PERGAMON

Neural Networks 13 (2000) 305–315

Neural
Networks

www.elsevier.com/locate/neunet

Contributed article

Position invariant recognition in the visual system with cluttered environments

S.M. Stringer, E.T. Rolls*

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

Received 29 September 1999; accepted 3 February 2000

Abstract

The effects of cluttered environments are investigated on the performance of a hierarchical multilayer model of invariant object recognition in the visual system (VisNet) that employs learning rules that utilise a trace of previous neural activity. This class of model relies on the spatio-temporal statistics of natural visual inputs to be able to associate together different exemplars of the same stimulus or object which will tend to occur in temporal proximity. In this paper the different exemplars of a stimulus are the same stimulus in different positions. First it is shown that if the stimuli have been learned previously against a plain background, then the stimuli can be correctly recognised even in environments with cluttered (e.g. natural) backgrounds which form complex scenes. Second it is shown that the functional architecture has difficulty in learning new objects if they are presented against cluttered backgrounds. It is suggested that processes such as the use of a high-resolution fovea, or attention, may be particularly useful in suppressing the effects of background noise and in segmenting objects from their background when new objects need to be learned. However, it is shown third that this problem may be ameliorated by the prior existence of stimulus tuned feature detecting neurons in the early layers of the VisNet, and that these feature detecting neurons may be set up through previous exposure to the relevant class of objects. Fourth we extend these results to partially occluded objects, showing that (in contrast with many artificial vision systems) correct recognition in this class of architecture can occur if the objects have been learned previously without occlusion. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Segmentation; Invariance; Attention; Object recognition

1. Introduction

1.1. Background

In this paper we investigate the effects of cluttered environments on the performance of a model of transform (e.g. position, size and view) invariant object recognition in the visual system (VisNet) proposed by Rolls (1992, 1994, 1995, 2000) that employs learning rules that utilise a trace of previous neural activity. The model architecture is based on the following: (i) A series of hierarchical competitive networks with local graded inhibition. (ii) Convergent connections to each neuron from a topologically corresponding region of the preceding layer, leading to an increase in the receptive field size of neurons through the visual processing areas. (iii) Synaptic plasticity based on a modified Hebb-like learning rule with a temporal trace of each neuron's previous activity. These hypotheses were incorporated into a simulation, VisNet, which was shown to be capable of producing stimulus-selective but translation and view invariant representations (Wallis & Rolls, 1997). Models with hierarchically organised competitive networks designed to study neurally plausible ways of forming invariant representations of stimuli have been studied by a number of investigators (Fukushima, 1980; Poggio & Edelman, 1990), but VisNet differs from other models in that it relies on the spatio-temporal statistics of natural visual inputs to be able to associate together different transforms of stimuli which will tend to occur in temporal proximity. Further work with VisNet is presented in Elliffe, Rolls and Stringer (2000), Rolls and Milward (2000) and Rolls and Stringer (2000). However, so far, all investigations with VisNet have involved the presentation of stimuli against a blank background, so that no segmentation of the object from its background is needed. The aim of the investigations described here is to compare such results with simulations performed in cluttered environments. Although there has been much work involving object recognition in cluttered environments with artificial vision systems, many such systems typically rely on some form of search and

* Corresponding author. Tel.: +44-1865-271348; fax: +44-1865-310447.

E-mail address: edmund.rolls@psy.ox.ac.uk (E.T. Rolls).

PII: S0893-6080(00)00017-4

template matching procedure (see Ullman (1996) for a general review). Such problems may involve the object appearing against a cluttered background or partial occlusion of the object. However, biological nervous systems operate in quite a different manner to those artificial vision systems that rely on search and template matching, and the way in which biological systems cope with cluttered environments is likely to be quite different also.

One of the factors that will influence the performance of the type of architecture considered here, hierarchically organised series of competitive networks, which form one class of approaches to biologically relevant networks for invariant object recognition (Fukushima, 1980; Poggio & Edelman, 1990; Rolls, 1992; Rolls & Treves, 1998; Wallis & Rolls, 1997), is how lateral inhibition and competition are managed within a layer. Even if an object is not obscured, the effect of a cluttered background will be to fire additional neurons, which will in turn to some extent compete with and inhibit those neurons that are specifically tuned to respond to the desired object. Moreover, where the clutter is adjacent to part of the object, the feature analysing neurons activated against a blank background may be different from those activated against a cluttered background, if there is no explicit segmentation process. In this paper we examine the performance of one network of this type, VisNet, taken as an exemplar of the class, when presented with stimuli to be identified invariantly even when presented in a cluttered background. From the experiments we are able to make some proposals about the operation of real nervous systems.

In Section 2 we show that whereas recognition of objects learned previously against a blank background is hardly affected by the presence of background noise, the ability to learn position invariant responses to new objects when presented against cluttered backgrounds is greatly reduced. This suggests that some form of attentional mechanism may be required during learning to highlight the current stimulus being attended to and suppress the effects of background noise. However, we also demonstrate that this problem may be ameliorated by the prior existence of stimulus tuned feature detecting neurons in the early layers of the VisNet, and that these feature detecting neurons may be set up through previous exposure to the relevant class of objects. Such feature detecting neurons may then help to suppress the effects of background clutter when the visual system is exposed to new members of that class. Hence, the findings predict that in real world cluttered environments, attention is more likely to be required for learning than for recognition; and that learning of new objects is facilitated in cluttered backgrounds if feature analysers useful for the new objects have been formed by previous exposure to other objects with similar features.

In Section 3 we examine the recognition of partially occluded stimuli. Many artificial vision systems that perform object recognition typically search for specific markers in stimuli, and hence their performance may become fragile if key parts of a stimulus are occluded. However, in contrast we demonstrate that the biologically inspired model discussed in this paper can continue to offer robust performance with this kind of problem, and that the model is able to correctly identify stimuli with considerable flexibility about what part of a stimulus is visible.

