runs of the network.

paradigm can radically alter the normal vector averaging, image classification, performed by a Hebbian based competitive network.

One question which emerges about the representation in the final layer of the network relates to how evenly the network divides up its resources to represent the learnt stimuli. It is conceivable that one stimulus stands out among the set of stimuli by containing very distinctive features which would make it easier to categorise. This may produce an unrepresentative number of neurons with high discrimination factors which are in fact all responding to the same stimulus. It is important that at least some cells code for (or provide information about) each of the stimuli. As a simple check on this, the preferred stimulus of each cell was found and the associated measure of discrimination added to a total for each stimulus, which in practice never varied by more than a factor of 1.3:1 for all stimuli.

Also with regard to the representation in the final layer, it would be informative to see how the overall pattern of activity varies as a stimulus changes location or as different stimuli are presented in a single location. One would expect to see consistent activity for any one stimulus across all locations, but variable activity when changing stimuli in any one location. Since there are over a thousand neurons in the final layer, it is not easy to discern changes in the patterns of activity simply by displaying a sequence of snap-shots of the neural responses. Instead, the average response of each neuron is recorded. In the upper left graph of figure 13 the neural responses are averaged for presentations of the 'T' stimulus across the middle three of the nine locations, and in the lower half they are averaged over three stimuli presented centrally. The large contrast between neighbouring cells in the top graph indicates that some cells are consistently more active than their neighbours across translation of the 'T' stimulus. In the lower left graph however, the average responses are less distinct suggesting that the average of neural activity across

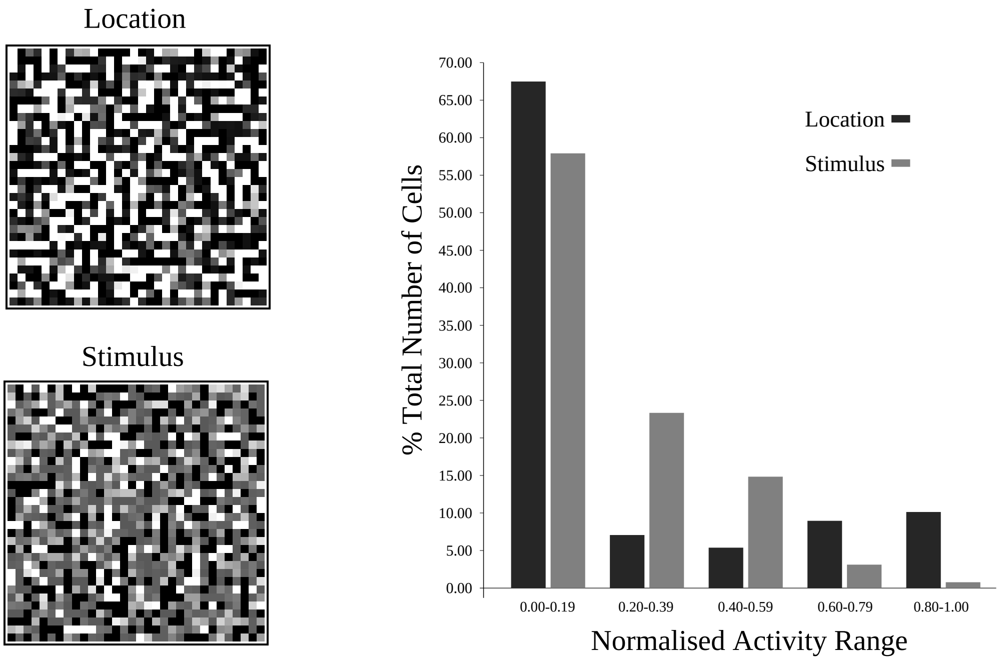



Figure 13: Comparison of the average activity seen in the fourth layer to multiple presentations. 'Location': the 'T' stimulus is presented in all 9 training locations. 'Stimulus': all three stimuli are presented in the same central location. The left hand graphs show actual averages for the 32x32 grid of cells. The right hand graph shows a binned histogram of average activity.

stimuli shows very little patterning - since all neurons exhibit similar average activity.

This contrast is brought out even more clearly in the right hand graph of this figure, in which the average activity histograms in both cases are plotted. Here the location histogram is the average of nine histograms gained from different combinations of responses to the 'T' stimulus across three of the nine training locations. The stimulus histogram is also the average of the nine histograms, each generated by presenting the three stimuli at one of the nine training locations. When looking across location one sees that nearly two thirds of the cells are largely inactive, many of which will be tuned to the other two stimuli and others of which have not learnt due to competition within the layers. The majority of cells which do respond across all locations do so strongly, having close to maximal average responses. Averaging across stimuli however, reveals that the majority of cells which respond do so only relatively weakly - possibly responding to only two of the three stimuli. Taken together, the results reveal that some 20% of cells in the final layer show consistent, active responses to a stimulus shown in the nine training locations, whereas only 4% appear to react comparably well to all three stimuli shown in one location. Both results lend further weight to our claim that a large number of neurons in the final layer have successfully learnt an invariant response to the translation of each of the stimuli.

As a final experiment in this first series we turn to the response of layer four neurons to shifts in the presentation location away from the training locations. Figure 14 shows the

