Pergamon

Progress in Neurobiology, Vol. 51, pp. 167 to 194, 1997

Printed in Great Britain
0301-0082/97/$32.00

PII: S0301-0082(96)00054-8

INVARIANT FACE AND OBJECT RECOGNITION IN THE VISUAL SYSTEM

GUY WALLIS* and EDMUND T. ROLLS†

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

(*Received* 8 *July* 1996)

Abstract—Neurophysiological evidence is described, showing that some neurons in the macaque temporal cortical visual areas have responses that are invariant with respect to the position, size and view of faces and objects, and that these neurons show rapid processing and rapid learning. A theory is then described of how such invariant representations may be produced in a hierarchically organized set of visual cortical areas with convergent connectivity. The theory proposes that neurons in these visual areas use a modified Hebb synaptic modification rule with a short-term memory trace to capture whatever can be captured at each stage that is invariant about objects as the object changes in retinal position, size, rotation and view. Simulations are then described which explore the operation of the architecture. The simulations show that such a processing system can build invariant representations of objects.

CONTENTS

1. INTRODUCTION

This paper draws together evidence on how information about visual stimuli is represented in the temporal cortical visual areas, and on how representations that are invariant with respect to the position, size and even view of objects are formed. The evidence comes from neurophysiological studies of single neuron activity in primates. It also comes from closely related theoretical studies which consider how the representations may be set up by learning in a multistage cortical architecture. The neurophysiological evidence considered comes in part from neural systems involved in processing information about faces, because with the large number of neurons devoted to this class of stimuli, this system has proved amenable to experimental analysis. However, recent evidence also described suggests that there may be a similar neural system that produces invariant representations of non-face objects (Rolls *et al.*, 1996b).

*Present address: Max-Planck Institut für biologische Kybernetik, Spemannstrasse 38, 72076 Tübingen, Germany.

†Author for correspondence. Tel: + 44-(0)1865-271348; Fax: + 44-(0)1865-310447; E-mail: Edmund.Rolls@psy.ox.ac.uk.

2. NEUROPHYSIOLOGY OF THE TEMPORAL CORTICAL VISUAL AREAS

2.1. Visual Cortical Areas in the Temporal Lobes

Visual pathways project via a number of cortico-cortical stages from the primary visual cortex until they reach the temporal lobe visual cortical areas (Seltzer and Pandya, 1978; Maunsell and Newsome, 1987; Baizer *et al.*, 1991). The inferior temporal visual cortex, area TE, is divided into a set of subareas, and in addition there is a set of different areas in the cortex in the superior temporal sulcus (Seltzer and Pandya, 1978; Baylis *et al.*, 1987) (see Fig. 1). Of these latter areas, TPO receives inputs from temporal, parietal and occipital cortex; PGa and IPa from parietal and temporal cortex; and TS and TAa primarily from auditory areas (Seltzer and Pandya, 1978). There is considerable specialization of function in these areas (Baylis *et al.*, 1987). For example, areas TPO, PGa and IPa are multimodal, with neurons which respond to visual, auditory and/or somatosensory inputs; the inferior temporal gyrus and adjacent areas (TE3, TE2, TE1, TEa and TEm) are primarily unimodal visual areas; areas in the cortex in the anterior and dorsal part of the superior temporal sulcus (e.g. TPO, IPa and IPg) have neurons specialized for the analysis of moving visual stimuli; and neurons responsive primarily to faces are found more frequently in areas TPO, TEa and TEm (Baylis *et al.*, 1987), where they comprise approximately 20% of the visual neurons responsive to stationary stimuli, in contrast with the other temporal cortical areas in which they comprise 4–10%. The neurons which respond to non-face stimuli and the other neurons that respond to faces often require two or more simple features to be present in the correct spatial relationship in order to respond (Perrett *et al.*, 1982; Tanaka *et al.*, 1990, Tanaka *et al.*, 1991; Rolls *et al.*, 1994).

2.2. Distributed Encoding of Identity

The neurons described as having responses selective for faces are selective in that they respond from two to 20 times more (and statistically significantly more) to faces than to a wide range of gratings, simple geometrical stimuli, or complex three-dimensional objects (see Rolls, 1984, 1992b; Baylis *et al.*, 1985, 1987). The selectivity of these neurons for faces has been quantified recently using information theory. This showed that these neurons reflected much more information about which (of 20) face stimuli had been seen (on average 0.4 bits) than about which (of 20) non-face stimuli had been seen (on average 0.07 bits) (Tovee and Rolls, 1995).

These neurons thus reflect information not just that a face has been seen, but about which face has been seen. They respond differently to different faces. An important question for understanding brain computation is whether a particular object (or face) is represented in the brain by the firing of one or a few gnostic (or "grandmother" or "cardinal") cells (Barlow, 1972), or whether instead the firing of a group or ensemble of cells, each with somewhat different responsiveness, provides the representation, as the data indicate for faces (Baylis *et al.*, 1985). A recent way in which the fineness of tuning of these neurons to individual faces has been quantified is by measurement of the sparseness of the representation, a:

$$a = (\Sigma_{s=1,S}\, r_s/S)^2/\Sigma_{s=1,S}\, (r_s^2/S)$$

where r_s is the mean firing rate to stimulus s in the set of S stimuli. The sparseness has a maximum value of 1.0 and a minimum value close to zero ($1/S$, if a neuron responded to only one of the S stimuli in a set of stimuli). [The interpretation of this measure can be made clear by means of an example. If a neuron had a binary firing rate distribution, with a high rate to some stimuli and no response to others, and the neuron responded to 50% of the stimuli, the sparseness of its representation would be 0.5 (fully distributed). If a neuron responded to just 10% of the stimuli, the sparseness of its representation would be 0.1 (sparse)]. For a sample of these cells for which the responses were tested to a set of 23 faces and 45 natural scenes, it was found that the sparseness of the representation of the 68 stimuli had an average for the set of neurons of 0.65 (Rolls and Tovee, 1995a). If the spontaneous firing rate was subtracted, then the "response sparseness" for these neurons was 0.33 (Rolls and Tovee, 1995a). It is suggested that the utility of this rather distributed encoding within the class faces is that it may enable the maximum information about a set of stimuli to be provided by a population of neurons (subject to a constraint on the average firing rate of the neurons — see Baddeley *et al.*, 1997). Such a distributed representation would be ideal for *discrimination*, for the maximum information suitable for comparing fine differences between different stimuli would be made available across the population (if 50% were active to each stimulus). In contrast, it is suggested that more sparse representations are used in memory systems such as the hippocampus, because this helps to maximize the number of different memories that can be stored (see Rolls and Tovee, 1995a; Treves and Rolls, 1994).

Although this rather distributed representation is present in the temporal cortical visual areas, it is

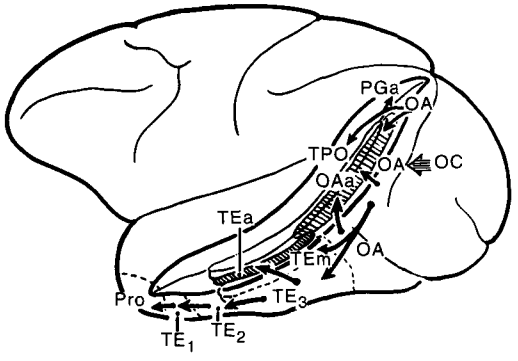


Fig. 1. Lateral view of the macaque brain (left) and coronal section (right) showing the different architectonic areas (e.g. TEm, TPO) in and bordering the anterior part of the superior temporal sulcus (STS) of the macaque (see text). The coronal section is through the temporal lobe 133 mm P (posterior) to the sphenoid reference (shown on the lateral view).

certainly not fully distributed. If the information provided by a single neuron about each of the stimuli in a set of stimuli is calculated, then it is found that the amount of information about individual stimuli can be as high as 1.5–2 bits for some stimuli (usually those which elicit the highest firing rate), and may approach zero for the stimuli in the set which produce responses that are close to the mean response of the neuron to the stimuli (Rolls *et al.*, 1997). The advantages of this type of sparse distributed representation for cognitive processing include generalization to similar stimuli (in the Hamming distance sense, see Rolls and Treves, 1997), graceful degradation (fault tolerance), and some locality to the representation, so that some single neurons which receive inputs from such a representation can obtain sufficient information without requiring an enormous fan in, that is number of synapses (Rolls *et al.*, 1996a). (The number of synapses per neuron in the cerebral cortex is in the order of 5000, and only a proportion of these inputs will be active in any one 20 msec period.)

This information-theoretic approach has focused on how visual information about what is being looked at in the world is represented by the activity of individual neurons. How does the process scale with more neurons than one? Evidence recently has been obtained that the information available about which visual stimulus (which of 20 equiprobable faces) had been shown increases linearly with the number of neurons in the sample (Rolls *et al.*, 1996a; Abbott *et al.*, 1996). Because information is a logarithmic measure, this indicates that the number of stimuli encoded rises approximately exponentially, as the number of cells in the sample increases. The consequence of this is that large numbers of stimuli, and fine discriminations between them, can be represented without having to measure the activity of an enormous number of neurons. For example, the results of the experiments of Rolls *et al.*, 1996a) indicate that the activity of 15 neurons would be able to encode 192 face stimuli (at 50% accuracy), of 20 neurons 768 stimuli, of 25 neurons 3072 stimuli, of 30 neurons 12288 stimuli, and of 35 neurons 49152 stimuli (Abbott *et al.*, 1996; the values are for an optimal decoding case). This means that it is now possible to read the code about face identity from the end of this part of the visual system. By measuring the firing rates of relatively small numbers (tens) of neurons, we know which (of potentially hundreds or thousands) of visual stimuli are being looked at by the monkey. It is of interest that much information is available from the firing rates of an ensemble of neurons, with no account being taken of the relative time of firing of the spikes in the different neurons (cf Engel *et al.*, 1992).

2.3. A Neuronal Representation of Faces and Objects Showing Invariance

One of the major problems which must be solved by a visual system used for object recognition is the building of a representation of visual information which allows recognition to occur relatively independently of size, contrast, spatial frequency, position on the retina, and angle of view, etc. We have shown that many of the neurons whose responses reflect face identity have responses that are relatively invariant with respect to size and contrast (Rolls and Baylis, 1986); spatial frequency (Rolls *et al.*, 1985, Rolls *et al.*, 1987); and retinal translation, i.e. position in the visual field (Tovee *et al.*, 1994; cf earlier work by Gross, 1973, Gross *et al.*, 1985). Some of these neurons even have relatively view-invariant responses, responding to different views of the same face but not of other faces (Hasselmo *et al.*, 1989a). It is clearly important that invariance in the visual system is made explicit in the neuronal responses, for this simplifies greatly the output of the visual system to memory systems such as the hippocampus and amygdala, which can then remember or form associations about *objects*. The function of these memory systems would be almost impossible if there were no consistent output from the visual system about objects (including faces), for then the memory systems would need to learn about all possible sizes, positions etc of each object, and there would be no easy generalization from one size or position of an object to that object when seen with another retinal size or position.

Although the neurons just described have view-invariant responses, there is another population of face-selective neurons, found particularly in the cortex in the superior temporal sulcus, which tends to have view-dependent responses (Perrett *et al.*, 1985a; Hasselmo *et al.*, 1989b). Some of these neurons have responses which reflect the facial expression but not the facial identity of the stimulus (Hasselmo *et al.*, 1989a). These neurons could be useful in providing information of potential use in social interactions (Rolls, 1984, 1990, 1992a; Perrett *et al.*, 1985b). Damage to this population and to brain areas to which these neurons project may contribute to the deficits in social and emotional behaviour produced by temporal or ventral frontal lobe damage (see Rolls, 1984,1990, 1991, 1992, 1995a, 1996b; Leonard *et al.*, 1985; Hornak *et al.*, 1996).

To investigate whether view-invariant representations of objects are also encoded by some neurons in the inferior temporal cortex of the rhesus macaque, the activity of single neurons was recorded while monkeys were shown very different views of 10 objects (Rolls *et al.*, 1996b). The stimuli were presented for 0.5 sec on a colour video monitor while the monkey performed a visual fixation task. The stimuli were images of 10 real plastic objects which had been in the monkey's cage for several weeks, to enable him to build view invariant representations of the objects. Control stimuli were views of objects which had never been seen as real objects. The neurons analysed were in the TE cortex in and close to the ventral lip of the anterior part of the superior temporal sulcus. Many neurons were found that responded to some views of some objects. However, for a smaller number of neurons, the responses occurred only to a subset of the objects, irrespective of the viewing angle. These neurons thus conveyed information about which object has been seen, independently of view, as confirmed by information theoretic analysis of the neuronal responses. Each neuron did not, in general, respond to only one object, but instead responded to a subset of the

objects. Thus, they showed ensemble, sparse-distributed, encoding. The information available about which object was seen increased approximately linearly with the number of neurons in the ensemble. These experiments provide preliminary evidence that there is a view-invariant representation of objects, as well as faces, in the primate temporal cortical visual areas.

Until now, research on translation invariance has considered the case in which there is only one object in the visual field. The question then arises of how the visual system operates in a cluttered environment. Do all objects that can activate an inferior temporal neuron do so whenever they are anywhere within the large receptive fields of inferior temporal cortex neurons? If so, the output of the visual system might be confusing for structures which receive inputs from the temporal cortical visual areas. To investigate this we measured the responses of inferior temporal cortical neurons with face-selective responses in rhesus macaques performing a visual fixation task. We found that the response of neurons to an effective face centred 8.5° from the fovea was decreased to 71% if an ineffective face stimulus for that cell was present at the fovea. If an ineffective stimulus for a cell is introduced parafoveally when an effective stimulus is being fixated, then there was a similar reduction in the responses of neurons. More concretely, the mean firing rate across all cells to a fixated effective face with a non-effective face in the periphery was 34 spikes/sec. On the other hand, the average response to a fixated non-effective face with an effective face in the periphery was 22 spikes/sec. (These firing rates reflected the fact that in this population of neurons, the mean response for an effective face was 49 spikes/sec with the face at the fovea, and 35 spikes/sec with the face 8.5° from the fovea.) Thus, these cells gave a reliable output about which stimulus is actually present at the fovea, in that their response was larger to a fixated effective face than to a fixated non-effective face, even when there are other parafoveal stimuli ineffective or effective for the cell (Rolls and Tovee, 1995b). Thus, the cell provides information biased towards what is present at the fovea, and not equally about what is present anywhere in the visual field. This makes the interface to action simpler, in that what is at the fovea can be interpreted (e.g. by an associative memory) partly independently of the surroundings, and choices and actions can be directed if appropriate to what is at the fovea (cf Ballard, 1993). These findings are a step towards understanding how the visual system functions in a normal environment.