1.2. The VisNet model

In this section we give an overview of the VisNet model; full details are provided by Rolls and Milward (2000) and Wallis and Rolls (1997). In particular, the simulations performed in this paper use the latest version of the VisNet model (VisNet2) with the same parameter values as given in Rolls and Milward (2000). The model consists of a feed-forward[1] hierarchical series of 4 layers of competitive networks (with 1024 neurons per layer), corresponding in the primate visual system to V2, V4, the posterior inferior temporal cortex, and the anterior inferior temporal cortex, as shown in Fig. 1. The V2 layer of the model receives its inputs from an input layer which provides a representation comparable to that found in V1. The forward connections to individual cells are derived from a topologically corresponding region of the preceding layer, using a Gaussian distribution of connection probabilities. Within each layer competition is graded rather than winner-take-all, and is implemented in two stages. First, to implement lateral inhibition the activation of neurons within a layer is convolved with a local spatial filter which operates over several pixels. Next, contrast enhancement is applied by means of a sigmoid activation function where the sigmoid threshold is adjusted to control the sparseness of the firing rates to values that are approximately 0.01 for the first two layers, and 0.1 for layers 3 and 4 (see for details Rolls and Milward (2000)).

The mechanism for transform invariant object recognition proposed by Földiák (1991) and Rolls (1992) relies on the spatio-temporal statistics of natural visual input. In particular, in the real world different views of the same object are likely to occur in temporal proximity to each other. Then if synaptic learning rules are utilised that encourage neurons to respond invariantly to temporally proximal input patterns, such neurons should learn to respond invariantly to different views of individual stimuli. The original trace learning rule used in the simulations of (Wallis & Rolls, 1997) took the form

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

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

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

and we have the following definitions: x_j, jth input to the

[1] Backprojections are not included in the current implementation of VisNet, because there is evidence that they are not necessary for rapid object identification (Rolls & Treves, 1998).

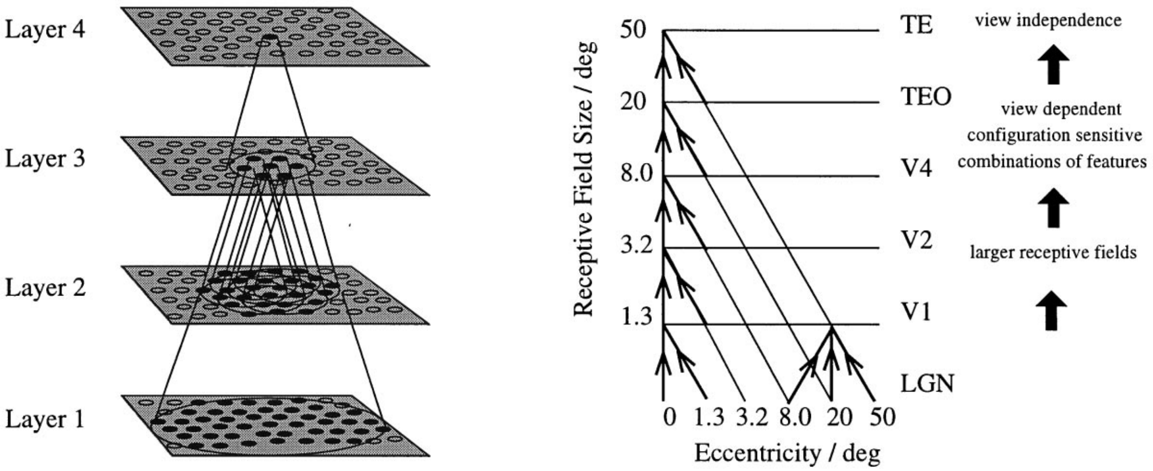

Fig. 1. Left: Stylised image of the VisNet four layer network. Convergence through the network is designed to provide 4th layer neurons with information from across the entire input retina. Right: Convergence in the visual system—adapted from Rolls (1992). V1, visual cortex area V1; TEO, posterior inferior temporal cortex; and TE, inferior temporal cortex (IT).

neuron; y, output from the neuron; $\bar{y}^\tau$, trace value of the output of the neuron at time step τ; α, learning rate. Annealed between unity and zero; w_j, synaptic weight between jth input and the neuron; η, 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 set to 0.8. However, for $\eta = 0$ Eq. (1) becomes the standard Hebb rule

$$\Delta w_j = \alpha y^\tau x_j^\tau. \tag{3}$$

Recently, however, it has been demonstrated (Rolls & Milward, 2000) that a modified Hebbian rule which incorporates a trace of activity calculated from the preceding presentations but not the current time step can offer substantially improved performance over the standard trace rules described in Földiák (1991) and Wallis and Rolls (1997). This rule takes the form

$$\Delta w_j = \alpha \bar{y}^{\tau-1} x_j^\tau. \tag{4}$$

In VisNet simulations discussed later in this paper, learning rule (4) is used to develop transform invariant neurons.