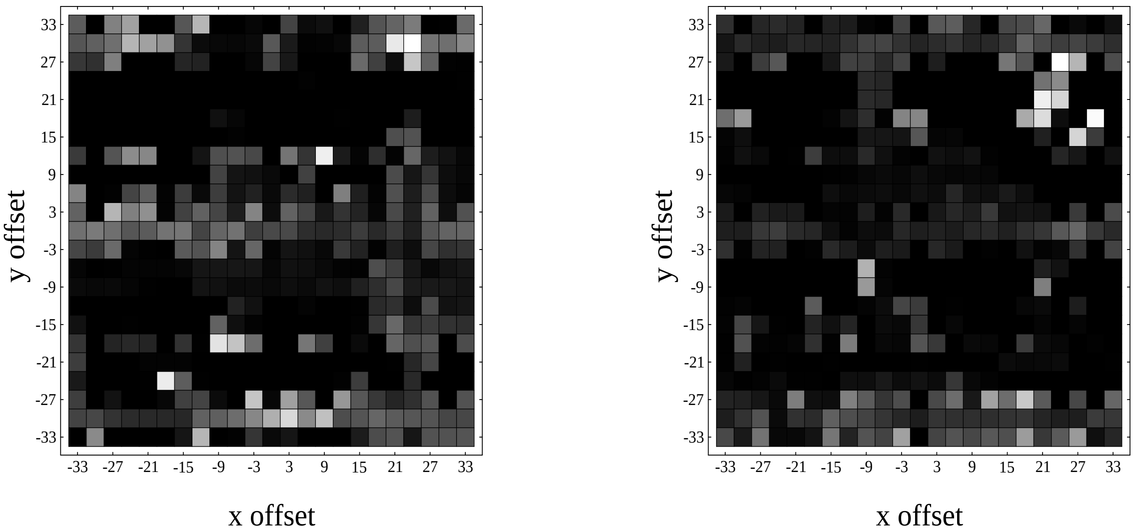



Figure 14: Response of two layer four neurons to their preferred stimuli presented over a range of x,y coordinate offsets from the central location. Lighter squares indicate presentation locations which produced high firing.

response of two layer four neurons as their preferred stimulus is presented in 596 locations arranged on a square grid over the input retina. The training locations correspond to offsets (-30,30), (0,30), (30,30), (-30,0), (0,0), (30,0), (-30,-30), (0,-30) and (30,-30). The cells selected are the same as those whose responses were tested earlier, and which appear in figure 10. We might predict some generalisation around the training locations, due to the spatial non-determinacy of the filters, particularly for shifts with only a horizontal or only a vertical component. In practice this is what emerges from the analysis, although there is a tendency for greater tolerance to horizontal shifts than vertical ones. This presumably reflects the significance of the vertical placement of the horizontal segment of the three stimuli for performing their dissociation. This analysis indicates that some local shift invariance emerges from the properties of the filters providing the input to the network, and that further, more global, shift invariance is provided by the trace learning rule, which enables the same stimulus placed at different parts of the retina to activate the same layer four neurons.

## 'T','L' and '+' as stimuli part II: Optimal network parameters

The second series of investigations using the 'T','L' and '+' stimuli, centres upon finding optimal parameters for elements of the network, such as the degree of non-linearity in the neurons, and the optimal trace time constant $\eta$, which controls the relative effect of previous activities on current learning as described in the introduction.

In the following series of results network performance is gauged in terms of a single 800 epoch run of the network with the average discrimination factor for the top sixteen cells

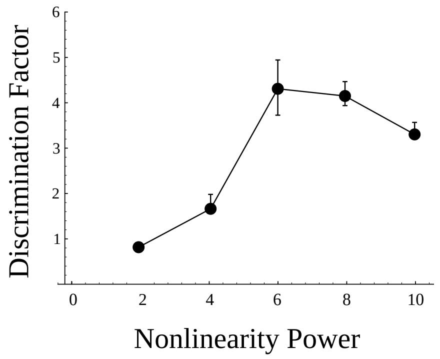



Figure 15: Variation in network performance as a function of the transfer function non-linearity in neurons of the first layer.

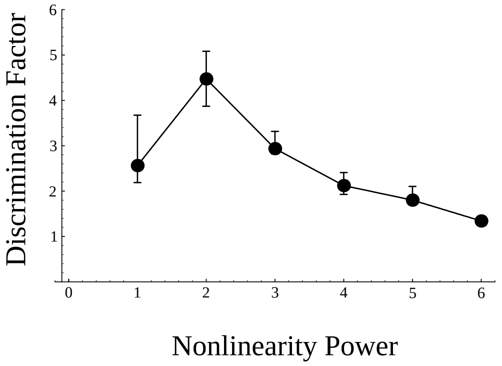



Figure 16: Variation in network performance as a function of the transfer function non-linearity in neurons of the second to fourth layers.

of the fourth layer being displayed at each parameter value. Since these are not repeated measures of the same quantity, the *median* value of the ranked cells, with error bars set at the upper and lower quartile values for the sixteen cells is displayed rather than a mean with standard error bars.

The first parameter to be tested was the non-linearity of the activation function for layer one neurons. The results are plotted in figure 15, and reveal a peak around a non-linearity value of 6. This value is surprisingly high, but presumably allows the first layer to provide a relatively clean, sparse output for unique features within each stimulus at each location, which subsequent layers can associate together across stimulus sweep sequences. Ultimately, there has to be a trade-off between increased firing sparseness and some other constraint on network performance, relating to the sparseness of inter-layer connectivity. If only a very few neurons in layer one fire, then only few cells in layer two will receive any input at all. This trade-off is reflected for non-linearity values greater than 6, for which a tailing off in overall network performance is observed. In practice, the required sparseness could also be achieved using a non-zero threshold, which might well reduce the optimal value of non-linearity found here.

Since the number of inputs to the first layer differ from the number of inputs to neurons in layers 2 to 4, it is not obvious that the degree of non-linearity of layer one cells will reflect an optimal value for later layers. This prompted optimisation of the non-linearity for neurons in layers two to four, the results for which appear in figure 16. The much lower value of non-linearity found to be optimal in these experiments for these layers suggests that the trade-off between signal clarity and excessive sparseness has shifted in favour of keeping relatively many neurons active. This is probably due to two factors. The first is that the first layer has already served to help orthogonalise the stimuli, which reduces the need for strong competition between neurons, and secondly, the relatively sparse inter-layer connectivity present in the network requires many neurons to be active

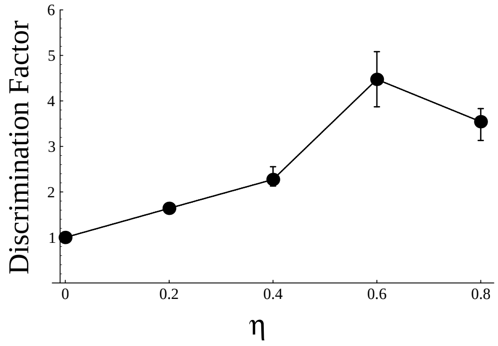



Figure 17: Variation in network performance as a function of the trace rule parameter $\eta$ in neurons of layers two to four - over nine locations.