2.4. Learning of New Representations in the Temporal Cortical Visual Areas

Given the fundamental importance of providing an ensemble-encoded representation of faces and objects which nevertheless has quite finely tuned neurons, experiments have been performed to investigate whether experience plays a role in determining the selectivity of single neurons which respond to faces. The hypothesis being tested was that visual experience might guide the formation of the responsiveness of neurons so that they provide an economical and ensemble-encoded representation of items actually present in the environment. To test this, Rolls *et al.* (1989) investigated whether the responses of temporal cortex face-selective neurons were at all altered by the presentation of new faces which the monkey had never seen before. It might be, e.g., that the population would make small adjustments in the responsiveness of its individual neurons, so that neurons would acquire tuning properties which would enable the population as a whole to discriminate between the faces actually seen. Thus, they investigated whether when a set of totally novel faces was introduced, the responses of these neurons were fixed and stable from the first presentation, or instead whether there was some adjustment of responsiveness over repeated presentations of the new faces. Firstly, it was shown for each neuron tested that its responses were stable over 5–15 repetitions of a set of familiar faces. Then a set of new faces was shown in random order (with 1 sec for each presentation), and the set was repeated with a new random order over many iterations. Some of the neurons studied in this way altered the relative degree to which they responded to the different members of the set of novel faces over the first few (from one to two) presentations of the set (Rolls *et al.*, 1989). If, in a different experiment, a single novel face was introduced when the responses of a neuron to a set of familiar faces was being recorded, it was found that the responses to the set of familiar faces were not disrupted, while the responses to the novel face became stable within a few presentations. Thus, there is now some evidence from these experiments that the response properties of neurons in the temporal lobe visual cortex are modified by experience, and that the modification is such that when novel faces are shown, the relative responses of individual neurons to the new faces alter (Rolls *et al.*, 1989). It is suggested that alteration of the tuning of individual neurons in this way results in a good discrimination over the population as a whole of the faces known to the monkey. This evidence is consistent with the categorization being performed by self-organizing competitive neuronal networks, as described below and elsewhere (Rolls, 1989a, 1989b, 1989c; Rolls and Treves, 1997).

Further evidence that these neurons can learn new representations very rapidly comes from an experiment in which binarized black and white images of faces which blended with the background were used. These did not activate face-selective neurons. Full grey-scale images of the same photographs were then shown for 10 0.5 sec presentations. It was found in a number of cases, if the neuron happened to be responsive to the face, that when the binarized version of the same face was shown next, the neurons responded to it (Rolls *et al.*, 1993; Tovee *et al.*, 1996). This is a direct parallel to the same phenomenon which is observed psychophysically, and provides dramatic evidence that these neurons are influenced by only a very few seconds (in this case 5 sec) of experience with a visual stimulus.

Such rapid learning of representations of new objects, which occurs in humans in a few seconds, appears to be a major type of learning in which the

temporal cortical areas are involved. Ways in which this learning could occur are considered below.

It is also the case that there is a much shorter-term form of memory in which some of these neurons are involved, for whether a particular visual stimulus (such as a face) has been seen recently, in that some of these neurons respond differently to recently seen stimuli in short-term visual memory tasks (Baylis and Rolls, 1987; Miller and Desimone, 1994). A tendency of some temporal cortical neurons to associate together visual stimuli when they have been shown over many repetitions separated by several seconds also has been described by Miyashita and Chang (1988); see also Miyashita (1993). In addition, Logothetis *et al.* (1994) using extensive training (600 training trials) showed that neurons could alter their responses to different views of computer-generated objects.

2.5. The Speed of Processing in the Temporal Cortical Visual Areas

An important constraint on the type of processing that could be involved in object recognition is provided by the speed of operation of each cortical stage involved in object recognition. There is evidence that it is very fast, as shown by the following. There is a whole sequence of visual cortical processing stages including V1, V2, V4, and the posterior inferior temporal cortex via which information reaches the anterior temporal cortical areas. Further, the response latencies of neurons in V1 are about 40–50 msec, and in the anterior inferior temporal cortical areas approximately 80–100 msec. This suggests that each stage may need to perform processing for only 15–30 msec before it has performed sufficient processing to start influencing the next stage. Consistent with this, response latencies between V1 and the inferior temporal cortex increase from stage to stage (Thorpe and Imbert, 1989). Because of the importance of the speed of processing, it has been investigated quantitatively as follows.

In a first approach, the information available in short temporal epochs of the responses of temporal cortical face-selective neurons about which face had been seen was measured. It was found that if a period of the firing rate of 50 msec was taken, then this contained 84.4% of the information available in a much longer period of 400 msec about which of four faces had been seen. If the epoch was as little as 20 msec, the information was 65% of that available from the firing rate in the 400 msec period (Tovee *et al.*, 1993). These high information yields were obtained with the short epochs taken near the start of the neuronal response, e.g. in the post-stimulus period 100–120 msec. Moreover, it was shown that the firing rate in short periods taken near the start of the neuronal response was highly correlated with the firing rate taken over the whole response period, so that the information available from a neuron was stable over the whole response period of the neurons (Tovee *et al.*, 1993). This finding was extended to the case of a much larger stimulus set, of 20 faces. Again, it was found that the information available in short (e.g. 50 msec) epochs was a considerable proportion (e.g. 65%) of that available in a 400 msec long firing rate analysis period (Tovee and Rolls, 1995). These investigations thus showed that there was considerable information about which stimulus had been seen in short time epochs near the start of the response of temporal cortex neurons.

The next approach was to address the issue of the length of the period for which a cortical area must be active to mediate object recognition. This approach used a visual backward masking paradigm. In this paradigm there is a brief presentation of a test stimulus which is rapidly followed (within 1–100 msec) by the presentation of a second stimulus (the mask), which impairs or masks the perception of the test stimulus. This paradigm used psychophysically leaves unanswered for how long visual neurons actually fire under the masking condition at which the subject can just identify an object. Although there has been a great deal of psychophysical investigation with the visual masking paradigm (Turvey, 1973; Breitmeyer, 1980; Humphreys and Bruce, 1989), there is very little direct evidence on the effects of visual masking on neuronal activity. For example, it is possible that if a neuron is well tuned to one class of stimulus, such as faces, that a pattern mask which does not activate the neuron, will leave the cell firing for some time after the onset of the pattern mask. In order to obtain direct neurophysiological evidence on the effects of backward masking on neuronal activity, we analysed the effects of backward masking with a pattern mask on the responses of single neurons to faces (Rolls and Tovee, 1994). This was performed to clarify both what happens with visual backward masking, and to show how long neurons may respond in a cortical area when perception and identification are just possible. When there was no mask the cell responded to a 16 msec presentation of the test stimulus for 200–300 msec, far longer than the presentation time. It is suggested that this reflects the operation of a short-term memory system implemented in cortical circuitry (e.g. by associatively modifiable connections between nearby pyramidal cells), the potential importance of which in providing a memory trace to guide learning is considered below. If the mask was a stimulus which did not stimulate the cell (either a non-face pattern mask consisting of black and white letters N and O, or a face which was a non-effective stimulus for that cell), then as the interval between the onset of the test stimulus and the onset of the mask stimulus (the stimulus onset asynchrony, SOA) was reduced, the length of time for which the cell fired in response to the test stimulus was reduced. This reflected an abrupt interruption of neuronal activity produced by the effective face stimulus. When the SOA was 20 msec, face-selective neurons in the inferior temporal cortex of macaques responded for a period of 20–30 msec before their firing was interrupted by the mask (Rolls and Tovee, 1994). (Comparable results also have been reported for neurons responding to non-face visual stimuli by Kovacs *et al.*, 1995). We went on to show that under these conditions (a test-mask stimulus onset asynchrony of 20 msec), human observers looking at the same displays could just identify which of six faces was shown (Rolls *et al.*, 1994).

These results provide evidence that a cortical area can perform the computation necessary for the

recognition of a visual stimulus in 20–30 msec, and provide a fundamental constraint which must be accounted for in any theory of cortical computation. The results emphasize just how rapidly cortical circuitry can operate. This rapidity of operation has obvious adaptive value, and allows the rapid behavioural responses to the faces and face expressions of different individuals which are a feature of primate social and emotional behaviour. Moreover, although this speed of operation does seem fast for a network with recurrent connections (mediated by e.g. recurrent collateral or inhibitory interneurons), recent analyses of networks with analog membranes which integrate inputs, and with spontaneously active neurons, show that such networks can settle very rapidly (Treves, 1993; Simmen *et al.*, 1996).

These experiments also have implications for visual processing in relation to top-down processing. The evidence just described indicates that visual recognition can occur (measured by the subjects saying which face they saw) with largely feed-forward processing. There is not time in the experiments described for visual information to pass from V1 to V2 to V4 and thus to posterior and then anterior inferior temporal cortex, and back again all the way to V1, before V1 has started to process the second visual input, that is to have its processing of the first visual stimulus cut off by the mask.

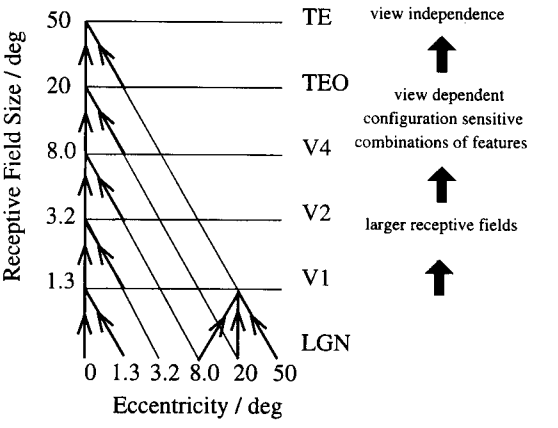


Fig. 2. Schematic diagram showing convergence achieved by the forward projections in the visual system, and the types of representation that may be built by competitive networks operating at each stage of the system from the primary visual cortex (V1) to the inferior temporal visual cortex (area TE) (see text). Area TEO forms the posterior inferior temporal cortex. The receptive fields in the inferior temporal visual cortex (e.g. in the TE areas) cross the vertical midline (not shown). Abbreviation: LGN, lateral geniculate nucleus.

2.6. Possible Computational Mechanisms in the Visual Cortex for Learning Invariant Representations

The neurophysiological findings described above, and wider considerations on the possible computational properties of the cerebral cortex (Rolls, 1989a, 1989, 1992b, 1994), lead to the following outline working hypotheses on object recognition by visual cortical mechanisms (Rolls, 1992b, 1994, 1995b). The principles underlying the processing of faces and other objects may be similar, but more neurons may become allocated to represent different aspects of faces because of the need to recognize the faces of many different individuals, i.e. to identify many individuals within the category faces.

Cortical visual processing for object recognition is considered to be organized as a set of hierarchically connected cortical regions consisting at least of V1, V2, V4, posterior inferior temporal cortex (TEO), inferior temporal cortex (e.g. TE3, TEa and TEm), and anterior temporal cortical areas (e.g. TE2 and TE1). (This stream of processing has many connections with a set of cortical areas in the anterior part of the superior temporal sulcus, including area TPO.) There is convergence from each small part of a region to the succeeding region (or layer in the hierarchy) in such a way that the receptive field sizes of neurons (e.g. 1° near the fovea in V1) become larger by a factor of approximately 2.5 with each succeeding stage (and the typical parafoveal receptive field sizes found would not be inconsistent with the calculated approximations of e.g. 8° in V4, 20° in TEO and 50° in inferior temporal cortex; Boussaoud *et al.*, 1991) (see Fig. 2). Such zones of convergence would overlap continuously with each other (see Fig. 2). This connectivity would be part of the architecture by which translation invariant representations are computed. Each layer is considered to act partly as a set of local self-organizing competitive neuronal networks with overlapping inputs. (The region within which competition would be implemented would depend on the spatial properties of inhibitory interneurons, and might operate over distances of 1–2 mm in the cortex.) These competitive nets operate by a single set of forward inputs leading to (typically non-linear, e.g. sigmoid) activation of output neurons; of competition between the output neurons mediated by a set of feedback inhibitory interneurons which receive from many of the principal (in the cortex, pyramidal) cells in the net and project back (via inhibitory interneurons) to many of the principal cells which serves to decrease the firing rates of the less active neurons relative to the rates of the more active neurons; and then of synaptic modification by a modified Hebb rule, such that synapses to strongly activated output neurons from active input axons strengthen, and from inactive input axons weaken (see Rolls, 1989c; Rolls and Treves, 1997). (A biologically plausible form of this learning rule that operates well in such networks is

$$\delta w_{ij} = k y_i (x_j - w_{ij})$$

where k is a learning rate constant, x_j is the jth input to the ith neuron, y_i is the output of the ith neuron, and w_{ij} is the jth weight on the ith neuron; see Rolls, 1989a, 1989b, 1989c; Rolls and Treves, 1997). Such competitive networks operate to detect correlations between the activity of the input neurons, and to allocate output neurons to respond to each cluster of such correlated inputs. Thus, these networks act as categorizers. In relation to visual information processing, they would remove redundancy from the input representation, and would develop low entropy

representations of the information (cf Barlow, 1985; Barlow *et al.*, 1989). Such competitive nets are biologically plausible, in that they utilize Hebb-modifiable forward excitatory connections, with competitive inhibition mediated by cortical inhibitory neurons. The competitive scheme suggested would not result in the formation of "winner-take-all" or "grandmother" cells, but would instead result in a small ensemble of active neurons representing each input (Rolls, 1989a, 1989b, 1989c). The scheme has the advantages that the output neurons learn better to distribute themselves between the input patterns (cf Bennett, 1990), and that the sparse distributed representations formed have utility in maximizing the number of memories that can be stored when, towards the end of the visual system, the visual representation of objects is interfaced to associative memory (Rolls, 1989a, 1989b; Rolls and Treves, 1990). In that each neuron has graded responses centred about an optimal input, the proposal has some of the advantages with respect to hypersurface reconstruction described by Poggio and Girosi, 1990b). However, the system proposed learns differently, in that instead of using perhaps non-biologically plausible algorithms to locate optimally the centres of the receptive fields of the neurons, the neurons use graded competition to spread themselves throughout the input space, depending on the statistics of the inputs received, and perhaps with some guidance from Backprojections (see Rolls, 1989a, 1989b). The finite width of the response region of each neuron which tapers from a maximum at the centre is important for enabling the system to generalize smoothly from the examples with which it has learned (cf Poggio and Girosi, 1990a, 1990b), to help the system to respond, e.g. with the correct invariances as described below.