1.3. Training and test procedure

The images used for training and testing VisNet in this paper are specially constructed for the cluttered environment problems described in Section 1.1. There are 7 face stimuli approximately 64 pixels in height constructed without backgrounds as shown in Fig. 2. In addition there are three possible backgrounds: a blank background (greyscale 127, where the range is 0–255), and 2 cluttered backgrounds as shown in Fig. 3 which are 128 × 128 pixels in size. Each image presented to VisNet's 128 × 128 input retina is then composed of a single face stimulus positioned at one of nine locations on either a blank or cluttered background. The cluttered background was intended to be like the background against which an object might be viewed in a natural scene. If a background is used in an experiment described here, the same background is always used, and it is always in the same position, with stimuli moved to different positions on it. The nine stimulus locations are arranged in a square grid across the background, where the grid spacings are 32 pixels horizontally or vertically. Before images are presented to VisNet's input layer they are preprocessed by a set of input filters which accord with the general tuning profiles of simple cells in V1 (Hawken & Parker, 1987); full details are given in Rolls and Milward (2000). To train the network a sequence of images is presented to VisNet's retina that corresponds to a single stimulus occurring in a randomised sequence of the nine locations across a background. 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 1 epoch of training. In this manner the network is trained one layer at a time starting with layer 1 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. For each cell the single cell

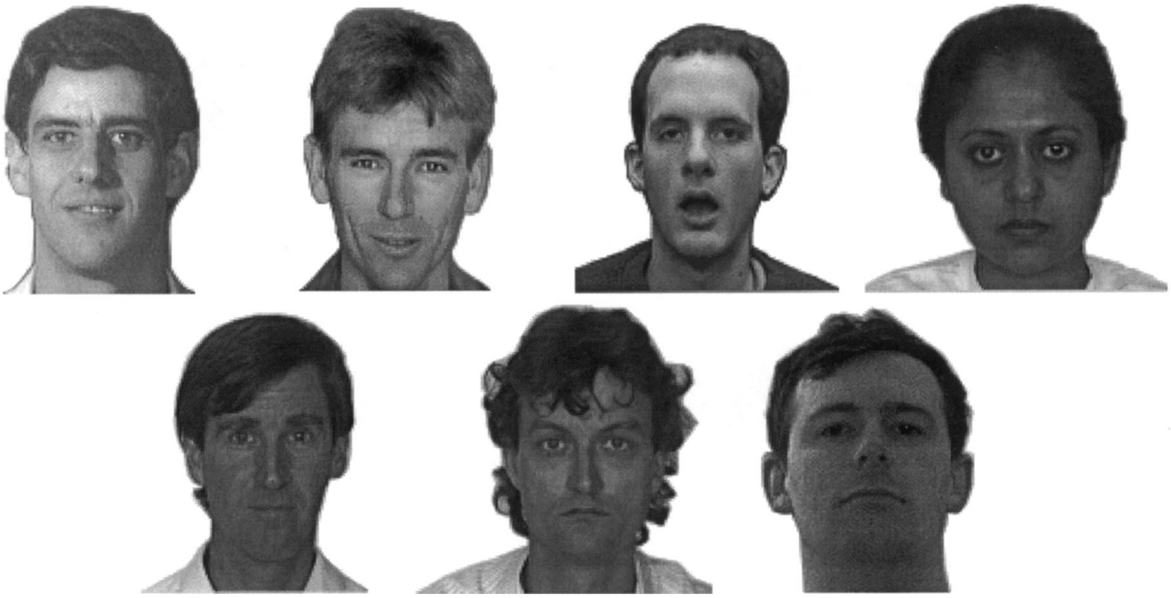

Fig. 2. Face stimuli used in VisNet simulations: faces are ordered from face 1 (top left) to face 7 (bottom right).

information measure used was the maximum amount of information a cell conveyed about any one stimulus. This is computed using the following formula with details given by Rolls, Treves, Tovee and Panzeri (1997) and Rolls and Milward (2000). The stimulus-specific information $I(s, R)$ is the amount of information the set of responses R has about a specific stimulus s, and is given by

$$I(s,R) = \sum_{r \in R} P(r|s) \log_2 \frac{P(r|s)}{P(r)} \tag{5}$$

where r is an individual response from the set of responses R.

However, the single cell information measure cannot give a complete assessment of VisNet's performance with respect to invariant object recognition. If all output cells learned to respond to the same stimulus then there would in fact be relatively little information available about the set of stimuli S, and single cell information measures alone would not reveal this. To address this issue, we also calculated a multiple cell information measure, which assesses the amount of information that is available about the whole set of stimuli from a population of neurons. Procedures for calculating the multiple cell information measure are described by Rolls, Treves and Tovee (1997) and Rolls and Milward (2000). In brief, we calculate the mutual information, that is, the average amount of information that is

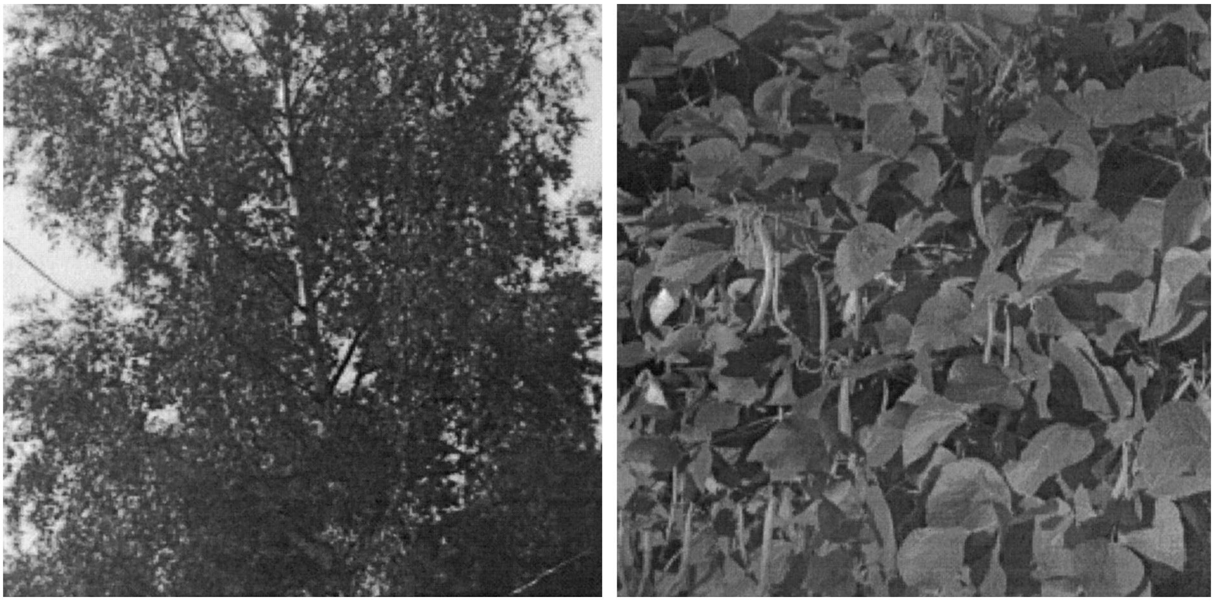

Fig. 3. Cluttered backgrounds used in VisNet simulations: backgrounds 1 and 2 are left and right, respectively.