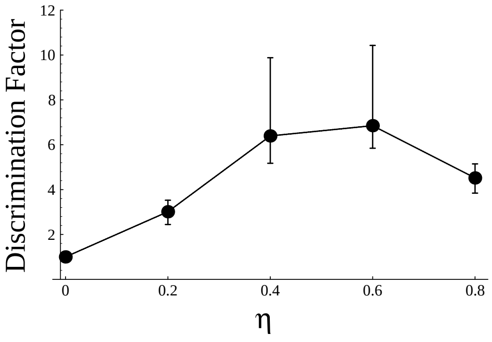



Figure 18: Variation in network performance as a function of the trace rule parameter $\eta$ in neurons of layers two to four - over five presentation locations.

in each layer to ensure that information about each input flows successfully through the hierarchy.

The next series of tests concern the effective length of the trace - controlled by the parameter $\eta$. Figure 17 displays the effect of varying the value of $\eta$ for the nine standard presentation locations. The optimal value of $\eta$ might conceivably change with the alteration of the number of training locations, and indeed one might predict that it would be smaller if the number of presentation locations were reduced. To confirm this network performance was also measured for presentation sweeps over only five locations. Figure 18 shows the results of this experiment, which confirm the expected shift in the general profile of the curve towards shorter time constant values. Of course, the optimal value of $\eta$ derived is in effect a compromise between optimal values for the three layers in which the trace operates. Since neurons in each layer have different effective receptive field sizes, one would expect each layer's neurons to be exposed to different portions of the full sweep of a particular stimulus. This would in turn suggest that the optimal value of $\eta$ will grow through the layers. However, no attempt was made to optimise these values independently, so as to be seen to keep the amount of parameter 'tweaking' in the network to a minimum.

The experiments described in this section represent a useful test case for performance of the network, and based upon this work it is now possible to turn to other important issues, such as how the performance of the network is affected by increasing the number of stimuli and whether other transformations can be learnt.

## Faces as stimuli part I: Translation

One of the consequences of the relatively small number of stimuli used in experiment 1, is that they are often encoded uniquely in the final layer, that is local encoding is used,

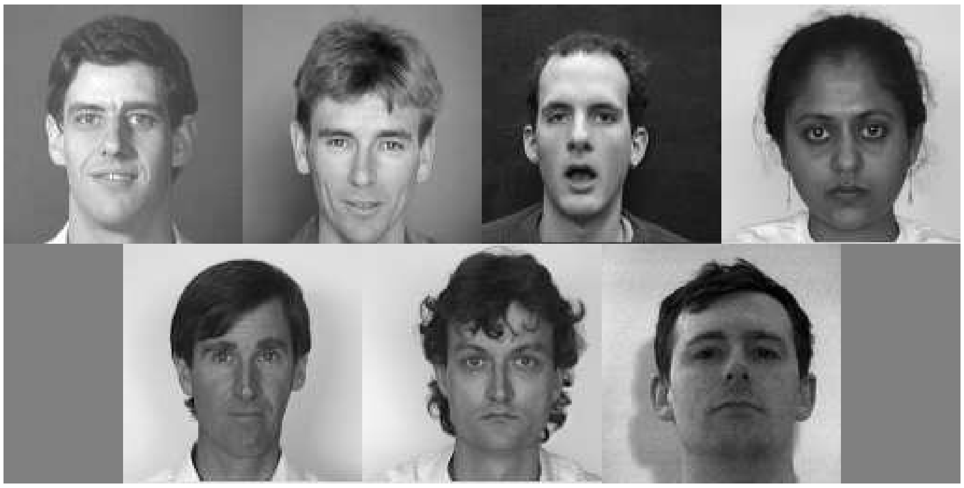


Figure 19: Seven faces used as stimuli in the face translation experiment.

with for example a 'T' detector or 'L' detector. Whilst this local or 'Grandmother Cell' type encoding is efficient with only a few stimuli, it is by no means what is observed in the tuning of IT neurons. Instead, face cells typically respond to several faces in a subset of all faces - as described in the introduction. The aim of this latest set of experiments was therefore to start to address the important issues of how the network operates when invariant representations must be learned for a larger number of stimuli, and how the neuronal encoding changes. The experiment also starts to address the issue of whether the network can learn when much more complicated, real biological stimuli (faces) are used.

The set of face images used appear in figure 19. In practice, to equalize luminance the dc component of the images was removed. In addition, so as to minimize the effect of cast shadows, an oval Hamming window was applied to the face image which also served to remove any hard edges of the image relative to the plain background upon which they were set.

The results of training in the translation invariance paradigm with 7 faces each in 9 locations are shown in figures 20, 21 and 22. The network produces neurons with high discrimination factors, and this only occurs if it is trained with the trace rule. Comparison of performance with the untrained net and a Hebb trained net in figure 22 reveals a slight change from the previous experiment, with the results for the random net having slipped below the Hebb results - indicating the greater level of complexity of the present problem. However, the most significant difference from the previous simulations is that instead of local encoding, with neurons in layer 4 responding only to one face independently of location, a more distributed representation was found, as illustrated in the examples of layer 4 neurons shown in figure 20.

The production of a distributed representation raises the question of how well the network has solved the translation invariance problem. In the previous experiment, using characters as stimuli, simple inspection of the response profiles of neurons confirmed that