Translation invariance would be computed in such a system by utilizing competitive learning to detect regularities in inputs when real objects are translated in the physical world. The hypothesis is that because objects have continuous properties in space and time in the world, an object at one place on the retina might activate feature analysers at the next stage of cortical processing, and when the object was translated to a nearby position, because this would occur in a short period (e.g. 0.5 sec), the membrane of the post-synaptic neuron would still be in its "Hebb-modifiable" state (caused e.g. by calcium entry as a result of the voltage dependent activation of NMDA receptors), and the presynaptic afferents activated with the object in its new position would thus become strengthened on the still-activated postsynaptic neuron. It is suggested that the short temporal window (e.g. 0.5 sec) of Hebb-modifiability helps neurons to learn the statistics of objects moving in the physical world, and at the same time to form different representations of different feature combinations or objects, as these are physically discontinuous and present less regular correlations to the visual system. Foldiak (1991) has proposed computing an average activation of the postsynaptic neuron to assist with the same problem. One idea here is that the temporal properties of the biologically implemented learning mechanism are such that it is well suited to detecting the relevant continuities in the world of real objects. Another suggestion is that a memory trace for what has been seen in the last 300 msec appears to be implemented by a mechanism as simple as continued firing of inferior temporal neurons after the stimulus has disappeared, as was found in the masking experiments described above (see also Rolls and Tovee, 1994; Rolls *et al.*, 1994). This would enable pairwise association of successive images of the same object. Rolls (1992b, 1994, 1995b) also has suggested that other invariances, e.g. size, spatial frequency and rotation invariance, could be learned by a comparable process. (Early processing in V1 which enables different neurons to represent inputs at different spatial scales would allow combinations of the outputs of such neurons to be formed at later stages. Scale invariance would then result from detecting at a later stage which neurons are almost conjunctively active as the size of an object alters.) It is suggested that this process takes place at each stage of the multiple-layer cortical processing hierarchy, so that invariances are learned first over small regions of space, and then over successively larger regions. This limits the size of the connection space within which correlations must be sought.

Increasing complexity of representations could also be built in such a multiple layer hierarchy by similar mechanisms. At each stage or layer the self-organizing competitive nets would result in combinations of inputs becoming the effective stimuli for neurons. In order to avoid the combinatorial explosion, it is proposed, following Feldman (1985), that low-order combinations of inputs would be what is learned by each neuron. (Each input would not be represented by activity in a single input axon, but instead by activity in a small set of active input axons.) Evidence consistent with this suggestion that neurons are responding to combinations of a few variables represented at the preceding stage of cortical processing is that some neurons in V2 and V4 respond to end-stopped lines, to tongues flanked by inhibitory subregions, or to combinations of colours (see references cited by Rolls, 1991); in posterior inferior temporal cortex to stimuli which may require two or more simple features to be present (Tanaka *et al.*, 1990); and in the temporal cortical face processing areas to images that require the presence of several features in a face (such as eyes, hair and mouth) in order to respond (see above and Yamane *et al.*, 1988). (Precursor cells to face-responsive neurons might, it is suggested, respond to combinations of the outputs of the neurons in V1 that are activated by faces, and might be found in areas such as V4.) It is an important part of this suggestion that some local spatial information would be inherent in the features which were being combined. For example, cells might not respond to the combination of an edge and a small circle unless they were in the correct spatial relation to each other. [This is, in fact, consistent with the data of Tanaka *et al.* (1990) and with our data on face neurons, in that some face neurons require the face features to be in the correct spatial configuration, and not jumbled; Rolls *et al.* (1994).] The local spatial information in the features being combined would ensure that the representation at the next level would contain some information about the (local spatial) arrangement of features.

Further low-order combinations of such neurons at the next stage would include sufficient local spatial information so that an arbitrary spatial arrangement of the same features would not activate the same neuron, and this is the proposed, and limited, solution which this mechanism would provide for the feature binding problem (cf von der Malsburg, 1990). By this stage of processing, a view-dependent representation of objects suitable for view-dependent processes such as behavioural responses to face expression and gesture would be available.

It is suggested that view-independent representations could be formed by the same type of computation, operating to combine a limited set of views of objects. 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 and van Doorn, 1979; Poggio and Edelman, 1990; Logothetis *et al.*, 1994). Consistent with the suggestion that the view-independent representations are formed by combining view-dependent representations in the primate visual system, is the fact that in the temporal cortical areas, neurons with view-independent representations of faces are present in the same cortical areas as neurons with view-dependent representations (from which the view-independent neurons could receive inputs) (Hasselmo *et al.*, 1989a; Perrett *et al.*, 1987). This solution to "object-based" representations is very different from that traditionally proposed for artificial vision systems, in which the relative coordinates in three-dimensional space of the different features 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 three-dimensional 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, but would be less good for making actions in three-dimensional space to particular parts of, or inside, objects, as the three-dimensional coordinates of each part of the object would not be explicitly available. It is proposed, therefore, that visual fixation is used to locate in foveal vision part of an object to which movements must be made, and that local disparity and other measurements of depth then provide sufficient information for the motor system to make actions relative to the small part of space in which a local, *view-dependent*, representation of depth would be provided (cf Ballard, 1990).

The computational processes proposed above operate by an unsupervised learning mechanism, which utilizes regularities in the physical environment to enable invariant representations to be built. In some cases, it may be advantageous to utilize some form of mild teaching input to the visual system, to enable it to learn for example that rather similar visual inputs have very different consequences in the world, so that different representations of them should be built. In other cases, it might be helpful to bring representations together, if they have identical consequences, in order to use storage capacity efficiently. It is proposed elsewhere (Rolls, 1989a, 1989b) that the Backprojections from each adjacent cortical region in the hierarchy (and from the amygdala and hippocampus to higher regions of the visual system) play such a role by providing guidance to the competitive networks suggested above to be important in each cortical area. This guidance, and also the capability for recall, are it is suggested implemented by Hebb-modifiable connections from the backprojecting neurons to the principal (pyramidal) neurons of the competitive networks in the preceding stages (Rolls, 1989a, 1989b; Rolls and Treves, 1997).

The computational processes outlined above use distributed coding with relatively finely tuned neurons with a graded response region centred about an optimal response achieved when the input stimulus matches the synaptic weight vector on a neuron. The distributed encoding would help to limit the combinatorial explosion, to keep the number of neurons within the biological range. The graded response region would be crucial in enabling the system to generalize correctly to solve e.g. the invariances. However, such a system would need many neurons, each with considerable learning capacity, to solve visual perception in this way. This is fully consistent with the large number of neurons in the visual system, and with the large number of, probably modifiable, synapses on each neuron (e.g. 5000). Further, the fact that many neurons are tuned in different ways to faces is consistent with the fact that in such a computational system, many neurons would need to be sensitive (in different ways) to faces, in order to allow recognition of many individual faces when all share a number of common properties.

3. A NETWORK MODEL OF INVARIANT VISUAL OBJECT RECOGNITION

To test and clarify the hypotheses just described about how the visual system may operate to learn

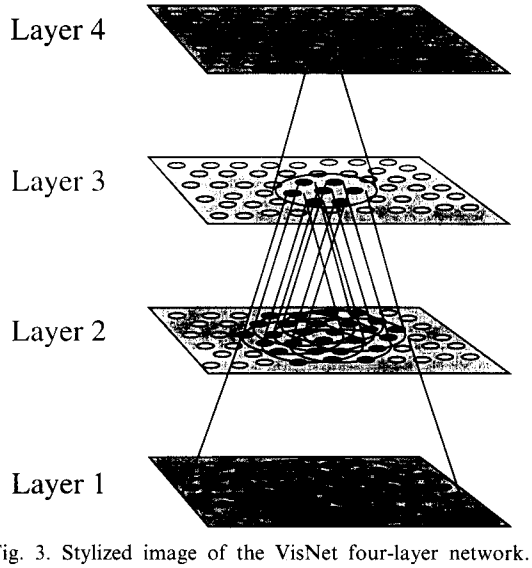
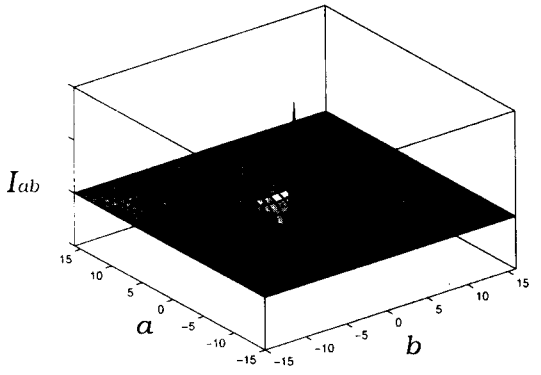


Fig. 3. Stylized image of the VisNet four-layer network. Convergence through the hierarchical network is designed to provide fourth layer neurons with information from across the entire input retina.



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

Fig. 4. Local lateral inhibition is implemented between nearby cells in a layer using this type of filter. For a given cell, the lateral inhibition is calculated based on the activity of nearby cells at distances away indexed by a and b. The parameters δ and σ are variables used to modify the amount and extent of inhibition, respectively.

invariant object recognition, Wallis and Rolls developed a simulation which implements many of the ideas just described, and is consistent with and based on much of the neurophysiology summarized above. The network simulated, visnet, can perform object, including face, recognition in a biologically plausible way, and after training shows for example translation and view invariance (Wallis *et al.*, 1993). The architecture and operation of this neural network are described next, for the simulation helps to define and test some of the hypotheses presented in Section 2 on how the cerebral cortex could perform invariant object recognition. We note that the most crucial part of the proposal is the use of the trace learning rule, described in Section 3.2.

3.1. VisNet Architecture

3.1.1. *Connectivity*

In the four-layer network, the successive layers correspond approximately to V2, V4, the posterior temporal cortex, and the anterior temporal cortex. The network is designed as a series of hierarchical, convergent, competitive networks. The network is constructed such that the convergence of information from the most disparate parts of the network's input layer can potentially influence firing in a single neuron in the final layer — see Fig. 3. This corresponds to the scheme described by many researchers (e.g. Van Essen *et al.*, 1992; Rolls, 1992b) as present in the primate visual system — see Fig. 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 neurons in adjacent 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 precludes the repeated connection of any pair of cells. Each cell receives 100 connections from a 32 × 32 grid of cells in the preceding layer, initially with a 67% probability that a connection comes from within four cells of the distribution centre — although the effective radius of convergence increases slightly through the layers. Figure 3 shows the general convergent network architecture used. Localization and limitation of connectivity in the network is intended to mimic cortical connectivity, partially because of the clear retention of retinal topology through regions of visual cortex. This architecture also encourages the gradual combination of features from layer to layer which has relevance to the binding problem, as described below*.

3.1.2. *Calculation of Neuronal Firing*

The activation h of each neuron in the plane of 32 × 32 neurons in each layer was calculated in the conventional way as the synaptically weighted sum of the input firings connected to each neuron,

$$h = \Sigma_j x_j w_j \quad (1)$$

where x_j is the firing rate of the jth input through the jth synaptic weight w_j to the neuron, and the sum is over all the inputs (indexed by j) connected to a neuron from the neurons in the preceding layer.

The next two steps implement short-range lateral inhibition between the neurons (performed in order to allow the neurons within a neighbourhood to reflect the strongest spatial information within that

*As neurons at the edge of the network would otherwise have fewer inputs the closer they are to the edge, in the actual network, simulated edge effects were precluded by wrapping the network into a toroid. This was performed by arranging the connections so that the top of the network was wrapped to the bottom, and the left to the right. This wrapping happens at all four layers of the network, and in the way an image on the "retina" is mapped to the input filters. This solution has the advantage of making all of the boundaries effectively invisible to the network. (This procedure does not itself introduce problems into evaluation of the network for the problems set, as many of the critical comparisons in VisNet involve comparisons between a network with the same architecture trained with the trace rule, or with the Hebb rule, or not trained at all, as described below.) In the real brain, such edge effects would be smoothed naturally by the transition of the locus of cellular input from the fovea to the lower acuity periphery of the visual field.

neighbourhood, and not to be suppressed by perhaps more active but distant neurons), and soft competition. The lateral inhibition helps to ensure that all parts of the stimuli presented are represented by the neurons in each layer. In the simulations, a local inhibitory function was applied to each neuron and its neighbouring cells, in a similar way to that used by von der Malsburg (1973). The local lateral inhibition was simulated via a linear local contrast-enhancing filter, consisting of a positive central spike surrounded by a negative gaussian field, the general shape and formula for which are given in Fig. 4. (As with the network connectivity, the inhibition acts toroidally.) The choice of parameters describing the mask $\delta = 10$ and $\sigma = 1$ meant that inhibition was largely restricted to the nearest neuronal neighbours. (In recent experiments by Milward and Rolls, using the sigmoid activation function, the range of the lateral inhibition was extended by increasing the value of σ to 1.4–4, and this improved the performance of VisNet.) The competition then applied was not winner-take-all (with only one neuron left active after the competition), but instead was graded, to produce a soft competitive network. This soft competition can be advantageous in the way neurons are allocated to stimuli (Bennett, 1990) and, in particular, has the important advantage of leading to distributed representations. After the competition, the average neuronal firing was scaled to a constant average value, to ensure that learning was similar for every presentation of a stimulus. The second step was, unless otherwise stated, implemented by raising the activity r of a neuron after the lateral inhibition to a power p greater than 1, and then rescaling the firing rates, to maintain the average firing rate of the neurons constant, i.e.

$$y = r^p/(\Sigma_i r_i^p) \tag{2}$$

where i indexes through the neurons in a layer, and would be represented in the brain by a shunting effect of inhibitory feedback neurons. In some simulations, an alternative activation function, a sigmoid, was used, as a check that the precise form of the competition was not crucial.

3.1.3. *Network Input*

In order that the results of the simulation might be made particularly relevant to understanding processing in higher cortical visual areas, the inputs to layer 1 come from a separate input layer which provides an approximation to the encoding found in visual area 1 (V1) of the primate visual system. 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 and Cooper, 1975; Linsker, 1986). VisNet does not attempt to train the response properties of simple cells, but instead starts with a fixed feature extraction level, as have some other researchers in the field (Hummel and Biederman, 1992; Buhmann *et al.*, 1991; Fukushima, 1980), with the intention of simulating the more complicated response properties of cells between V1 and the inferior temporal cortex (IT).

The response characteristics of neurons in the input layer are therefore provided by a series of spatially tuned filters with image contrast sensitivities chosen to accord with the general tuning profiles observed in the simple cells of V1. Currently, only even-symmetric (bar-detecting) filter shapes are used. The precise filter shapes were computed by weighting the difference of two Gaussians by a third orthogonal Gaussian according to the following:

$$\Gamma_{ab}(\rho,\theta,f) = \rho\left(e^{-\left(\frac{a\cos\theta + b\sin\theta}{\sqrt{2}^f}\right)^2} - \frac{1}{1.6}e^{-\left(\frac{a\cos\theta + b\sin\theta}{1.6\sqrt{2}^f}\right)^2}.e^{-\left(\frac{a\cos\theta + b\sin\theta}{3\sqrt{2}^f}\right)^2}\right) \tag{3}$$

where f is the filter spatial frequency (four frequencies over four octaves in the range 0.0625–0.5 pixels^{-1}), θ is the filter orientation (0–135° over four orientations), and ρ is the sign of the filter, i.e. ± 1*. Cells of layer 1 receive a topologically consistent, localized, random selection of the filter responses in the input layer, under the constraint that each cell samples every filter spatial frequency and receives a constant number (272 unless otherwise specified) of inputs. Oriented difference of gaussian filters were chosen in preference to the often used Gabor filter on the grounds of their better fit to available neurophysiological data including the zero DC response (Hawken and Parker, 1987; Wallis, 1994). (Any zero DC 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 by other neurons to the network.) The filter outputs also are normalized across scales to compensate for the low frequency bias in the images of natural objects. Figure 5 shows pictorially the general filter sampling paradigm.