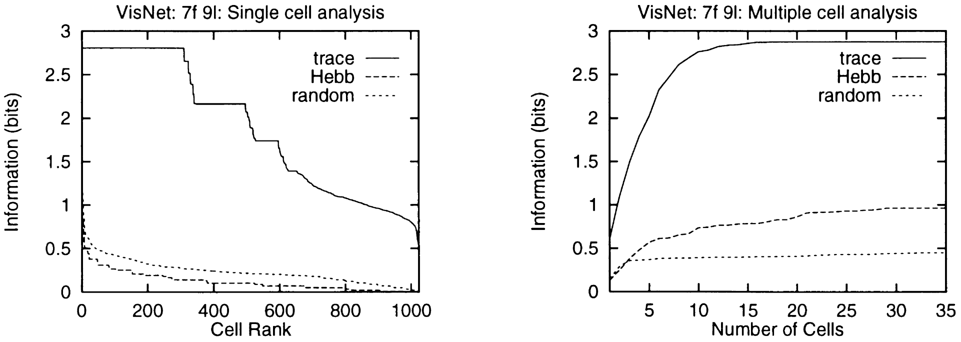

Fig. 4. Numerical results for experiment 1, with the 7 faces presented on a blank background during both training and testing. On the left are single cell information measures, and on the right are multiple cell information measures. Results are presented for the network trained with the trace rule (4), the Hebb rule (3), and random weights.

obtained about which stimulus was shown from a single presentation of a stimulus from the responses of all the cells. That is, the mutual information between the whole set of stimuli S and of responses R is the average across stimuli of this stimulus-specific information. This is achieved through a decoding procedure, in which the stimulus s' that gave rise to the particular firing rate response vector on each trial is estimated. A probability table is then constructed of the real stimuli s and the decoded stimuli s'. From this probability table, the mutual information is calculated as

$$I(s,s') = \sum_{s,s'} P(s,s')\log_2 \frac{P(s,s')}{P(s)P(s')}. \tag{6}$$

Multiple cell information values were calculated for the subset of cells which had as single cells the most information about which stimulus was shown. In particular, we calculated the multiple cell information from the first five cells for each stimulus that had maximal single cell information about that stimulus, that is from a population of 35 cells given that there were seven stimuli. This was found to be a sufficiently large subset of the cells to enable a check that the cells did indeed include some tuned to each of the different stimuli. The criterion for perfect performance to all stimuli was that the multiple cell information should reach the information needed to fully discriminate the set of stimuli, that is $\log_2 S$ bits.

2. VisNet simulations with stimuli in cluttered backgrounds

2.1. Previously trained stimuli tested in cluttered backgrounds

In the simulations in this section we begin by testing with VisNet2 how, after the network has been trained with stimuli presented on a blank background, testing with the stimuli presented in cluttered backgrounds affects the performance. Is invariant recognition still possible? Previous investigations with VisNet have involved the presentation of stimuli against a blank (e.g. greyscale 127) background. In the simulations described here we compare such results with simulations performed with the cluttered backgrounds as shown in Fig. 3.

Experiment 1 involves testing the network with the 7 face stimuli shown in Fig. 2 presented during training and recognition on a blank background. This experiment provides a baseline performance with which to compare results from later experiments with cluttered backgrounds or partially occluded stimuli. Numerical results for experiment 1 are presented in Fig. 4. On the left are the single cell information measures for all top (4th) layer neurons ranked in order of their invariance to the faces, while on the right are the multiple cell information measures. Results are presented for the network trained with the trace rule (4) or the Hebb rule (3), or untrained with the initial random weights. It may be seen that view invariant neurons with high single cell information measures only develop with the network trained with the trace rule (4). The results with trace rule (4) show a mildly reduced performance compared to results given in Rolls and Milward (2000). This is because in this present work the face stimuli are carefully separated from their original 64 × 64 natural backgrounds before being inserted into the 128 × 128 backgrounds used here. However, it can still be seen that a number of cells have reached the maximum possible single cell information measure of 2.8 bits ($\log_2$ of the number of stimuli) for this test case, and that the multiple cell information measures also reach the 2.8 bits indicating perfect performance.

In experiment 2, VisNet is trained with the 7 face stimuli presented on a blank background, but tested with the faces presented on each of the 2 cluttered backgrounds. Fig. 5 shows results for experiment 2, with single and multiple cell information measures on the left and right respectively. Comparing Figs. 4 and 5 shows that there is very little deterioration in performance when testing with the faces presented on either of the 2 cluttered backgrounds. This is an interesting result to compare with many artificial vision systems that would need to carry out computationally

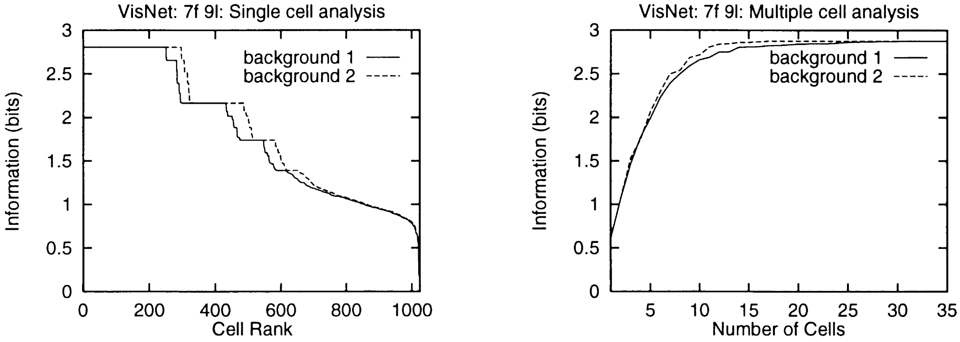

Fig. 5. Numerical results for experiment 2, with the 7 faces presented on a blank background during training and a cluttered background during testing. On the left are single cell information measures, and on the right are multiple cell information measures.