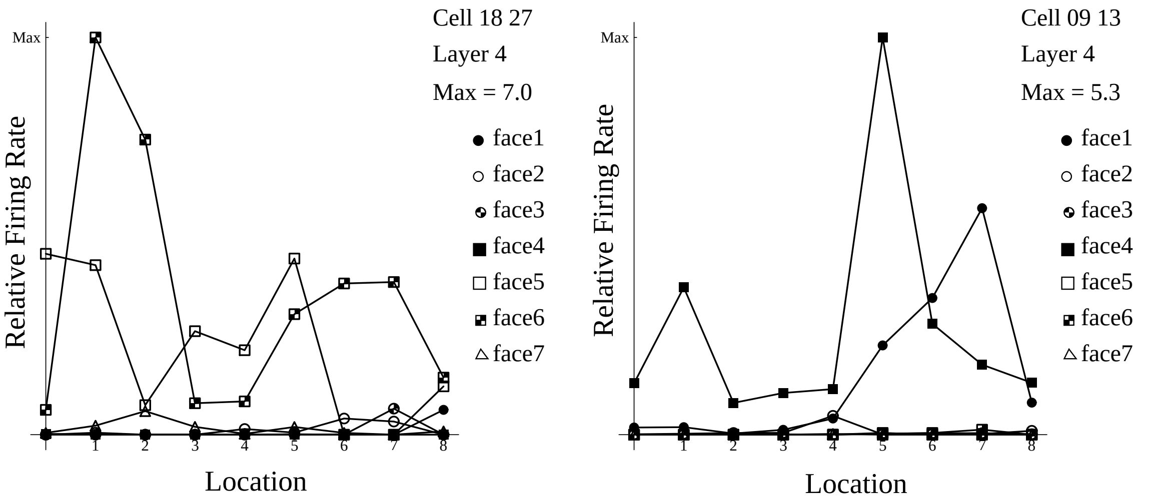



Figure 20: Response profiles for two neurons in the fourth layer - discrimination factors 2.64 and 2.10.

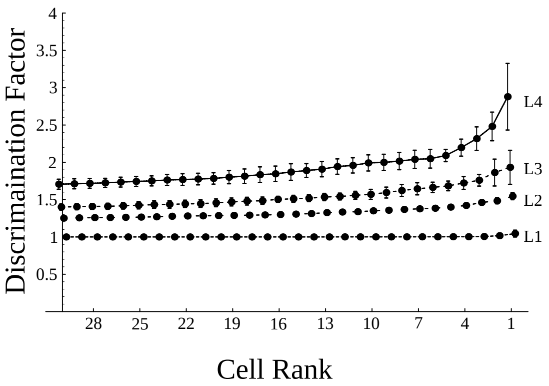



Figure 21: Variation in network performance for the top 30 most highly discriminating cells through the four layers of the network, averaged over five runs of the network.

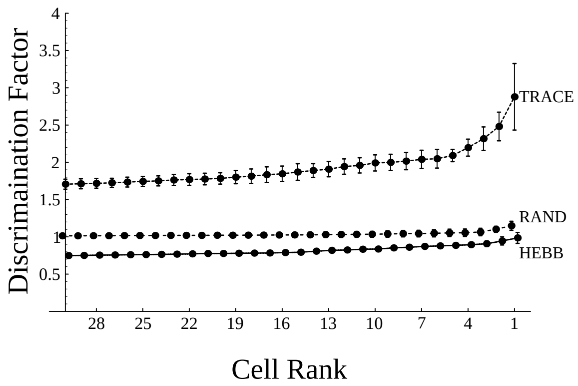



Figure 22: Variation in network performance for the top 30 most highly discriminating cells in the fourth layer for the three training regimes, averaged over five runs of the network.

the information about stimulus identity was present in an invariant form. However, in the present case we now have a distributed code in which each cell, although largely invariant for two or so preferred stimuli, is not in general selective for a single stimulus. As a consequence, the fourth layer as a whole may or may not retain the information required to dissociate the stimuli to which it responds. It is only appropriate to conclude that there is a distributed code for stimuli which is truly invariant if it can be successfully decoded, and each stimulus be discriminated from the others, irrespective of its original position on the retina. In order to decode the information in the firing of the fourth layer neurons, a fifth layer was added to the net which fully sampled the fourth layer cells. This layer was in turn trained in a supervised manner using gradient descent. Note that the fifth layer is intended purely as a tool for analysis and for decoding the representation found in the final layer of the network, and should not be interpreted as being part of the VisNet model of the primate visual system. If information about the identity of individual stimuli had been irrevocably lost by the representation built in layer four, due to the consistent pairing of stimuli by the neurons, then the new decoding network should not be able to extract information about individual stimulus identity.

Figure 23 shows the classification performance of the 5th layer for nets trained with the Hebb and trace rules as well as for the untrained net. The high performance for the Hebb trained net 84%, and the untrained net 92%, show the power of the gradient descent learning and should serve to put the perfect 100% performance on the trace rule trained net into perspective. However, the performance is distinct and the main attempt of doing this analysis, to confirm that information about stimulus identity is in fact still retrievable in the fourth layer, has been fulfilled.

Further evidence for the quality of this representation was obtained by measuring how well the cells responded to each stimulus by accumulating the number of neurons responding most strongly to each stimulus, multiplied by the associated discrimination factor - as was done in the previous experiment. The result is this time plotted, in figure 24, and shows that each of the stimuli had a high discrimination total. This evidence, taken together with the evidence shown in figures 21 and 22, indicates that a representation with some invariant properties of each stimulus is established in layer four, provided that the trace learning rule is used.

## Faces as stimuli part II: Rotation

Now that the network has been shown to be able to operate usefully with a more difficult translation invariance problem, we now address the question of whether the network can solve other types of transform invariance, as we had intended. The next experiment addresses this question, by training the network on the problem of 3D stimulus rotation, which produces non-isomorphic transforms, to determine whether the network can build a view-invariant categorisation of the stimuli. The trace rule learning paradigm should, in conjunction with the architecture we describe here, prove capable of learning any of the transforms tolerated by IT neurons, so long as each stimulus is presented in short sequences during which the transformation occurs and can be learned. This experiment continues with the use of faces but now presents them centrally in the retina in a sequence of views of a rotating face. The images used are shown in figure 25. The faces were again

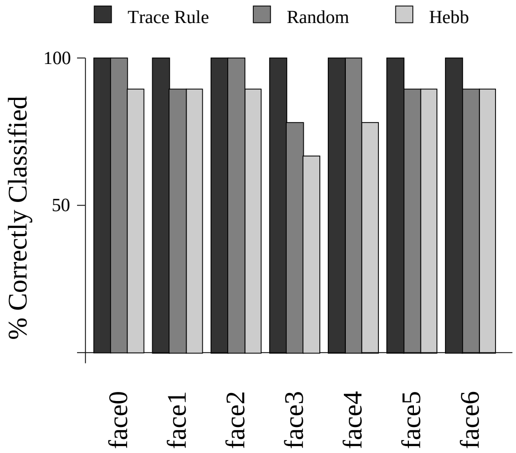



Figure 23: Stimulus classification achieved for each of the three training regimes by supervised training of a fifth layer on the outputs of the fourth layer.