3.2. The Trace Learning Rule

The learning rule implemented in the simulations utilizes the spatio-temporal constraints placed upon the behaviour of "real-world" objects to learn about natural object transformations. By presenting consistent sequences of transforming objects the cells in the network can learn to respond to the same object through all of its naturally transformed states, as described by Foldiak (1991), Rolls (1992b, 1994, 1995b, 1996b) and Wallis (1996b). 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 described here is intended in principle to enable learning of any of the transforms tolerated by inferior temporal cortex neurons (see above).

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

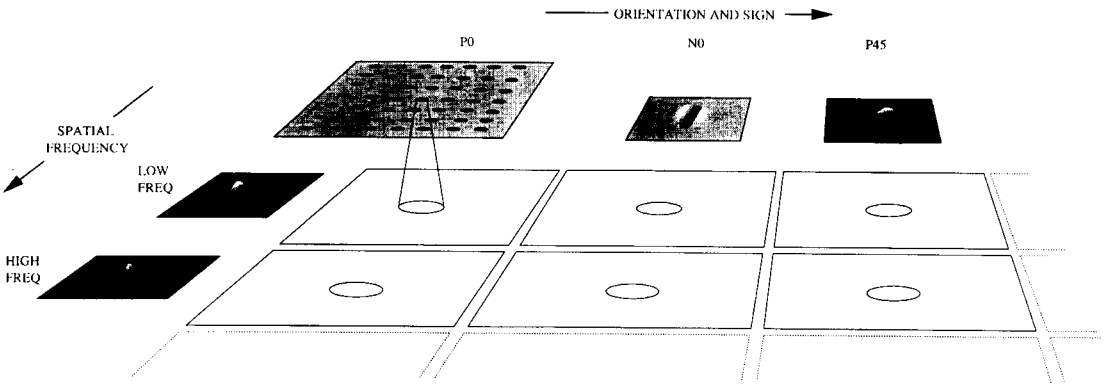


Fig. 5. The input filters for VisNet. There are four spatial frequencies and four orientations of the oriented difference of Gaussian — "bar detecting" — filters. Here, each square represents the retinal image presented to the network after being filtered by an oriented difference of Gaussian filter 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. From left to right, the orientation tuning increases from 0 in steps of four, with segregated pairs of positive (P) and negative (N) filter responses.

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 0.5 sec, 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. Making such associations works in practice because it is very likely that within short time periods different aspects of the same object will be being inspected. The cell will not, however, tend to make spurious links across stimuli that are part of different objects because of the unlikelihood in the real world of one object consistently following another.

The trace update rule used in these simulations is equivalent to both Foldiak's (1991) and to the earlier rule of Sutton and Barto (1981), and can be summarized as follows:

$$\delta w_{ij} = k\bar{y}_i^{(t)} x_j \tag{4}$$

where

$$\bar{y}_i^{(t)} = (1 - \eta)y_i^{(t)} + \eta\bar{y}_i^{(t-1)} \tag{5}$$

and x_j is the jth input to the neuron, y_i is the output of the ith neuron, w_{ij} is the jth weight on the ith neuron, η governs the relative influence of the trace and the new input (typically 0.4–0.6), and $\bar{y}_i^{(t)}$ represents the value of the ith cell's memory trace at time t. (The optimal value for η varies with presentation sequence length.)

To bound the growth of each cell's dendritic weight vector, the length of the weight vector of each neuron is explicitly normalized, a method in standard use for competitive nets (see Hertz *et al.*, 1991; Rolls and Treves, 1997). An alternative, more biologically relevant implementation, using a local weight bounding operation which utilizes a form of heterosynaptic long-term depression (see Brown *et al.*, 1990; Rolls, 1996a), has in part been explored using a rule similar to the Oja rule (see Oja, 1982; Hertz *et al.*, 1991; Rolls and Treves, 1997). The rule implemented for such tests was

$$\delta w_{ij} = k\bar{y}_i^{(t)}(x_j - w_{ij}) . \tag{4a}$$

This rule tends to keep the sum of the synaptic weights on each dendrite constant when the average firing rate of the inputs on the x lines is kept constant, as would be the case in the brain if the x inputs came from a population of cells with negative feedback operating through inhibitory feedback neurons (Rolls and Treves, 1997). This modified rule that performs automatic weight scaling implies long-term potentiation if x_j is greater than its average, and heterosynaptic long-term depression if x_j is below its average, for a given strong post-synaptic activation (Rolls, 1996a; Rolls and Treves, 1997).

To train the network to produce a translation invariant representation, one stimulus was placed successively in a sequence of (e.g. nine) positions across the input, then the next stimulus was placed successively in the same sequence of positions across the input, and so on through the set of stimuli. The idea was to enable the network to learn whatever was common at each stage of the network about a stimulus shown in different positions. To train on view invariance, different views of the same object were shown in succession, then different views of the next object were shown in succession, and so on.

3.2.1. *Measurement of Network Performance*

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 to all other stimuli irrespective of presentation location. Note that we state "set of stimuli" since neurons in inferior temporal cortex are not generally selective for a single stimulus but rather

Fig. 6. The three stimuli used in the first set of experiments on translation invariance by VisNet.

a sub-population of stimuli (Baylis *et al.*, 1985; Rolls and Tovee, 1995a; 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). One way to assess this was to run a two-way ANOVA on the set of responses of each cell, with one factor being stimulus type, and the other the position of the stimulus on the "retina". A high F ratio for stimulus type (F_s), and a low F ratio for stimulus position (F_p) would imply a position invariant representation of the stimuli. The "discrimination factor" of a particular cell was then gauged as the ratio F_s/F_p. This measure was supplemented in the simulations by a very analogous "relative amount of information" metric described in the Appendix, and by information measures identical to those used for real neurons by Tovee *et al.* (1994).

3.3. Translation Invariance with Simple Stimuli "T", "L" and " + "

A first test of the network used a set of three stimuli ("T", "L" and " + " shapes) based upon probable three-dimensional edge cues. [Chakravarty (1979) describes the application of these shapes as cues for the three-dimensional interpretation of edge junctions, and Tanaka *et al.* (1991) have demonstrated the existence of neurons responsive to such stimuli in the inferior temporal visual cortex.] The actual stimuli used are shown in Fig. 6. 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 cannot 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 by Foldiak (1991). The expectation is that layer 1 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"-shaped 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 unless otherwise stated. The sequence described above was followed for each stimulus, with the sequence start point and direction of sweep being chosen at random.

It was found following training with the trace rule that some layer 4 neurons responded to one of the stimuli whatever its location, with only small responses to the other stimuli (see examples in Fig. 7). These invariant responses were built up gradually over the layers of the network, with neurons in layers 2 and 3 starting to show some translation invariance (see examples in Fig. 8), which was necessarily limited to part of the "retina" because the convergence allowed by the architecture did not allow full translation invariance at these early stages (see Fig. 3). In layer 1, the neurons typically responded to inputs from an even more limited area of the "retina", but combined inputs appropriately from the different spatial filters in a region. (For example, after training, neurons in layer 1 frequently came to respond to inputs from spatial frequency filters of similar orientations but different spatial frequencies

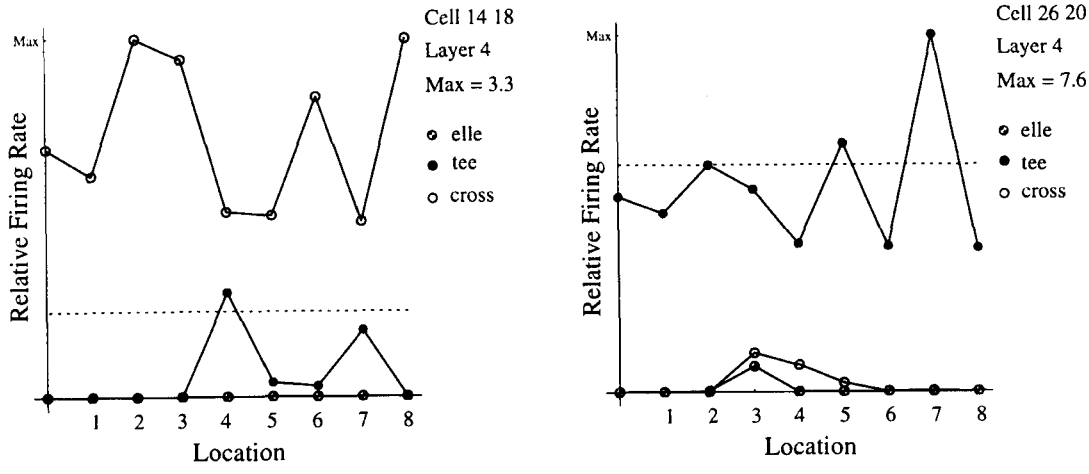


Fig. 7. Response profiles for two fourth layer neurons. The discrimination factors were 4.07 and 3.62. The firing rates for each of the three stimuli (L, T and +) at each of the nine training locations are shown.

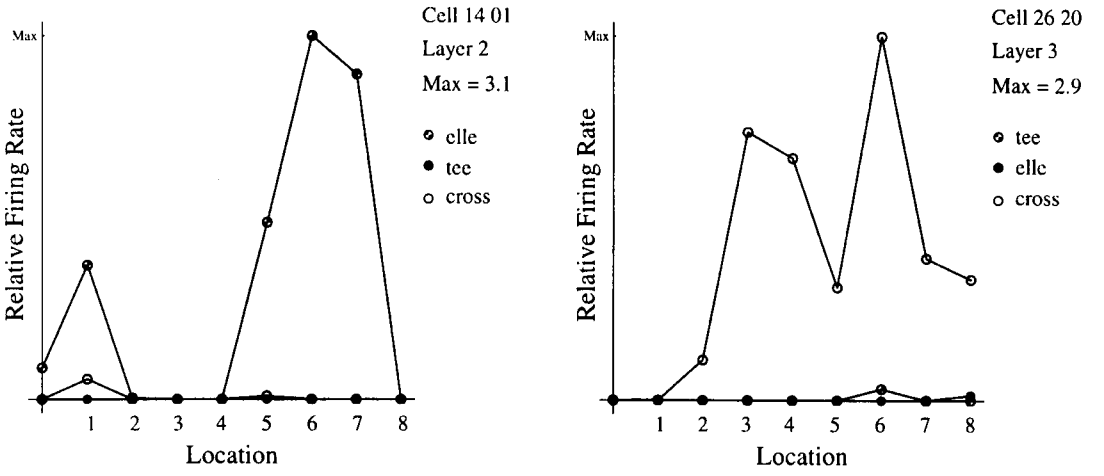


Fig. 8. Response profiles for a neuron in layer 2 (left, discrimination factor 1.34) and for a neuron in layer 3 (discrimination factor 1.64).

over a small retinal area, such as might be produced by an oriented edge at one position on the "retina".) The gradual emergence from layer to layer of the network of translation invariance over the whole "retina" is documented in Fig. 9, which shows the discrimination factor for the 30 most invariant cells in each of the four layers of the network. The values of discrimination factor in the range 2–5 reached by neurons in layer 4 indicate excellent translation invariant discrimination between the patterns, as can be seen by comparison with the values of the discrimination factor of the neurons shown in Figs 7 and 8. It is useful to note that because VisNet operates as a competitive network, it is expected, and desired, that only some of the neurons provide a good representation of the input stimuli: the other neurons remain unallocated, available for further patterns to be learned later.

It was next shown that use of the trace learning rule was essential for the invariant representations found in single neurons in layer 4 of VisNet. This was shown by testing the network under two new conditions. Firstly, the performance of the network was measured before learning occurs, that is with its initially random connection weights. Second, the network was trained with η in the trace rule set to 0, which causes learning to proceed in a traceless, standard Hebbian fashion. Figure 10 shows the results under the three 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 have discrimination factors which are barely higher than 1. This value of discrimination factor 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

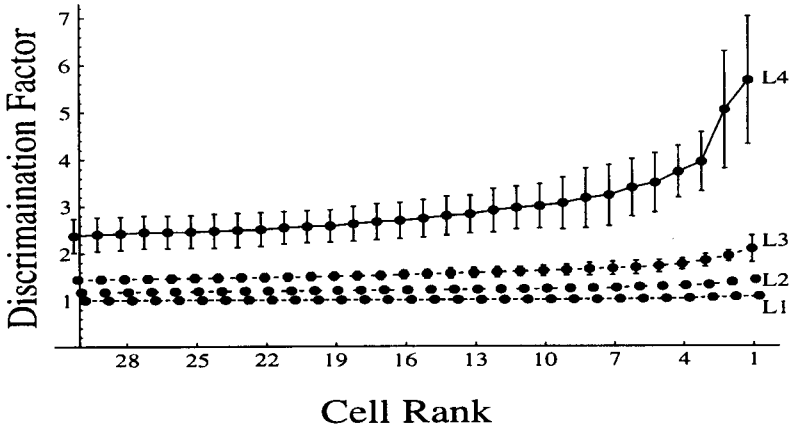


Fig. 9. The network discrimination factor for each of the 30 best cells in each of layers 1–4 (L1 etc) are shown. Training was with the trace learning rule, with three stimuli, +, T and L, at nine different locations The means and standard errors of five separate runs of the network are shown.

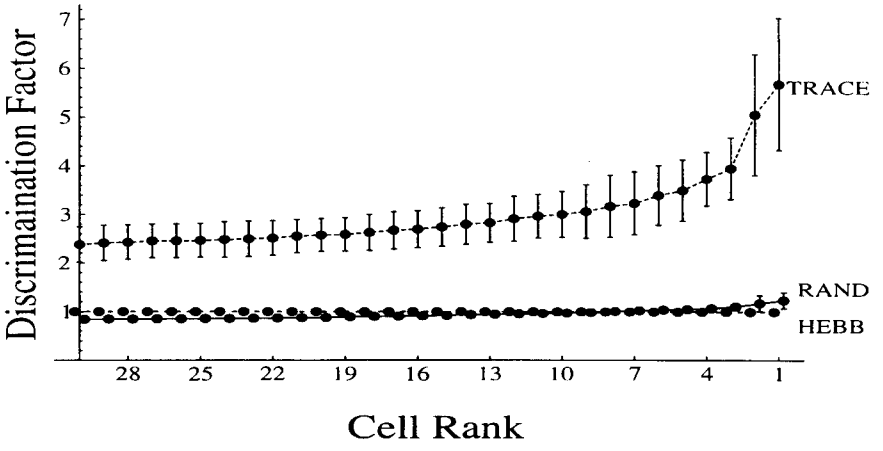


Fig. 10. Comparison of network discrimination when trained with the trace learning rule, with a Hebb rule (No trace), and when not trained (Random) on three stimuli, +, T and L, at nine different locations. The means and standard errors of five separate runs of the network are shown. The values for the 30 most discriminating cells in layer 4 are shown.

competitive network using just Hebbian learning will thus tend to categorize images by *where* they are rather than what they are — the exact opposite of what the net was intended to learn. This result is important, since it indicates that a small memory trace acting in the standard Hebbian learning paradigm can alter radically the normal vector averaging, image classification, performed by a Hebbian-based competitive network.