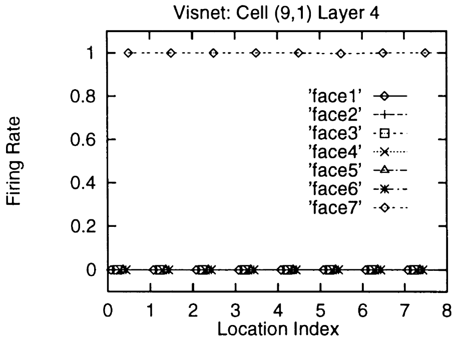

Fig. 6. Response profiles of a top layer neuron to the 7 faces from experiment 2, with the faces presented against cluttered background 1 during testing.

intensive serial searching and template matching procedures in order to achieve such results. In contrast, the VisNet neural network architecture is able to perform such recognition relatively quickly through a simple feedforward computation. Further results from experiment 2 are presented in Fig. 6 where we show the response profiles of a 4th layer neuron to the 7 faces presented on cluttered background 1 during testing. It can be seen that this neuron achieves excellent invariant responses to the 7 faces even with the faces presented on a cluttered background. The response profiles are independent of location but differentiate between the faces in that the responses are maximal for only one of the faces and minimal for all other faces.

2.2. Training with stimuli presented in cluttered backgrounds

In experiment 3, VisNet was trained with the 7 face stimuli presented on either one of the 2 cluttered backgrounds, but tested with the faces presented on a blank background. Results for experiment 3 are shown in Fig. 7, with single and multiple cell information measures on the left and right, respectively. This time, however, performance is very significantly degraded, with no cells reaching the maximum possible single cell information measure of 2.8 bits. This is in stark contrast to results from experiment 2, and reveals a significant asymmetry in terms of the effect of a cluttered background during learning and recognition. In this case, during training the network has learned to respond to a combination of a face and a cluttered background, and does not produce perfect recognition when the face is presented alone on a plain background.

The difficulty in producing good results when the stimuli are presented during training in cluttered backgrounds is even more evident in experiment 4, in which VisNet is both trained and tested with the 7 face stimuli presented on either one of the 2 cluttered backgrounds, with the same background being used for both training and testing.

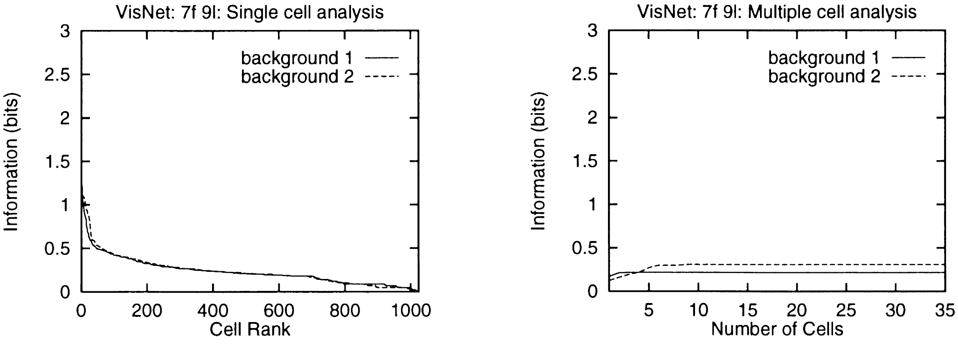

Fig. 7. Numerical results for experiment 3, with the 7 faces presented on a cluttered background during training and a blank background during testing. On the left are single cell information measures, and on the right are multiple cell information measures.

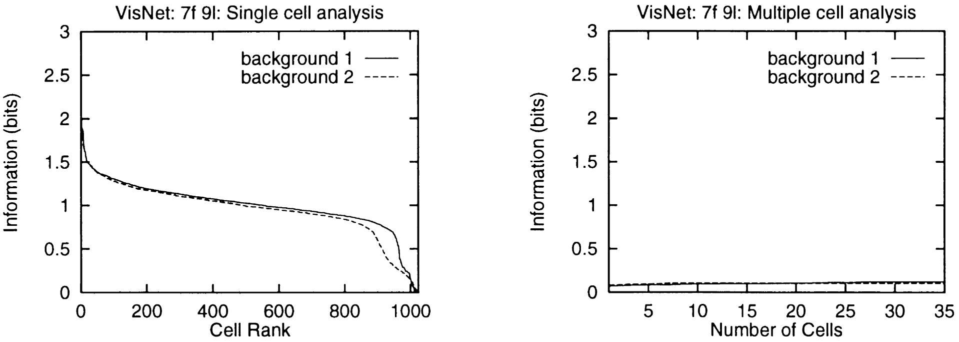

Fig. 8. Numerical results for experiment 4, with the 7 faces presented on the same cluttered background during both training and testing. On the left are single cell information measures, and on the right are multiple cell information measures.

Results for experiment 4 are shown in Fig. 8, with single and multiple cell information measures on the left and right, respectively. In particular, Fig. 9 shows the response profiles of a typical 4th layer neuron to the 7 faces presented on cluttered background 1 during training and testing. It can be seen that this neuron has learned to respond to all of the faces in all of the positions, which implies that the cell has simply learned to respond to the background.