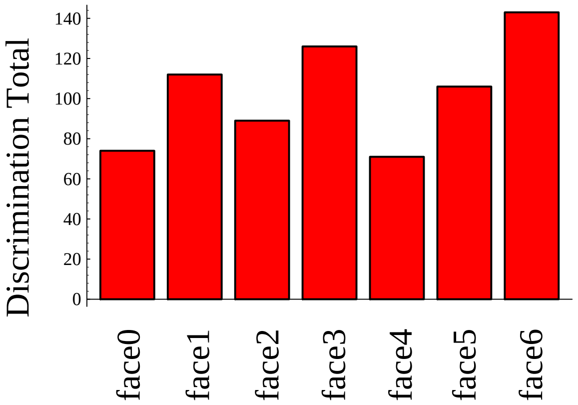



Figure 24: Discrimination totals for each stimulus across all neurons in the fourth layer trained with the trace rule.

smoothed at the edges to erase the harsh image boundaries, and the dc term was removed. During the 800 epochs of learning, each stimulus was chosen at random, and a sequence of preset views of it was shown, sweeping the face either to the left or right.

Although the actual number of images being presented is smaller, some 21 views in all, there is good reason to think that this problem may be harder to solve than the previous translation experiments. This is simply due to the fact that all 21 views exactly overlap with one another, whereas a maximum of seven images exactly overlapped in the previous experiment. Partly because of this fact, and also partly so as to use the capacity of the network more fully, the images presented were twice as large as those used in the translation experiments. This also permitted the net to discern finer feature detail in the individual faces.

Since this experiment employs a totally new training paradigm the usual series of optimal parameter testing was performed on the neural activation function and the size of the trace parameter $\eta$. The optimal non-linearity value remained unchanged but the optimal value of $\eta$ was slightly lower than for the previous two experiments. This shift in $\eta$ was almost certainly a function of the shorter sweep lengths, which have fallen from nine to seven steps. The fact that the $\eta$ value was even lower with only five presentations per sweep, confirms the general relationship between the optimal value of $\eta$ and the number of steps involved in the presentation of any single stimulus in its various transformed forms.

The net was indeed able to solve the invariance problem, with examples of invariant layer four neuron response profiles appearing in figure 26. One major difference from the results of previous experiments was that some cells in the first layer showed limited

Figure 25: Three faces in seven different views used as stimuli in Experiment II.

tolerance to shifts in viewing angle. This is to be expected since slightly rotated views of a face will share many of the same basic features in the same location, which results in the observed generalisation. Although, true generalisation across all views was not achieved until higher layers, the contribution of local generalisation provided by the cells in layer one for this problem meant that cells in layer three already exhibit full rotational invariance. This result is also in part due to the fact that the images, though twice as large as in the previous experiment, did not extend as far out across the retina as in the translation invariance experiment, allowing convergence of the information relevant to solving the problem to occur earlier in the hierarchy. Although view invariance was partially solved by layer three of the network, a further improvement was found with layer four neurons, where there are more cells with high discrimination factors and flatter response profiles than those observed in layer three.

Figure 27 confirms the improvement in invariant stimulus representation found through the layers and that layer four provides a considerable improvement in performance over all layers. This despite the emergence of invariant cells in the previous layer. Figure 28 shows the Hebb trained and untrained nets performing equally poorly, whilst the trace trained net shows good invariance across the entire 30 cells selected. Since the number of stimuli has been reduced, training has again established a local and sparse representation of the stimuli.

# 3 Discussion

The most significant conclusion to draw from these experiments is that the network can solve invariance problems, and more to the point, only by virtue of inclusion of the trace rule. The trace rule is both local and hence biologically plausible, in that the signals required to alter the synaptic strength during learning are the presynaptic firing and the postsynaptic activation, both available locally at the synapse. The use of such

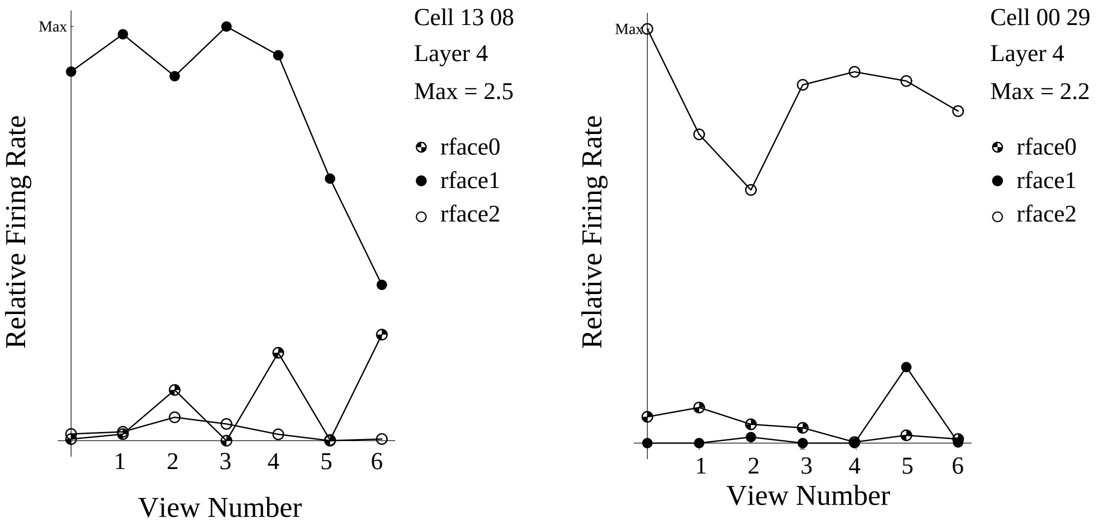



Figure 26: Response profiles for cells in the last two layers of the network - discrimination factors 11.12 and 12.40.