A property of competitive networks is that they should allocate at least some neurons to each input stimulus, so that the output of the competitive net provides information about all input stimuli (see e.g. Hertz *et al.*, 1991; Rolls and Treves, 1997). It was confirmed that VisNet did allocate neurons to each of the input stimuli on which it was trained with the trace rule. (One simple way that this was shown was by finding the preferred stimulus for each cell, and adding its discrimination factor value to a total for each stimulus. This measure in practice never varied by more than a factor of 1.3: 1 for all stimuli.)

Although the global translation invariance achieved by VisNet is a result of the trace rule enabling neurons to set up invariant representations over large shifts in the stimulus position (see Figs 7,10), there is in addition some local generalization to untrained locations near to trained locations, which arises because the input filters illustrated in Fig. 5 have, particularly for the low spatial frequency filters, a significant receptive field size.

Having established that VisNet could learn translation invariant representations, we next investigated how two parameters of the network, the length or time constant of the trace in the learning rule [set by η in equation (5)], and the nonlinearity of the activation function [set e.g. by p in equation (2)], affected the performance of the network. (For these investigations, the network performance is shown by the median and interquartile range of the discrimination factor for the best 16 cells of the fourth layer for each parameter value from a run with 800 training trials. The network was again trained on the three stimuli T, L and +, at each of nine locations placed as before at horizontal, vertical coordinates on the retina −30,30; −30,0; 30,30; 30,0; 0,0; 30,0; −30, −30; 0, −30; and 30, −30.)

The effects of varying the effective length of the trace, controlled by the parameter η, with larger values implementing a longer trace, are shown in Fig. 11. It is shown that values of 0.6–0.8 are best for training with the nine standard presentation locations. A large value of η will have the advantage of allowing stimulus presentations far apart in the run of nine locations to be associated together, but the disadvantage that then there will be some spurious association between stimuli, in that when a new stimulus is chosen during training, for its first few presentations, some trace activity will persist from the previous stimulus. (No explicit trace reset when a new stimulus is presented is used in VisNet, to simulate a "worst case" condition. It is, of course, possible that in the brain, if the eyes are shifted to a new object, there is some resetting produced by factors such as saccadic suppression and masking of previous neuronal activity produced by what might be a completely new visual input produced when the eyes saccade to a new object. Such resetting would serve to improve the performance of the network.) The optimal length of the trace for the reason just discussed is likely to depend on the number of presentations of each stimulus in any one run, before a different stimulus is shown. This is confirmed in Fig. 12, which shows that a smaller value of η, 0.4–0.6, is better if the run length is 5, with each stimulus being shown in five different locations before the next stimulus is chosen. In fact, the way that η operates in equation (5) implies exponential decay of the trace, and this has been shown to be close to optimal when the system must operate with different run lengths for any one stimulus before another stimulus is shown (Wallis, 1996a). It is also the case that the optimal value of η may be different for each layer of VisNet, at least for translation invariance with steady progression across the retina. This arises because the neurons in each layer have different effective receptive field sizes, so that each layer's

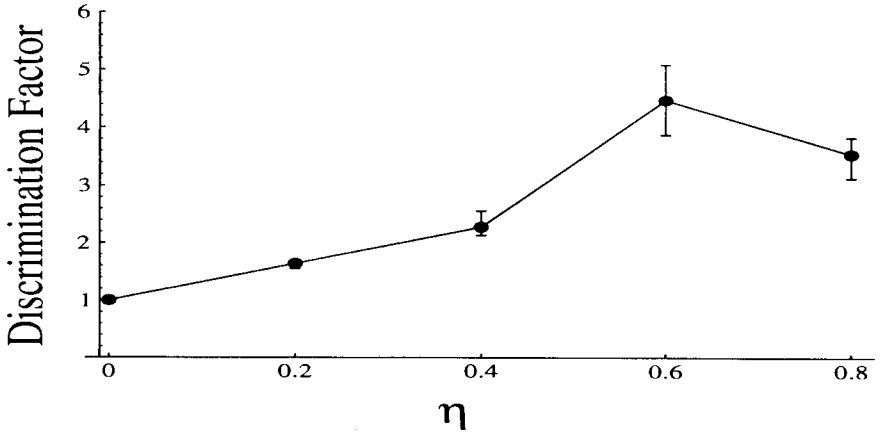


Fig. 11. The effects of varying the effective length of the trace, controlled by the parameter η, with larger values implementing a longer trace, on the performance of VisNet trained with T, L and + stimuli in each of nine different locations.

neurons are exposed to different proportions of the full sweep of a particular stimulus across the retina. This indicates that the optimal value of η will grow through the layers, with training on this type of problem. In fact, the receptive fields of neurons in layer 1 of VisNet are so small that there is only little scope for invariance learning in layer 1, and in practice it is found that VisNet operates well with only Hebbian learning in layer 1 (η set to 0), allowing neurons in layer 1 to learn to respond to combinations of simultaneously active filter inputs, without encouraging them to learn invariant representations. In accordance with this, η is set to 0 for layer 1 for the simulations described here. Unless otherwise stated, it was set to 0.6 for layers 2–4 of VisNet.

The effect of altering the degree of non-linearity of the activation function of the neurons, which controls the strength of the competition between the neurons, is shown in Fig. 13. High values of p in equation (2) tend to make the network winner-take-all, with one neuron left active after the competition, while lower values (e.g. 2) tend to produce a much more distributed representation, with many neurons left active after the "soft" competition. It is shown in Fig. 13 that a value of 2 for the non-linearity power p was optimal for layers 2–4. To provide a quantitative measure of the sparseness of the representation with which VisNet operated well, we calculated a measure of the population sparseness as

$$a_p = (\Sigma_{n=1,N}\, r_n/N)^2/\Sigma_{n=1,N}\, (r_n^2/N) \qquad (6)$$

where r_n is the firing rate of the nth neuron in the population of N neurons in a layer. This population sparseness has a maximum value of 1.0 if all the neurons are equally active when a stimulus is shown, and a minimum value of $1/N$ if only one neuron in the population of N neurons is active, that is if there is one winner. The value for N was 1024 neurons per layer for the simulations described, so that a value of 0.001 would correspond to only one neuron active in a layer. With a value for p of 2 for layers 2–4, typical values for a_p were 0.1–0.3.

As the inputs to layer 1 of VisNet were from spatial frequency filters that were simply convolved with the image, and had no mutual inhibition, a greater degree

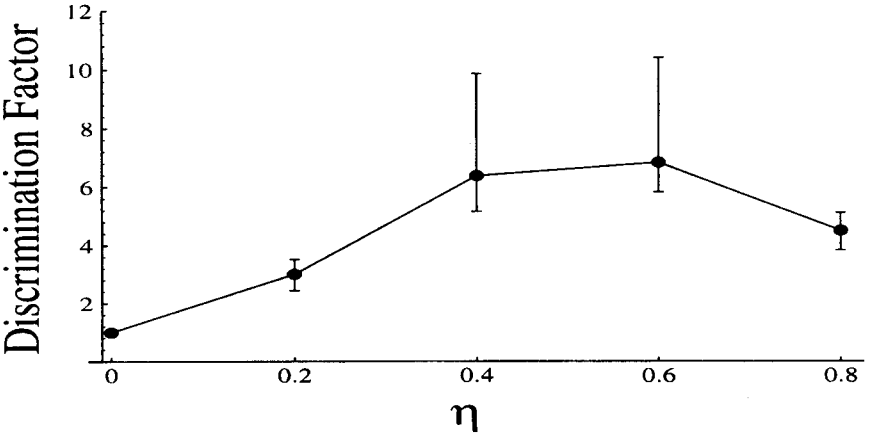


Fig. 12. The effects of varying the effective length of the trace, controlled by the parameter η, on the performance of VisNet trained with T, L and + stimuli in each of five different locations.

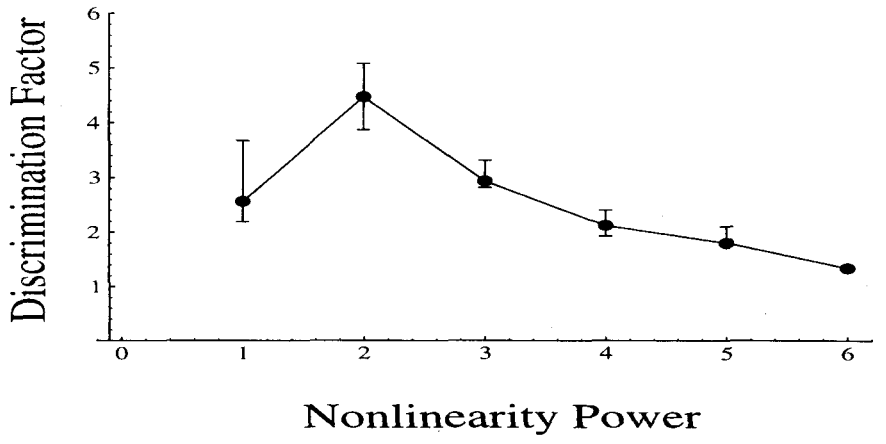


Fig. 13. The effect of altering the degree of non-linearity of the activation function of the neurons, which controls the strength of the competition between the neurons.

of non-linearity was required in the competition between the neurons of layer 1 to bring down the sparseness of the representation in layer 1 to appropriate values. The non-linearity power p of equation (2) used for layer 1 was typically 6, and this produced a sparseness a_p which was typically in the region of 0.02 for layer 1.

To ensure that there was no special dependence on the type of activation function and competition implemented between the neurons, some runs were performed with another activation function with a physiologically plausible shape, a sigmoid:

$$1/(1 + e^{-2\beta(x-\alpha)}) \quad (7)$$

where α is a threshold, β is the slope, and x is the activation. To apply this activation function, α was set to the 98th percentile of the activations in that layer. (This procedure results in 98% of the firing rates being below the mid-point of the sigmoid, 0.5.) The slope β was set to a fixed value for each layer which resulted in approximately 4% of the firing rates lying within or above the linear part of the sigmoid. Comparably good performance to that already described was obtained with the use of this sigmoid activation function.

The VisNet simulations just described with three simple stimuli provided a useful test case for performance of the network. We next tested whether the network could operate with much more complex, real biological, stimuli, faces, which in many cases were the same as those used as stimuli in the neurophysiological experiments on the temporal cortical visual areas described above; and whether the architecture could learn other types of invariance, such as view invariance.

3.4. Translation Invariance with Faces

VisNet was trained with seven faces each in nine locations on the retina. The set of face images used is shown in Fig. 14. In practice, to equalize luminance, the DC component of the images was

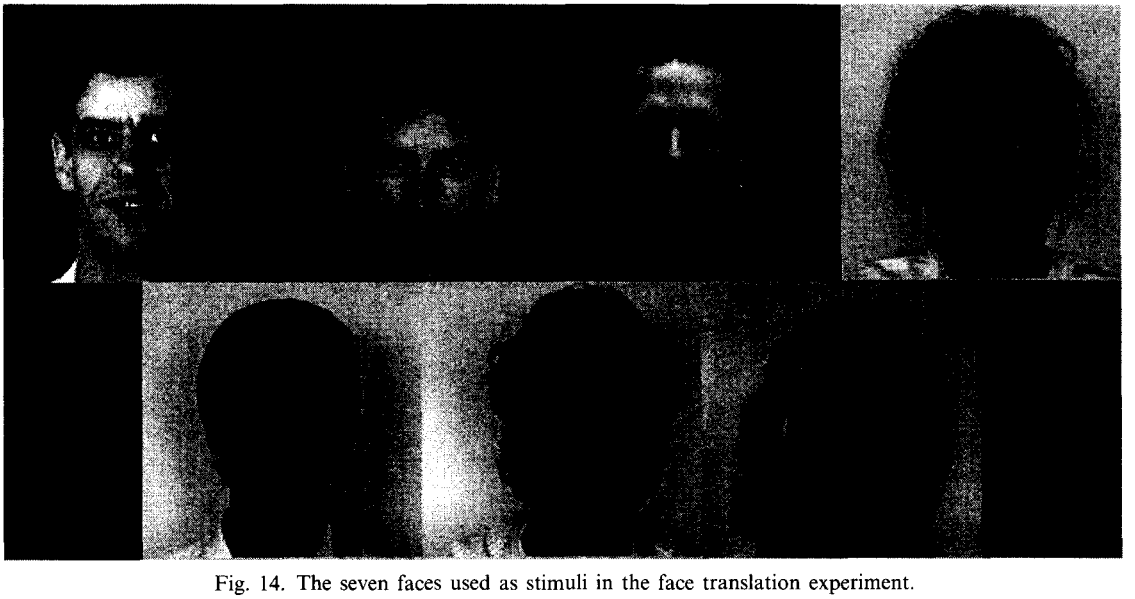


Fig. 14. The seven faces used as stimuli in the face translation experiment.

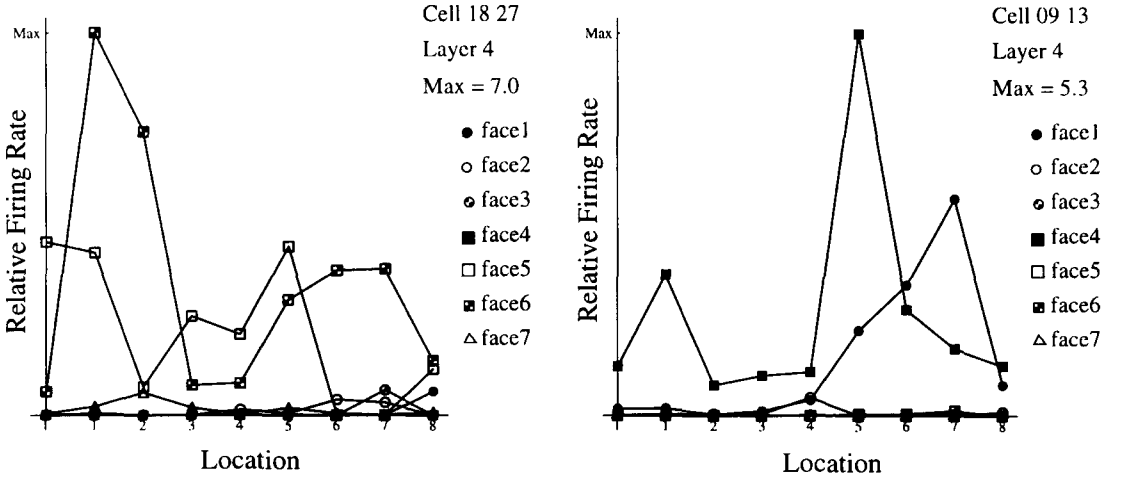


Fig. 15. The response profiles for two neurons in the fourth layer after training with seven faces at each of nine locations. Their discrimination factors were 2.64 and 2.10.