Part of the difficulty that hierarchical multilayer competitive networks have with learning in cluttered environments may be that without explicit segmentation of the stimulus from its background, at least some of the features that should be formed to encode the stimuli are not formed properly, because the neurons learn to respond to combinations of inputs which come partly from the stimulus, and partly from the background. To investigate this, we performed experiment 5 in which we pretrained layers 1–3 with stimuli to ensure that good feature combination neurons for stimuli were available, and then allowed learning in only layer 4 when stimuli were presented in the cluttered backgrounds.

In experiment 5 VisNet is first exposed to a completely random sequence of the face stimuli in different positions against a blank background during which layers 1–3 are allowed to learn for the usual number of epochs. The effect of this is to set up feature detecting neurons in the early layers that are tuned to this general class of stimulus. However, this initial random exposure to the face stimuli cannot develop position invariant responses among top layer neurons since there is no temporal structure to the order of the different positions in which the faces are presented at this stage. That is, the different positions in which the presentations of a given stimulus are presented are not constrained to occur in temporal proximity during this initial exposure. However, the presence of the stimulus tuned feature detecting neurons in the early layers has a significant impact on the subsequent ability of VisNet to develop invariant neurons when the 4th layer is properly trained with the stimuli presented against cluttered backgrounds. The next step, then, is to train layer 4 in the usual way with the 7 faces presented against a cluttered background, where the images are now presented such that different positions for the same face occur close together in time. Results for experiment 5 are shown in Fig. 10, with single and multiple cell information measures on the left and right, respectively. Comparing Figs. 8 and 10 shows that prior random exposure to the face stimuli has led to much improved performance. Indeed, it can be seen that a number of cells have reached the maximum possible single cell information measure of 2.8 bits for this test case, although the multiple cell information measures do not quite reach the 2.8 bits that would indicate perfect performance for the complete face set. Response profiles of a top layer neuron to the 7 faces from experiment 5, with the faces presented against cluttered background 1 during training of layer 4 and testing are shown in Fig. 11. It can be seen that this neuron has developed excellent invariant responses to the 7 faces.

These results demonstrate that the problem of developing position invariant neurons to stimuli occurring against cluttered backgrounds may be ameliorated by the prior existence of stimulus tuned feature detecting neurons in the early layers of the visual system, and that these feature detecting neurons may be set up through previous exposure to the relevant class of objects.

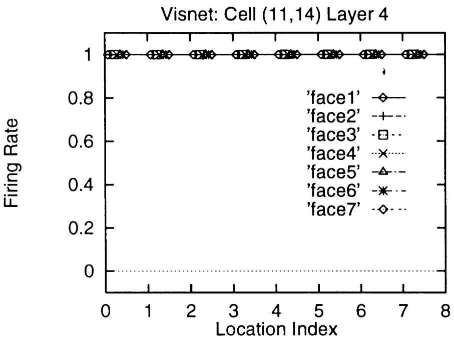

Fig. 9. Response profiles of a top layer neuron to the 7 faces from experiment 4, with the faces presented against cluttered background 1 during training and testing.

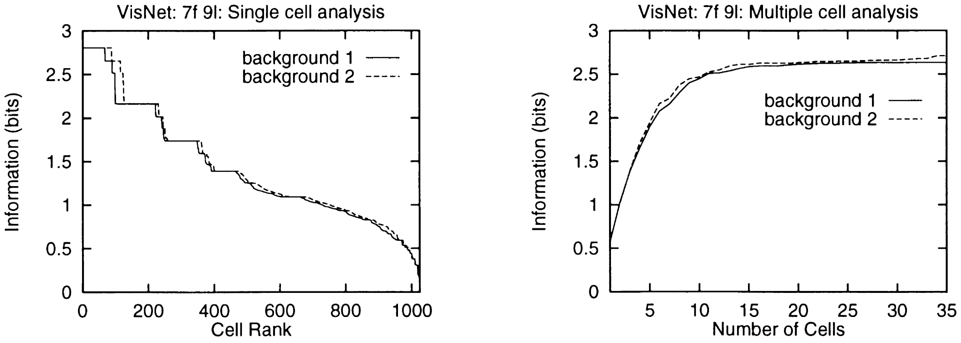

Fig. 10. Numerical results for experiment 5. In this experiment VisNet is first exposed to a completely random sequence of faces in different positions against a blank background during which layers 1–3 are allowed to learn. This builds general features detecting neurons in the lower layers that are tuned to the face stimuli, but cannot develop view invariance since there is no temporal structure to the order in which different views of different faces occur. Then layer 4 is trained in the usual way with the 7 faces presented against a cluttered background, where the images are now presented such that different views of the same face occur close together in time. On the left are single cell information measures, and on the right are multiple cell information measures.

3. VisNet simulations with partially occluded stimuli

In this section we examine the performance of VisNet2 tested with partially occluded stimuli. In these simulations, training and testing is performed with a blank background to avoid confounding the two separate problems of occlusion and background clutter. In object recognition tasks, artificial vision systems may typically rely on being able to locate a small number of key markers on a stimulus in order to be able to identify it. This approach can become fragile when a number of these markers become obscured. In contrast, biological vision systems may generalise or complete from a partial input as a result of the use of distributed representations in neural networks (for introduction see Rolls & Treves, 1998 and for early contributions see Kohonen, 1989 and Willshaw, Buneman & Longuet-Higgins, 1969) and this could lead to greater robustness in situations of partial occlusion.