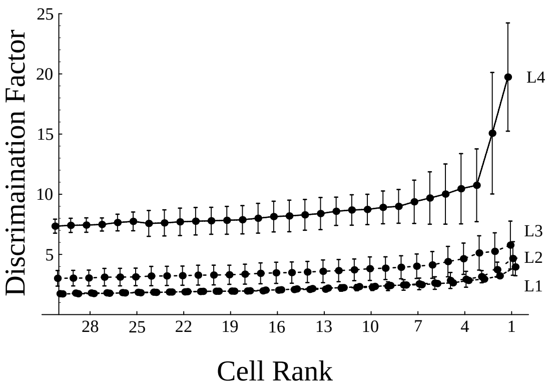



Figure 27: Variation in network performance for the top 30 most highly discriminating cells through the four layers of the network, averaged over five runs of the network.

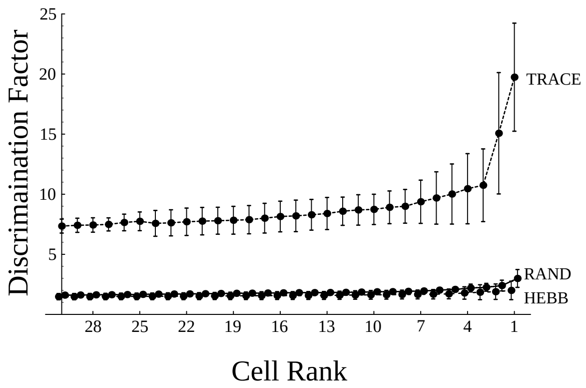



Figure 28: Variation in network performance for the top 30 most highly discriminating cells in the fourth layer for the three training regimes, averaged over five runs of the network.

a learning rule sets this proposal apart from most other proposals for how invariant representations might be formed. The system also operates by self-organizing competitive learning, which is also biologically plausible, in that the learning can be driven by the actual inputs received with no external teacher needed, and in that lateral inhibition, which implements competition, is a well known property of cortical architecture. Other models have typically combined various less attractive elements such as supervised or non-local learning (Poggio & Edelman, 1990; Fukushima, 1980; Mel, 1996), extremely idealised or simplified stimuli (Földiák, 1991; Hinton, 1981), prohibitive object by object matching processes (Olhausen *et al.*, 1993; Buhmann *et al.*, 1990) or non-localised connectivity (Hummel & Biederman, 1992). In fairness many of these models have advantages over the model described here, such as Olhausen's model which more explicitly tackles the problem of locating and attending to objects throughout the visual field. Our model only really deals with recognition within the high acuity centre of the visual field and would require some other mechanism for locating and fixating objects. As described here, our model has also only been exposed to relatively few stimuli. Mel and Fukushima's models for example, have been successfully trained on much larger data sets. However, although not reported here, our net has been successfully trained to 96% accuracy on a set of 100 hand-written digits (Wallis, 1994).

One further element which our model lacks is an active solution for the feature binding problem, as addressed by the theories and models of von der Malsburg (1981; 1986) for example. If a cell is trained to respond to the appearance of a set of features in, say, a translation invariant way, then the neuron may well respond to novel rearrangements of these features. This is not in fact the case for real neurons responsive to faces, in that neurons often reduce their response to faces in which the features appear jumbled up (Perrett *et al.*, 1992), nor is it true for other more abstract features (Tanaka *et al.*, 1991). Successfully determining the features present in an invariant manner, whilst still retaining spatial configuration leads to the binding problem. Models which throw away spatial information so as to achieve translation invariance will run into the problem of 'recognising' rearrangements of the features triggering recognition. This is certainly true for models which attempt to learn invariance in one stage (Mel, 1996; Cavanagh, 1978). In his recent paper, Mel (1996) records the recognition of cells to jumbled features at over 95% of the recognition rate to trained configurations.

With this in mind we have to emphasise an important feature of the model proposed here, namely the convergent processing hierarchy. The purpose of this hierarchy is to gradually build representations of ever increasing complexity within the self-organizing competitive net layers. These nets are intended to produce neurons which respond to combinations of inputs themselves forming the effective stimuli for later neurons. By responding to local combinations of neurons co-active in the previous layer, arbitrary spatial arrangements of the same features should fail to activate the same neuron. This should then reduce the chance of finding the trigger features supporting recognition in new arrangements of the features, revealing a partial solution to the feature binding problem. Indeed, it is an important part of this suggestion that some local spatial information would be inherent in the features which were being combined - far more of the form described by Tanaka et. al. (1991) in their recordings from IT neurons.

Following the feature combination argument, a view-dependent representation of objects suitable for view-dependent processes such as behavioural responses to face expression and gesture would only be available after several stages of processing. View-independent representations could then be formed by the same type of computation, operating to combine a limited set of views of objects. Indeed, neurons with view-independent responses are present in the visual system, and evidence suggests that they receive their inputs from view-dependent neurons in the same region (Hasselmo *et al.*, 1989; Perrett *et al.*, 1987). The plausibility of providing view-independent recognition of objects by combining a set of different views of objects has been proposed by a number of investigators (Koenderink & van Doorn, 1979; Tarr & Pinker, 1989; Bülthoff & Edelman, 1992), and the network described here reveals how such a representation might be set up, without recourse to gradient decent type learning (Poggio & Edelman, 1990; Logothetis *et al.*, 1994). This solution to 'object-based' representations is very different from that traditionally proposed for artificial vision systems, in which the coordinates in 3D-space of objects are stored in a database, and general-purpose algorithms operate on these to perform transforms such as translation, rotation, and scale change in 3D space (e.g. Marr, (1982)). In the present, much more limited but more biologically plausible scheme, the representation would be suitable for recognition of an object, and for linking associative memories to objects as described in more detail by Rolls (1995)). The solution to invariant recognition proposed here would certainly need a large number of neurons, but this is simply consistent with the fact that perhaps half of the cortex of non-human primates is devoted to vision. This in turn leads to one aim of future work, namely to discover the capacity of the system, in terms of the number of objects or stimuli about which it could learn. The hypothesis is that invariant properties common to many objects can be learned in early layers of the type of network described here, with information about particular objects only being represented in the last few layers.