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 are shown in Figs 15 – 17. The network produced neurons with high discrimination factors, and this only occurred if it was trained with the trace rule (Fig. 17). A difference from the previous simulations was that with more stimuli, the neurons did not typically respond to only one stimulus independently of location, but instead a more distributed representation was found, as illustrated in the examples of layer 4 neurons shown in Fig. 15.

To check that information was present in the type of distributed representation found in layer 4 that could be decoded easily from many neurons to indicate which stimulus was presented independently of location, 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 (i.e. a delta rule, with one neuron in the fifth layer for each stimulus). (The fifth layer was intended purely as a tool for analysis and for decoding the representation found in the fourth layer of the network. 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 layer 5 should not be able to extract information about individual stimulus identity.) Fig. 18 shows the classification performance of the fifth layer for nets trained with the Hebb and trace rules as well as for the untrained net. Performance on every stimulus was perfect for nets trained with the trace rule, confirming that information about stimulus identity was present in the representation built by the trace rule by layer 4. In contrast, the performance, although quite good because of the power of the supervised learning used in layer 5, was not perfect for the Hebb trained net (84%), or for the untrained net (92%). Further evidence that all the stimuli were well represented by layer 4 was that taking the preferred stimulus for each cell, and adding its discrimination factor value to a

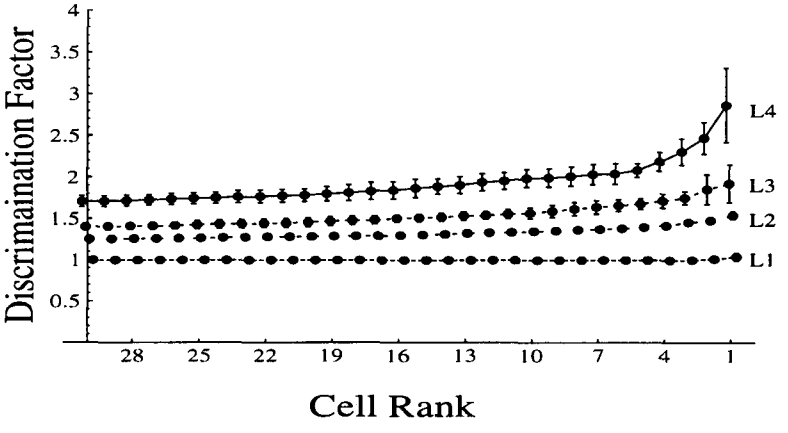


Fig. 16. The network performance shown by that of the 30 most highly discriminating cells for each of the four layers of the network, averaged over five runs of the network. The training set was seven faces at each of nine locations.

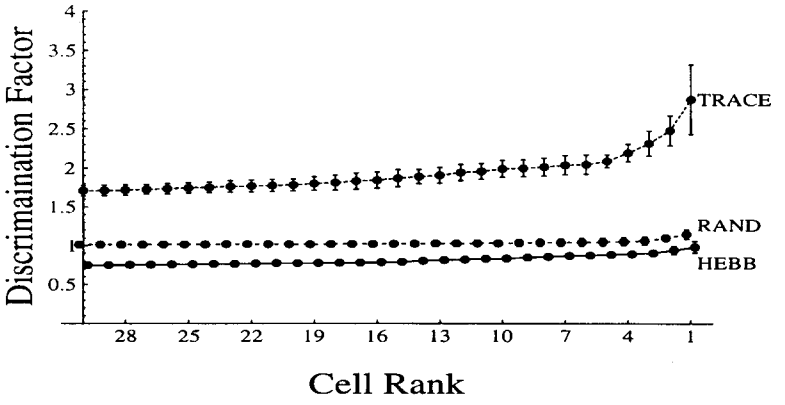


Fig. 17. The network performance shown by that of the 30 most highly discriminating cells in the fourth layer for the three training regimes, averaged over five runs of the network. The training set was seven faces at each of nine locations.

total for each stimulus, produced high discrimination scores for every stimulus when the net was trained with seven faces at each of nine locations.

3.5. View Invariance

To investigate how well VisNet might solve other types of invariance, the network was trained on the problem of three-dimensional stimulus rotation, which produces non-isomorphic transforms, to determine whether the network can build a view-invariant representation of 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 its different transformed states close together in time. Seven different views of three different faces presented centrally on the retina were used. The images used are shown in Fig. 19. The faces were

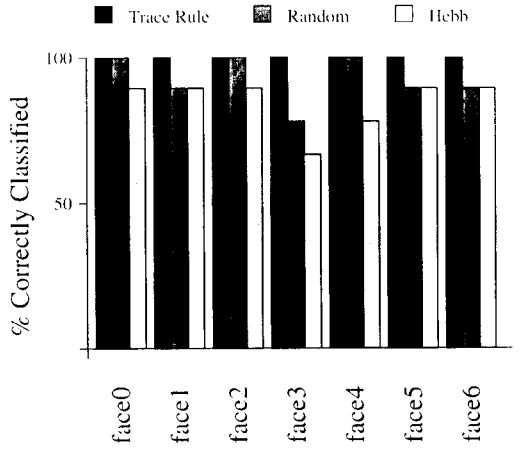


Fig. 18. The stimulus classification achieved for a fifth layer trained with a delta rule to classify faces based on the representation set up by VisNet in layer 4 when trained with the trace or Hebb rule, or not trained (Random). There were seven faces each shown in nine different locations.

again smoothed at the edges to erase the harsh image boundaries, and the DC term was removed. To use the capacity of the network fully, given that the images were presented only centrally, 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. 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 clockwise or counter-clockwise.

The net was able to solve the view invariance problem. Examples of invariant layer four neuron response profiles are shown in Fig. 20. One difference from the results of the translation invariance experiments was that some cells in the first layer showed limited 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 generalization. Although true generalization across all views was not achieved until higher layers, the contribution of local generalization provided by the cells in layer one for this problem meant that some cells in layer three already exhibit view 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 (see Fig. 21, where there are more cells with high discrimination factors and flatter response profiles than those observed in layer three). Figure 22 shows that only the net trained with the trace rule can solve the problem, and that the nets trained with the Hebb rule or untrained (random connectivity) perform equally poorly. Because there were fewer stimuli (seven) in a view set than in a translation invariance set, it was found that the optimal value of the trace parameter, η, was a little lower than the 0.6 found to be optimal for nine locations.

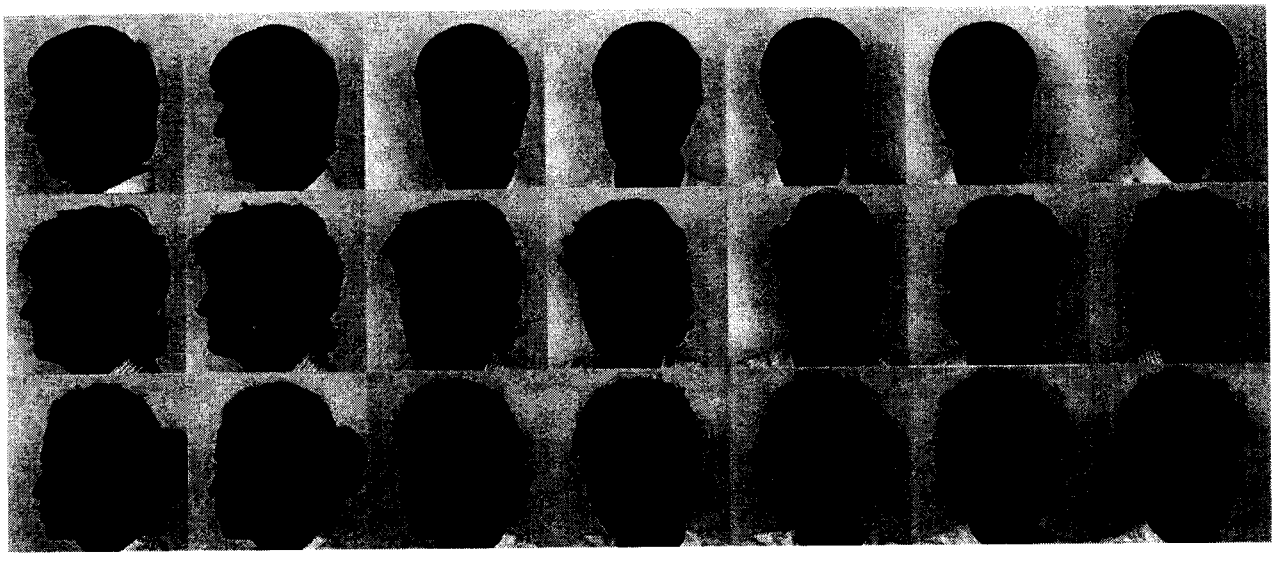


Fig. 19. The three faces each with seven different views used as the stimuli in the view invariance learning experiment. The goal of the net was to learn to recognize each face independently of view.

3.6. Size Invariance

VisNet also has been trained successfully to produce size invariant representations. In one experiment, by M. Elliffe and Rolls, VisNet produced perfect discrimination of seven faces each trained with seven different sizes (from 1/4 to 7/4 of the normal size).

3.7. Translation Invariance with Seven Faces and 49 Training Locations

In recent simulations, Rolls and T. Milward have extended the analysis of VisNet to investigate whether it can still form invariant responses when there are many more locations over which it must show translation invariant representations of objects such as faces. In one such investigation, Rolls and Milward trained VisNet on seven faces shown at each of 49 training locations. Each face was a 32 × 32 pixel image with 256 grey scale values. The image was presented in each of 49 locations in a 64 × 64 part of the retina during training. With a value of $\eta = 0.6$, the trace effect remaining from a previous presentation of a stimulus decays to a small value after the stimulus image has been presented in seven different retinal locations. Therefore, they did not present the 49 locations for any one image during training in a standard serial sequence, but instead used a set of short-range movements across the retina, followed by a longer jump. The idea here is that during inspection of an object during learning, there is a set of small eye movements, followed by a longer saccade to another part of the object, which occurs several times. In detail, the sequence of presentations of any image consisted of seven small

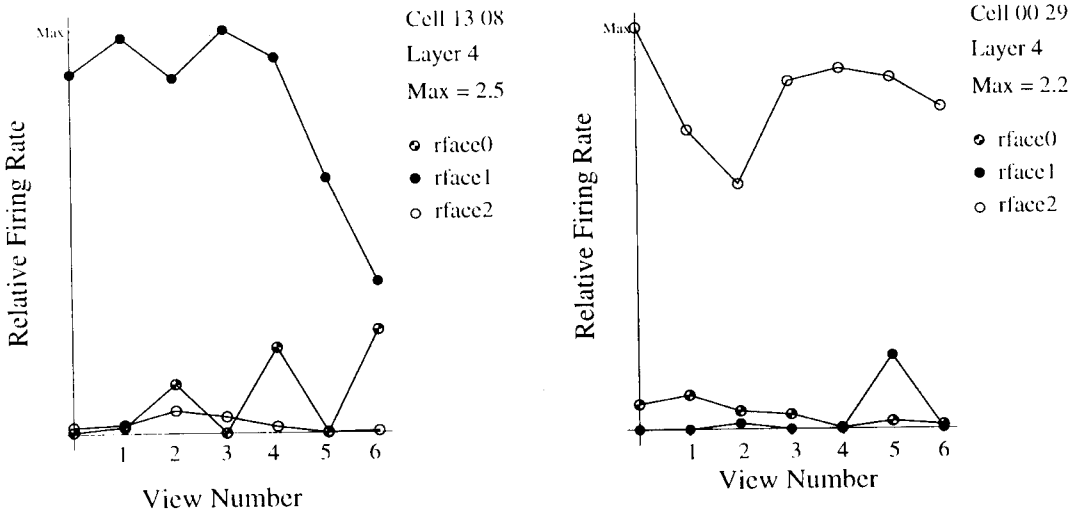


Fig. 20. Response profiles for cells in the layer 4 of VisNet when trained on three faces each with seven different views. The discrimination factors of the cells were 11.12 and 12.40.

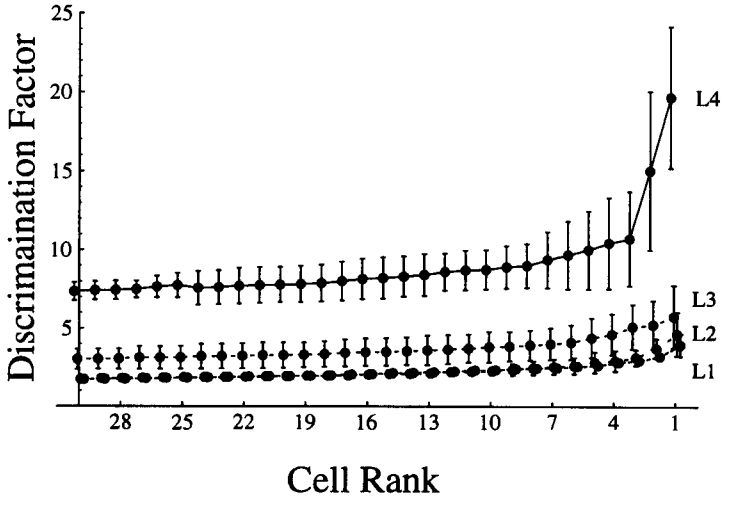


Fig. 21. The discrimination factors for the 30 best cells in each of the four layers L1–L4 of VisNet trained on three faces each with seven views. The means and standard errors based on five runs are shown.

movements to adjacent testing locations (which were arranged in a pattern which consisted of seven rows of seven points in the 64 × 64 grid), followed by a random long jump to another training location to start on another set of small movements. Each of the 49 locations was visited once per training epoch for each image. 2000 such training trials were run for each layer. During testing, each face was presented at the 49 training locations, and the responses of the cells in layer 4 were measured to determine whether they showed responses which displayed selectivity for one of the faces but invariance with respect to where that face was shown. Trace reset between stimuli was used for these and later simulation runs.