In experiment 6, the network is first trained with the 7 face stimuli without occlusion, but during testing there are two options: either (i) the top halves of all the faces are occluded, or (ii) the bottom halves of all the faces are occluded. Since VisNet is tested with either the top or bottom half of the stimuli no stimulus features are common to the two test options. This ensures that if performance is good with both options, the performance cannot be based on the use of a single feature to identify a stimulus. Results for experiment 6 are shown in Fig. 12, with single and multiple cell information measures on the left and right, respectively. Comparing Figs. 4 and 12 show that there is only a modest drop in performance in the single cell information measures when the stimuli are partially occluded. For both options (i) and (ii), even with partially occluded stimuli, a number of cells continue to respond maximally to one preferred stimulus in all locations, while responding minimally to all other stimuli. However, comparing results from options (i) and (ii) shows that the network performance is better when the bottom half of the faces is occluded. This is consistent with psychological results showing that face recognition is performed more easily when the top halves of faces are visible rather than the bottom halves (see Bruce, 1988). The top half of a face will generally contain salient features, e.g. eyes and hair, that are particularly helpful for recognition, and it is interesting that these simulations appear to further demonstrate this point. Furthermore, the multiple cell information measures confirm that performance is better with the upper half of the face visible (option (ii)) than the lower half (option (i)), in that when the top halves of the faces are occluded the multiple cell information measure asymptotes to a sub-optimal value reflecting the difficulty of discriminating between these more difficult images. Further results from experiment 6 are presented in Fig. 13 where we show the response profiles of a 4th layer neuron to the 7 faces, with the bottom half of all the faces occluded during testing. It can be seen that this neuron continues to respond invariantly to the 7 faces, responding maximally to one of the faces but minimally for all other faces.

4. Discussion

The results of experiments 1 and 2 show that once this

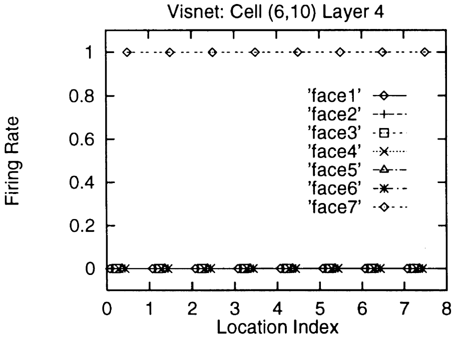

Fig. 11. Response profiles of a top layer neuron to the 7 faces from experiment 5, with the faces presented against cluttered background 1 during training of layer 4 and testing.

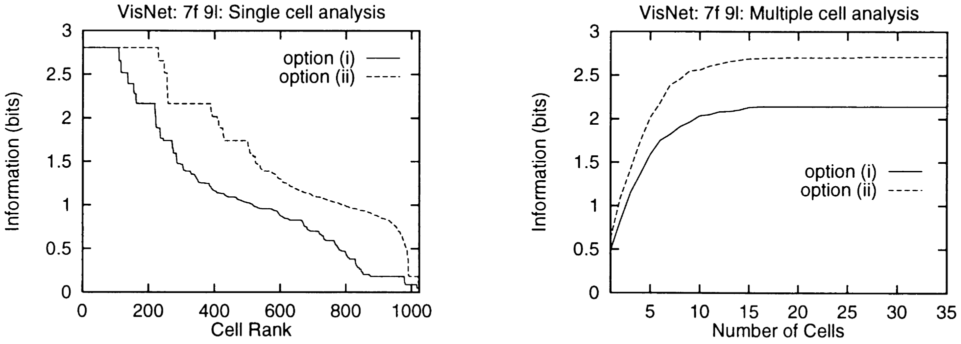

Fig. 12. Numerical results for experiment 6, with the 7 faces presented on a blank background during both training and testing. However, during testing there are two options: either (i) the top half of all the faces are occluded; or (ii) the bottom half of all the faces are occluded. On the left are single cell information measures, and on the right are multiple cell information measures.

class of network has been trained on a set of stimuli presented against a blank background, then it can later recognise a stimulus with invariance even when the stimulus is presented in a cluttered background. This is an interesting and important result, for it shows that after learning, special mechanisms for segmentation and for attention are not needed in order for neurons already tuned by previous learning to the stimuli to be activated correctly in the output layer. Although the experiments described here tested for position invariance, we predict and would expect that the same results would be demonstrable for size and view invariant representations of objects.

When tested in cluttered environments, the background clutter may of course activate some other neurons in the output layer, but at least the neurons that have learned to respond to the trained stimuli are activated. The result of this activity is sufficient for the activity in the output layer to be useful, in the sense that it can be read off correctly by a pattern associator connected to the output layer. Indeed, we have tested this by connecting a pattern associator to layer 4 of VisNet. The pattern associator has seven neurons, one for each face, and 1024 inputs, one from each neuron in layer 4 of VisNet. The pattern associator learned when trained with a simple associative Hebb rule (3) to activate the correct output neuron whenever one of the faces was shown in any position in the uncluttered environment. This ability was shown to be dependent on invariant neurons for each stimulus in the output layer of VisNet, for the pattern associator could not be taught the task if VisNet had not been previously trained to produce invariant representations. Then it was shown that exactly the correct neuron was activated when any of the faces was shown in any position with the cluttered background. This read-off by a pattern associator is exactly what we hypothesize takes place in the brain, in that the output of the inferior temporal visual cortex (where neurons with invariant responses are found) projects to structures such as the orbitofrontal cortex and amygdala, where associations between the invariant visual representations and stimuli such as taste and touch are learned (Rolls, 1999; Rolls & Treves, 1998). Thus testing whether the output of an architecture such as VisNet can be used effectively by a pattern associator is a very biologically relevant way to evaluate the performance of this class of architecture.