One aspect of the model that has not been treated in detail here is the optimal form of the trace rule, and of the parameter $\eta$ which controls the length of the trace. In the real world, objects may typically be viewed for 0.5-1 s or more, with fixation durations between saccades often 200-300 ms. A trace of up to 1 s in the real world might thus be satisfactory. In the simulations described here, optimal values of $\eta$ rose to produce a somewhat longer trace when each object was shown for nine sequential time steps than for five, as one might expect. A more detailed and quantitative approach to the optimal form of the trace rule is being pursued (Wallis, 1996). This has shown that under a variety of probabalistic stimulus presentation paradigms, the form of the trace rule used here, which weights events with exponentially decreasing strength the more distant they were in the past, is close to optimal. Of course, this discussion assumes that there is no active resetting of the visual system between the inspection of different objects. If the eye movements that accompany orientation to a new object are great, then active suppression might be caused by the complete change in the inputs reaching the visual system, and transient feedback inhibition produced by the large visual input produced by the re-orientation. Such a resetting between objects would help the operation of the model described here, but the model by no means needs this, and operates well if the trace is of a fixed duration, in general just shorter than the average time with which any stimulus is inspected.

Since the trace rule is seen to play a crucial role in the successful learning of object invariant responses it is worth considering the major tenets of the learning as they apply to the primate visual system. Firstly, it should be possible to process several different images of an object within the several hundred ms for which an object may be viewed, and secondly, that at least some learning can take place rapidly in the cortex, within this order of time. Both tenets appear to be satisfied. Several hundred ms is sufficient time for the visual stimulus to process several different images of an object (Rolls & Tovee, 1994; Thorpe & Imbert, 1989). Moreover, learning of new representations appears to occur with as little as several 1 s presentations of new stimuli, such as faces, that have never been seen before (Rolls *et al.*, 1993).

The third and final tenet of the learning used here is that there should be some means for previous cellular activity to affect learning in the visual cortex. This question which has been considered by several researchers (Rolls, 1992; Földiák, 1991; Rhodes, 1992; Montague *et al.*, 1991; O'Reilly & Johnson, 1994) and is summarised in the section on the trace rule earlier in this paper. One point not raised in that discussion however, is that a trace learning rule would be appropriate for the ventral visual system, concerned with invariant form representation, but *not* for the dorsal visual system, in which motion and location are processed (Ungeleider & Mishkin, 1982; Ungerleider & Haxby, 1994). Indeed, the importance of having a trace rule in the part of the visual system involved in invariant object recognition, and the importance of *not* having such a rule in the part of the visual system involved in processing motion and location, might be a fundamental reason for keeping these two processing streams apart.

# A Appendix: Measure of network performance

In these experiments, a neuron's response varies as a function of location (or rotation) and stimulus type. A simple measure for how invariant the response of a neuron is to shifts in the location of a stimulus is simply the average variance across location for each stimulus $s_L^2$. If the response profiles are flat this variance should be low. In addition, the distinctiveness of the response to any particular stimulus can be gauged by the average variance in response across stimuli $s_S^2$. If the responses for each stimulus are very different and allow easy discrimination between stimulus groups this variance should be high. Any variance left unaccounted for by $s_S^2$ and $s_L^2$ we can denote $s_{error}^2$. This variance acts as a measure of the reliability of the other two variances measured, and should be low if the response of the neuron is consistent across location and stimulus.

An ideal, translation invariant, highly discriminating neuron can thus be determined by seeking a high value for the ratio between $s_S^2$ and $s_L^2$, along with a generally low value of $s_{error}^2$. Snedecor (1989) describes how to combine the error and each variance measure to derive a measure of the 'relative amount of information' (RAI) for each of the two factors. The formula normalises the measured variance with regard to the error, number of stimuli and number of locations. By converting $s_S^2$ and $s_L^2$ into RAI measures, a direct quotient of the two corrected variances can now be taken:

$$\text{DiscriminationFactor} = \mathcal{F}_C \frac{(N_S - 1)s_S^2 + N_S(N_L - 1)s_{error}^2}{(N_L - 1)s_L^2 + N_L(N_S - 1)s_{error}^2}$$

| | | | |
|---|---|---|---|
| $N_S$ | Number of stimulus classes | $N_L$ | Number of examples of each stimulus |
| $s_S^2$ | Stimulus class sample variance | $s_L^2$ | Location sample variance |
| $s_{error}^2$ | Sample error | $\mathcal{F}_C$ | Correction factor |

One attractive quality of this measure is that it evaluates simply to the correction factor $\mathcal{F}_C$ in the case that a cell only responds to one stimulus in one location. The formula for $\mathcal{F}_C$ is not given since in all of the cases studied here it is only very slightly different from one, $0.975 < \mathcal{F}_C < 1$.

## Acknowledgements

The authors are grateful to Peter Földiák and two anonymous reviewers for help and advice in preparing this manuscript. Guy Wallis was supported by a SERC grant whilst conducting the experimental work described in this paper.

# References


Abbott, L.A., Rolls, E.T., & Tovee, M.J. 1996. Representational capacity of face coding in monkeys. *Cerebral Cortex*, **In press**.

Anderson, J.A., & Rosenfeld, E. (eds). 1988. *Neurocomputing: Foundations of Research.* Cambridge: MIT Press.

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.

Bennett, A. 1990. Large competitive networks. *Network*, **1**, 449–462.

Brown, T.H., Kariss, E.W., & Keenan, C.L. 1990. Hebbian synapses: Biological mechanisms and algorithms. *Annual Review of Neuroscience*, **13**, 475–511.

Buhmann, J., Lades, M., & von der Malsburg, C. 1990. Size and distortion invariant object recogntion by hierarchical graph matching. *Pages 411–416 of: International Joint Conference on Neural Networks.* New York: IEEE.

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. *Pages 121–159 of:* Kosko, B. (ed), *Neural Networks for Signal Processing.* Englewood Cliffs, New Jersey: Prentice Hall.

Bülthoff, H., & Edelman, S. 1992. Psychophysical support for a two–dimensional view interpolation theory of object recognition. *Pages 60–64 of: Proceedings of the National Academy of Science, USA*, vol. 92.

Cavanagh, P. 1978. Size and location invariance in the visual system. *Perception*, **7**, 167–177.