The results of training VisNet on seven faces at 49 locations are shown in Fig. 23. The discrimination factor for the 30 most translation invariant cells in layer 4 when VisNet was trained with the trace rule, or was untrained (random weights) as a control, are shown. The results from one of the cells are shown in Fig. 24. This cell responded to one of the faces only at all locations, and not to any of the other faces at any location. Figure 25 shows the results from the same simulation expressed as the amount of information in bits about which of the seven faces had been shown (calculated across all 49 training locations) represented by the 30 most selective cells. (The application of information theory to analyse translation invariant neuronal responses has been described by Tovee *et al.*, 1994.) The results are shown separately for VisNet trained with the trace rule, and left untrained as a control with random weights. The network could also generalize to other locations at which it was not trained, as shown in Fig. 26, which shows performance when tested at all 1024 locations after training at 49 locations. These results show that this architecture can still perform reasonably at the very difficult task of learning

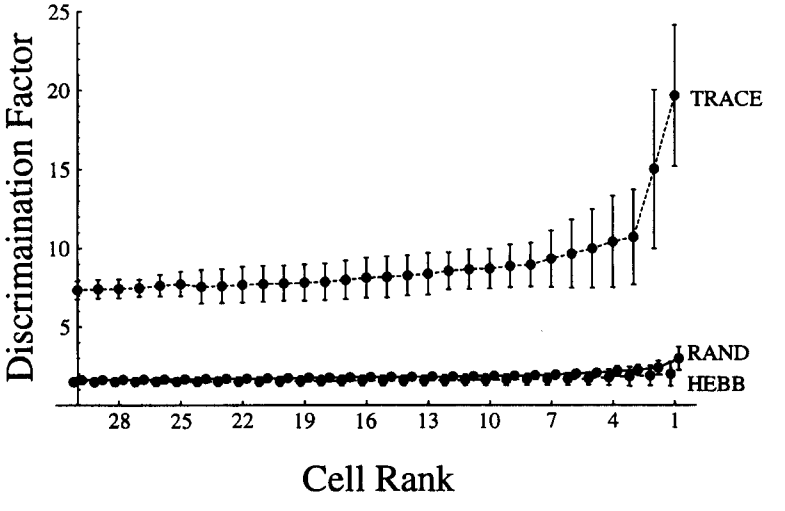


Fig. 22. The discrimination factors for the 30 best cells in layer 4 of VisNet with three faces each with seven views. The means and standard errors based on five runs are shown when training was with the Trace or Hebb rule, or the network was left untrained with its random initial connectivity.

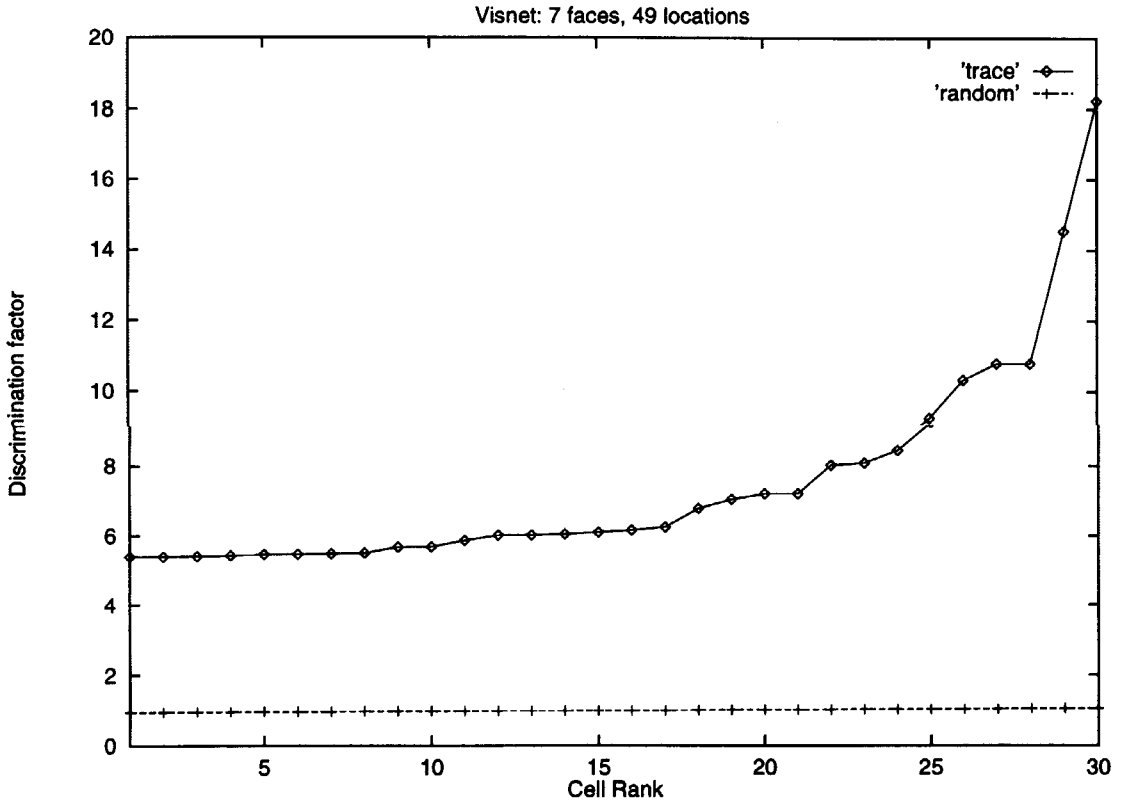


Fig. 23. The results of training VisNet on seven faces at 49 locations. The discrimination factor for the 30 most translation invariant cells in layer 4 when VisNet was trained with the trace rule, or was untrained (random weights) as a control, are shown.

translation invariant representations over 49 training locations of seven different complex images, faces. To enable it to learn, the eyes had effectively to jump to different parts of the object several times, so that the trace rule could make associations not just over short distances across the retina, but also (using its higher layers) over longer distances across the retina. The network also can perform reasonably at the difficult task of learning seven faces when trained with every face shown in every one of 1024 locations (Fig. 27).

4. COMPARISON OF DIFFERENT APPROACHES TO INVARIANT OBJECT RECOGNITION

The findings described in Section 3 show that the proposed trace learning mechanism and neural architecture can produce cells with responses selective for stimulus type with considerable position, view and size invariance. We now compare to other approaches the proposal made here and by Rolls (1992b, 1994,1995, 1996) and investigated by simulation using VisNet, about how the visual cortical areas may solve the problem of forming invariant representations.

The trace rule is 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 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 typically have combined various less attractive elements such as supervised or non-local learning (Poggio and Edelman, 1990; Fukushima, 1980; Mel, 1996), extremely idealized or simplified stimuli (Foldiak, 1991; Hinton, 1981), prohibitive object by object matching processes (Olhausen *et al.*, 1993; Buhmann *et al.*, 1990); or non-localized connectivity (Hummel and Biederman, 1992). On the other hand, some of these models have some advantages over the model described here. For example, the model of Olhausen *et al.* (1993) more explicitly addresses the issue of locating and attending to objects throughout the visual field. The model described here only addresses object recognition within the high acuity centre of the visual field, and would require some other mechanism for locating and fixating stimuli. It is possible that in the brain this process is not performed by the ventral visual system, but is instead performed by the dorsal visual system. Another issue about the current simulation, VisNet, is that it has only been trained with relatively few stimuli. The models of Mel (1996) and Fukushima (1980) have e.g.

been successfully trained on much larger data sets. However, in work in progress, VisNet has been successfully trained to 96% accuracy on a set of 100 hand-written digits (Wallis, 1994; Wallis, Rolls and Milward, in preparation), and other investigations using the trace rule have shown better categorization performance than Fukushima's NeoCognitron, and better performance than the delta rule on a cross-validation test (Wallis, 1996b).

One important aspect of any model of invariant shape processing is that it should address the feature binding problem. The essence of the problem is that the local spatial arrangements of features should be conserved, but the system must respond to these local spatial arrangements (which together might define an object) independently of where they are. Real neurons do solve the problem, in that neurons responsive to faces (Perrett *et al.*, 1982, 1992), or objects (Tanaka *et al.*, 1991) respond less when the features are jumbled. Models which throw away the relative spatial arrangement of features so as to achieve translation invariance will run into the problem of false "recognition" of stimuli in which the features have been rearranged. This is certainly true for models which attempt to learn invariance in one stage (Mel, 1996; Cavanagh, 1978). For example, in a recent paper, Mel (1996) records that in his model the cells "recognize" stimuli as the original configurations even when the features are jumbled. An active dynamic linking of features is proposed by von der Malsburg (1981, 1990) and von der Malsburg and Schneider (1986) as one solution to this problem. The solution proposed by Rolls 1992b), 1994),1995, 1996 and incorporated into the network described here, is that competitive learning should allow neurons to learn to respond to *combinations* of their inputs. Given that their inputs are restricted spatially within the network, only low-order combinations of spatially arranged features or inputs are learned by each neuron. The suggestion is that with the redundancy present in the real world (i.e. the world does not consist of random arrangements of pixels, but instead has regularities such as edges and combinations of edges), a multilevel architecture operating with the same competitive scheme repeated at each level, can learn sufficient about the world to represent its local spatial properties, in such a way that rearrangements of the features will not lead to the same response. The multilevel architecture helps in this process, by enabling only short range (in terms of connection space) spatial arrangements to be learned at any one stage, even though by higher layers of the network these may represent combinations of features present over considerable parts of the input layer.

Part of the reason for simulating the network described here was to determine whether invariant representations which retain information about the local spatial arrangement of features can be learned with real images, provided as examples of the real statistics present in the world. The results of the simulation do in fact support this proposal. For example, the network was able to discriminate the "T", "L" and " + " stimuli, which, after all, are

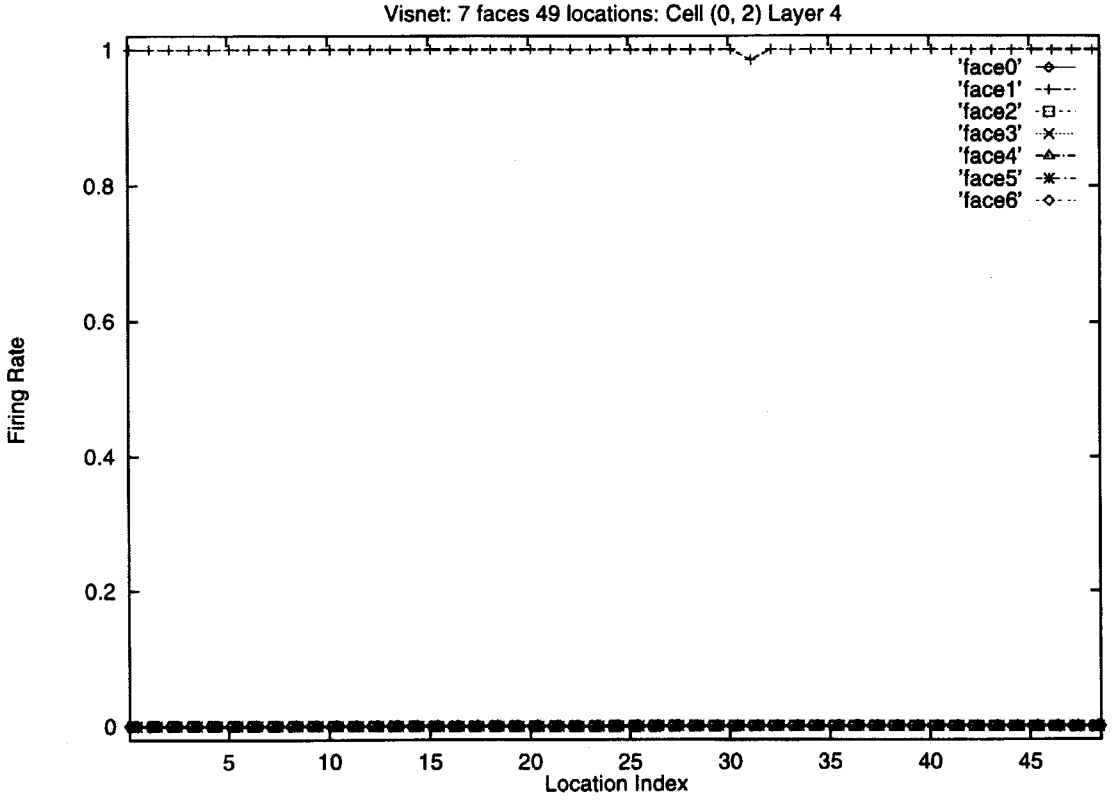


Fig. 24. Response for a cell in layer 4 of VisNet when trained on seven faces in each of 49 locations.

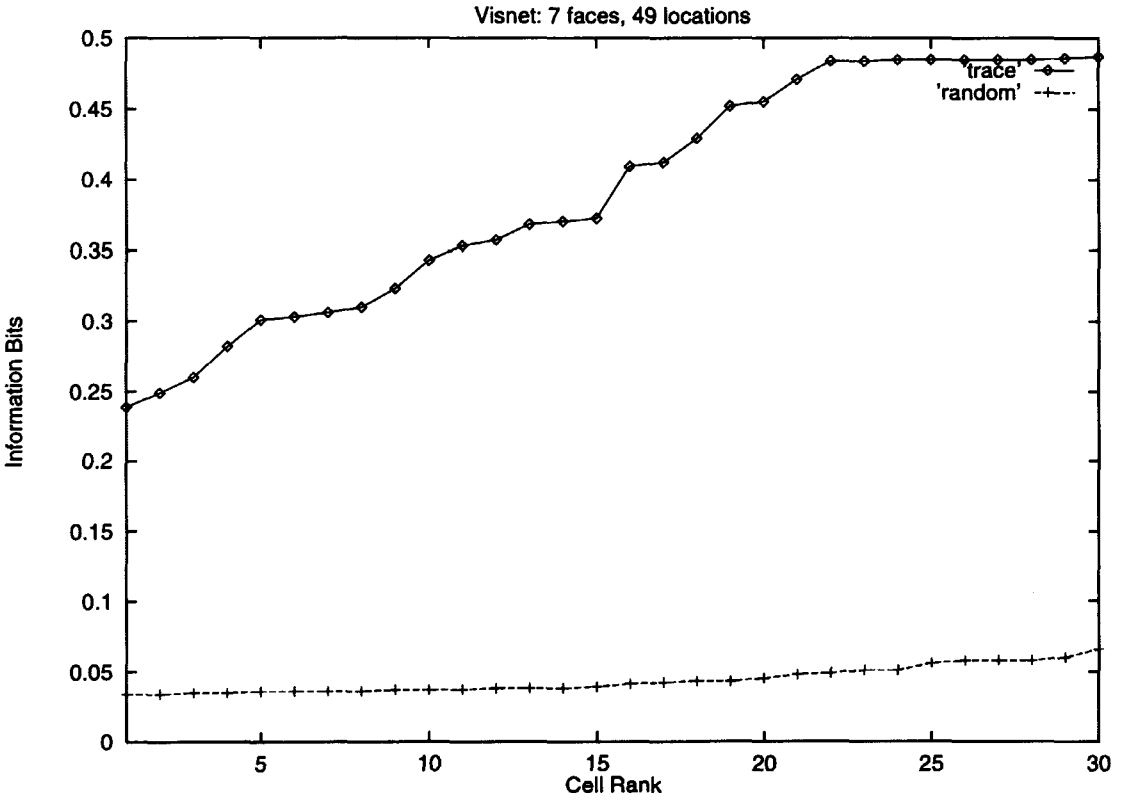


Fig. 25. The results from the same simulation as that shown in Fig. 23 expressed as the average amount of information in bits about which of the seven faces had been shown (calculated across all 49 training locations) represented by each of the 30 most selective cells. The results are shown separately for VisNet trained with the trace rule, and left untrained as a control with random weights.

merely rearrangements of two bar features. The whole spirit of this proposal is very much in line with the discrimination exhibited by temporal cortex neurons to jumbled features.