The results of experiments 3 and 4 suggest that in order for a cell to *learn* invariant responses to different transforms of a stimulus when it is presented during training in a cluttered background, some form of segmentation is required in order to separate the figure (i.e. the stimulus or object) from the background. This segmentation might be performed using evidence in the visual scene about different depths, motions, colours, etc. of the object from its background. In the visual system, this might mean combining evidence represented in different cortical areas, and might be performed by cross-connections between cortical areas to enable such evidence to help separate the representations of objects from their backgrounds in the form-representing cortical areas. However, we note that it was not possible in the experiments described here to change the background from trial to trial during learning due to the complexity of VisNet's image construction and pre-processing stage. It is possible that if the background did continually change during learning, whereas the object being learned about tended to be present (though in a somewhat transformed version) from trial to trial, then the architecture would

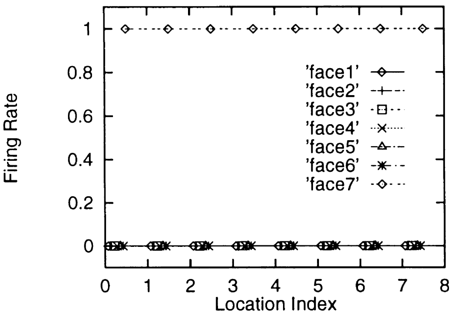

Fig. 13. Response profiles of a top layer neuron to the 7 faces from experiment 6, with the bottom half of all the faces occluded during testing.

have little that it could learn across trials by trace learning about the backgrounds, but would learn about the transforming object. Hence, it may be possible that an architecture such as VisNet cannot only recognise previously learned objects when presented in a cluttered background as shown here, but may also be able to learn invariant representations of objects provided that the background is not constant.

Another mechanism that might help the learning of new objects is attention. An attentional mechanism might highlight the current stimulus being attended to and suppress the effects of background noise, providing a training representation of the object more like that, which would be produced when it is presented against a blank background. With respect to attention, many neurophysiological experiments have demonstrated an attentional modulation of neuronal responses in visual areas V1, V2, V4, MT and MST (Luck, Chelazzi, Hillyard & Desimone, 1997; Motter, 1993; Reynolds, Chelazzi & Desimone 1999; Treue & Maunsell, 1996), and models of these sorts of attentional processes have been proposed by Hahnloser, Douglas, Mahowald and Hepp (1999) and Olhausen, Anderson and Van Essen (1993). Such attentional mechanisms could help subsequent brain structures to associate for example a fixated object with reward, and not to associate other objects in the visual field with reward, by making the response to the object being attended to considerably larger than for other objects. However, attention would also be useful for the trace-based learning of view invariant responses to novel objects in the environment. In this case, the output of the visual system for a particular object must be associated with all the different views of the same object, but not associated with other objects in the field of view. Hence, we propose that one way in which the visual system may solve the problem of learning view invariant responses to individual objects in cluttered environments uses an attentional mechanism in a similar way to that suggested for object-reward reinforcement learning. If such an attentional mechanism is required for the development of view invariance, then it follows that cells in the temporal cortex may only develop transform invariant responses to objects to which attention is directed.

The implication of the findings described here is that we have shown that explicit attentional and segmentation processes are not required in the type of architecture described for invariant responses to previously learned objects to be obtained. There may be some advantage to having foveally weighted processing, or an explicit attentional mechanism, to facilitate the readout of information under these circumstances when the previously learned objects are presented in cluttered environments, as the background will activate some output neurons. On the other hand, learning invariant representations of new objects in cluttered backgrounds may be simplified by having processes that either perform segmentation, or focus attention on one part of the input, or perform both (perhaps together).

Acknowledgements

This research was supported by the Medical Research Council, grant PG9826105, and by the MRC Interdisciplinary Research Centre for Cognitive Neuroscience.

References

Bruce, V. (1988). *Recognising faces*. Hillsdale, NJ: Erlbaum.

Elliffe, M.C.M., Rolls, E.T. & Stringer, S.M. (2000). Invariant recognition of feature combinations in the visual system (submitted for publication).

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

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.

Hahnloser, R., Douglas, R. J., Mahowald, M., & Hepp, K. (1999). Feedback interactions between neuronal pointers and maps for attentional processing. *Nature Neuroscience*, *2* (8), 746–752.

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

Kohonen, T. (1989). *Self-organization and associative memory*. (3rd ed.). Berlin: Springer.

Luck, S. J., Chelazzi, L., Hillyard, S. A., & Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas v1, v2, and v4 of macaque visual cortex. *Journal of Neurophysiology*, *77*, 24–42.

Motter, B. C. (1993). Focal attention produces spatially selective processing in visual cortical areas v1, v2, and v4 in the presence of competing stimuli. *Journal of Neurophysiology*, *70*, 909–919.

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.

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

Reynolds, J. H., Chelazzi, L., & Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas v2 and v4. *Journal of Neuroscience*, *19*, 1736–1753.

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. (1994). Brain mechanisms for invariant visual recognition and learning. *Behavioural Processes*, *33*, 113–138.

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

Rolls, E. T. (1999). *The brain and emotion*. Oxford: Oxford University Press.

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

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

Rolls, E.T. & Stringer, S.M. (2000). Invariant object recognition in the visual system with error correction and temporal difference learning (submitted for publication).

Rolls, E. T., & Treves, A. (1998). *Neural networks and brain function*. Oxford: Oxford University Press.

Rolls, E. T., Treves, A., & Tovee, M. J. (1997). The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. *Experimental Brain Research*, *114*, 177–185.

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

Treue, S., & Maunsell, J. H. (1996). Attentional modulation of visual motion processing in cortical areas mt and mst. *Nature*, *382*, 539–541.

Ullman, S. (1996). *High-level vision*. Cambridge, MA: MIT press.

Wallis, G., & Rolls, E. T. (1997). A model of invariant object recognition in the visual system. *Progress in Neurobiology*, *51*, 167–194.

Willshaw, D. J., Buneman, O. P., & Longuet-Higgins, H. C. (1969). Non-holographic associative memory. *Nature*, *222*, 960–962.