Chakravarty, I. 1979. A generalized line and junction labeling scheme with applications to scene analysis. *IEEE Transactions PAMI*, April, 202–205.

Desimone, R. 1991. Face–selective cells in the temporal cortex of monkeys. *Journal of Cognitive Neuroscience*, **3**, 1–8.

Földiák, P. 1991. Learning invariance from transformation sequences. *Neural Computation*, **3**, 194–200.

Földiák, P. 1992. *Models of Sensory Coding.* Tech. rept. CUED/F–INFENG/TR 91. University of Cambridge, Department of Engineering.

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.

Hasselmo, M.E., Rolls, E.T., Baylis, G.C., & Nalwa, V. 1989. Object–centred encoding by face–selective neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research*, **75**, 417–429.

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

Hinton, G.E. 1981. A parallel computation that assigns canonical object based frames of reference. *In: Proceedings of the 9th International Joint Conference on Artificial Intelligence*. Reviewed in (Rumelhart *et al.*, 1986).

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

Koenderink, J.J., & van Doorn, A.J. 1979. The internal respresentation of solid shape with respect to vision. *Biological Cybernetics*, **32**, 211–216.

Linsker, E. 1986. From basic network principles to neural architecture. *Proceedings of the National Academy of Sciences, USA*, **83**, 7508–7512, 8390–8394, 8779–8783.

Logothetis, N.K., Pauls, J., Bülthoff, H.H., & Poggio, T. 1994. View–dependent object recognition by monkeys. *Current Biology*, **4**(5), 401–414.

Marr, D.C. 1982. *Vision*. San Francisco: W.H. Freeman and Co.

Mel, B.W. 1996. *SEEMORE: Combining color, shape, and texture histogramming in a neurally-inspired approach to visual object recognition*. Unpublished Manuscript.

Miyashita, Y., & Chang, H.S. 1988. Neuronal correlate of pictorial short–term memory in the primate temporal cortex. *Nature*, **331**, 68–70.

Montague, R., Gally, J., & Edelman, G. 1991. Spatial signalling in the development and function of neural connections. *Cerebral Cortex*, **1**, 199–220.

Nass, M.M., & Cooper, L.N. 1975. A theory for the development of feature detecting cells in visual cortex. *Biological Cybernetics*, **19**, 1–18.

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

Olhausen, 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.

O'Reilly, R., & Johnson, M. 1994. Object recognition and sensitive periods: A computational analysis of visual imprinting. *Neural Computation*, **6**, 357–389.

Perrett, D.I., Mistlin, A.J., & Chitty, A.J. 1987. Visual cells responsive to faces. *Trends in Neurosciences*, **10**, 358–364.

Perrett, D.I., Hietanen, J.K., Oram, M.W., & Benson, P.J. 1992. Organisation and functions of cells responsive to faces in the temporal cortex. *Philosophical Transactions of the Royal Society, London [B]*, **335**, 23–30.

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

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.

Rolls, E.T. 1992. Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical areas. *Philosophical Transactions of the Royal Society, London [B]*, **335**, 11–21.

Rolls, E.T. 1995. Learning mechanisms in the temporal lobe visual cortex. *Behavioural Brain Research*, **66**, 177–185.

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

Rolls, E.T., Tovee, M.J., & Ramachandran, V.S. 1993. Visual learning reflected in the responses of neurons in the temporal visual cortex of the macaque. *Society for Neuroscience Abstracts*, **19**, 27.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. 1986. Learning internal representations by error propagation. *Chap. 8 of:* Rumelhart, D.E., & McClelland, J.L. (eds), *Parallel Distributed Processing*, vol. 1: Foundations. Cambridge, Massachusetts: MIT Press.

Snedecor, G.W., & Cochran, W.G. 1989. *Statistical Methods*. 8 edn. Ames, Iowa: Iowa State University Press.

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

Tanaka, K., Saito, H., Fukada, Y., & Moriya, M. 1991. Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology*, **66**, 170–189.

Tarr, M.J., & Pinker, S. 1989. Mental rotation and orientation–dependence in shape recognition. *Cognitive Psycholgy*, **21**, 233–282.

Thorpe, S.J., & Imbert, M. 1989. Biological constraints on connectionist models. *Pages 63–92 of:* Pfeifer, R., Schreter, Z., & Fogelman-Soulie, F. (eds), *Connectionism in Perspective*. London: John Wiley and Sons.

Tovee, M.J., Rolls, E.T., & Azzopardi, P. 1994. Translation invariance in the response to faces of single neurons in the temporal visual cortical areas of the alert macaque. *Journal of Neurophysiology*, **72**, 1049–1060.

Ungeleider, L.G., & Mishkin, M. 1982. Two cortical visual systems. *Pages 549–586 of:* Ingle, D.J., Goodale, M.A., & Mansfield, R.J.W. (eds), *Analysis of Visual Behaviour*. Cambridge, Massachusetts, USA: MIT press.

Ungerleider, L.G., & Haxby, J.V. 1994. 'What' and 'Where' in the human brain. *Current Opinion in Neurobiology*, **4**, 157–165.

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.

von der Malsburg, C. 1973. Self–organization of orientation sensitive cells in the striate cortex. *Kybernetik*, **14**, 85–100. Reprinted in (Anderson & Rosenfeld, 1988).

von der Malsburg, C. 1981. *The correlation theory of brain function*. Tech. rept. 81-2. Department of Neurobiology, Max-Planck-Institute for Biophysical Chemistry, Göttingen.

von der Malsburg, C., & Schneider, W. 1986. A neural cocktail-party processor. *Biological Cybernetics*, **54**, 29–40.

Wallis, G. 1994. *Neural Mechanisms Underlying Processing in the Visual Areas of the Occipital and Temporal Lobes*. Ph.D. thesis, Department of Experimental Psychology, Oxford Univeristy. www: ftp://ftp.mpik-tueb.mpg.de/pub/guy/all.ps.Z.

Wallis, G. 1996. Optimal unsupervised learning in invariant object recognition. *Submitted for review*. www: ftp://ftp.mpik-tueb.mpg.de/pub/guy/nc.ps.Z.

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