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.*, 1989a; 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 and van Doorn, 1979; Tarr and Pinker, 1989; Bülthoff and Edelman, 1992), and the network described here reveals how such a representation might be set up, without recourse to gradient descent algorithms used in other models (Poggio and 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 three-dimensional space of descriptors 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 three-dimensional 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 (1994, 1995b). 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 one-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. An important part of the hypothesis is that invariant properties (e.g. feature combinations) 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. In line with this, learning a representation of a new object can be fast in a system already trained on other objects, for new feature learning is unlikely to be required in intermediate layers, so that it is only necessary to link neurons in intermediate layers to new neurons in or close to the final layer of the system.

One aspect of the model that has not been treated

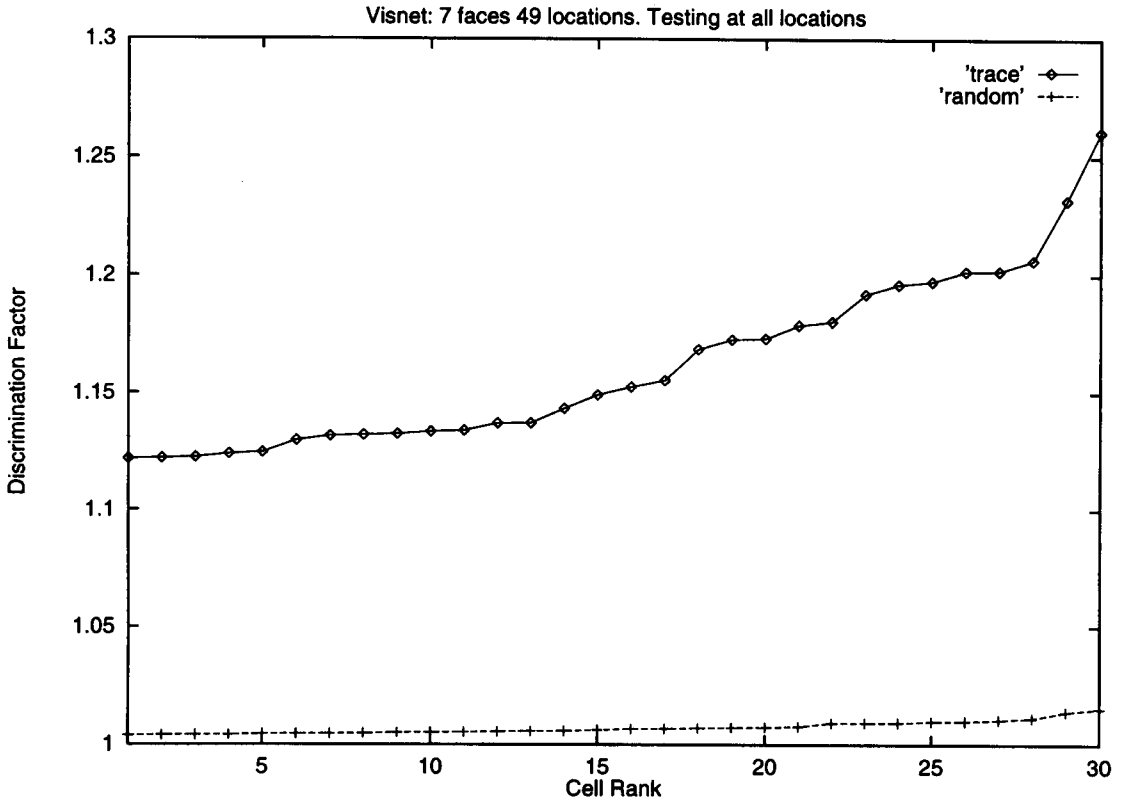


Fig. 26. The results of training VisNet on seven faces at 49 locations, and then testing at all 1024 locations. The discrimination factor for the 30 most translation invariant cells in layer 4 when VisNet was trained with the trace rule, or was untrained (random weights) as a control, are shown. This shows generalization to locations at which the net was not trained.

in detail here is the optimal form of the trace rule, and of the parameter η which controls the length of the trace. In the real world, objects may typically be viewed for 0.5–1 sec or more, with fixation durations between saccades often 200–300 msec. A trace of up to 1 sec in the real world might thus be satisfactory. In the simulations described here, optimal values of η 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 approach to the optimal form of the trace rule is being pursued (Wallis, 1996a). This has shown that under a variety of probabilistic 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 object is inspected.

Since the trace rule is seen to be capable of playing a crucial role in the successful learning of invariant responses to objects, it is worth considering the major requirements of the learning that would be needed in 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. This requirement appears to be satisfied. Several hundred ms is sufficient time for the visual stimulus to process several different images of an object (Rolls and Tovee, 1994; Thorpe and Imbert, 1989; see section 1.5). Rapid primarily feedforward processing is the computational style of competitive neuronal networks, and the lateral interactions necessary within each stage for the competition could be implemented rapidly by integrate-and-fire neurons (see Section 1.5 and Simmen *et al.*, 1996). Secondly, at least some learning should take place rapidly in the cortex, based e.g. on several rapidly changing views of an object seen one or twice. This requirement also appears to be satisfied, in that learning of new representations appears to occur with as little as

* The consequences of using a trace rule for invariant object recognition in humans is considered further by Wallis, 1996c), who describes psychophysical results consistent with the trace rule theory.

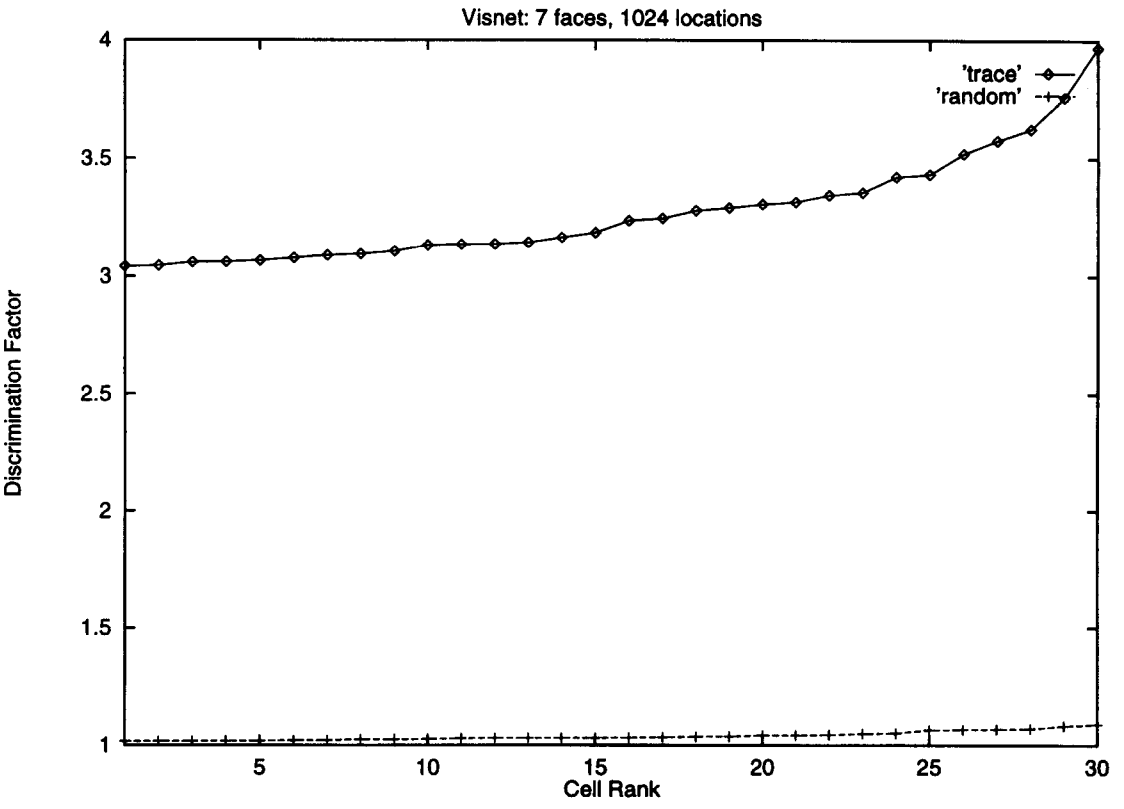


Fig. 27. The results of training VisNet on seven faces at 1024 locations, and then testing at all 1024 locations. The discrimination factor for the 30 most translation invariant cells in layer 4 when VisNet was trained with the trace rule, or was untrained (random weights) as a control, are shown.

several 1 sec presentations of new stimuli, such as faces, that have never been seen before (see Section 1.4).

The third requirement for the learning paradigm described here is that there should be some means for previous cellular activity to affect learning in the visual cortex. There are a number of possible mechanisms for this. One is that a memory trace for what has been seen in the last 300 msec appears to be implemented by a mechanism as simple as continued firing of inferior temporal neurons after the stimulus has disappeared (see Rolls and Tovee, 1994; Rolls *et al.*, 1994), probably as a result of attractor states being set up by Hebb-modifiable synapses being present between nearby cortical pyramidal cells (see Rolls and Treves, 1997). This mechanism would facilitate pairwise association between successive images of the same object. A second is that the binding period of glutamate to the NMDA receptors (which may last for 100 or more msec) and the entry of calcium to the postsynaptic neuron might implement a trace rule by producing a narrow time window over which the average activity at each presynaptic terminal affects learning (Rolls, 1992b; Rhodes, 1992). A third is that 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 (Foldiak, 1991; Montague *et al.*, 1991). It is of interest 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 (Ungerleider and Mishkin, 1982; Ungerleider and 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.

Acknowledgements—The authors have worked on some of the investigations described here with P. Azzopardi, G. C. Baylis, M. Booth, M. Elliffe, P. Foldiak, M. Hasselmo, C. M. Leonard, G. Littlewort, T. J. Milward, D. I. Perrett, M. J. Tovee and A. Treves, and their collaboration is sincerely acknowledged. The authors are grateful to Dr Peter Foldiak for help and advice in preparing this manuscript, and to Dr Roland Baddeley of the MRC Interdisciplinary Research Centre in Brain and Behaviour at Oxford, and Dr L. Abbott, a McDonnell–Pew Visiting Fellow at Oxford, for many helpful comments. Guy Wallis was supported by an SERC grant. Different parts of the research described were supported by the Medical Research Council, PG8513790; by a Human Frontier Science Programme grant; by an EC Human Capital and Mobility grant; by the MRC Oxford Interdisciplinary Research Centre in Brain and Behaviour; and by the Oxford McDonnell–Pew Centre in Cognitive Neuroscience.

REFERENCES

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

Baddeley, R. J., Wakeman, E., Booth, M., Rolls, E. T. and Abbott, L. F. (1997) The distribution of firing rates of primate temporal lobe visual neurons to "natural" scenes (in preparation).

Baizer, J. S., Ungerleider, L. G. and Desimone, R. (1991) Organization of visual inputs to the inferior temporal and posterior parietal cortex in macaques. *J. Neurosci.* **11,** 168–190.

Ballard, D. H. (1990) Animate vision uses object-centred reference frames. In: *Advanced Neural Computers*, pp. 229–236. Ed. R. Eckmiller. North-Holland: Amsterdam.

Ballard, D. H. (1993) Subsymbolic modelling of hand–eye co-ordination. In: *The Simulation of Human Intelligence*, Ch. 3, pp. 71–102. Ed. D. E. Broadbent. Blackwell: Oxford.

Barlow, H. B. (1972) Single units and sensation: a neuron doctrine for perceptual psychology?. *Perception* **1,** 371–394.

Barlow, H. B. (1985) Cerebral cortex as model builder. In: *Models of the Visual Cortex*, pp. 37–46. Eds D. Rose and V. G. Dobson. Wiley: Chichester.

Barlow, H. B., Kaushal, T. P. and Mitchison, G. J. (1989) Finding minimum entropy codes. *Neural Computat.* **1,** 412–423.

Baylis, G. C., Rolls, E. T. and 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 Res.* **342,** 91–102.

Baylis, G. C., Rolls, E. T. and Leonard, C. M. (1987) Functional subdivisions of temporal lobe neocortex. *J. Neurosci.* **7,** 330–342.

Baylis, G. C. and Rolls, E. T. (1987) Responses of neurons in the inferior temporal cortex in short term and serial recognition memory tasks. *Expl Brain Res.* **65,** 614–622.

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

Boussaoud, D., Desimone, R. and Ungerleider, L. G. (1991) Visual topography of area TEO in the macaque. *J. Comp. Neurol.* **306,** 554–575.

Breitmeyer, B. G. (1980) Unmasking visual masking: a look at the "why" behind the veil of the "how". *Psychol. Rev.* **87,** 52–69.

Brown, T. H., Kariss, E. W. and Keenan, C. L. (1990) Hebbian synapses: biological mechanisms and algorithms. *Ann. Rev. Neurosci.* **13,** 475–511.

Buhmann, J., Lades, M. and von der Malsburg, C. (1990) Size and distortion invariant object recognition by hierarchical graph matching. In: *International Joint Conference on Neural Networks*, pp. 411–416. IEEE: New York.

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

Bülthoff, H. and Edelman, S. (1992) Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc. natn. Acad. Sci. U.S.A.* **92,** 60–64.

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

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

Engel, A. K., Konig, P., Kreiter, A. K., Schillen, T. B. and Singer, W. (1992) Temporal coding in the visual system: new vistas on integration in the nervous system. *Trends Neurosci.* **15,** 218–226.

Feldman, J. A. (1985) Four frames suffice: a provisional model of vision and space [see p. 279]. *Behav. Brain Sci.* **8,** 265–289.

Foldiak, P. (1991) Learning invariance from transformation sequences. *Neural Comp.* **3,** 193–199.

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

Gross, C. G. (1973) Visual functions of the inferotemporal cortex. In: *Handbook of Sensory Physiology*, pp. 451–482. Springer-Verlag: Berlin.

Gross, C. G., Desimone, R., Albright, T. D. and Schwartz, E. L. (1985) Inferior temporal cortex and pattern recognition. *Expl Brain Res. Suppl.* **11,** 179–201.

Hasselmo, M. E., Rolls, E. T. and Baylis, G. C. (1989a) The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behav. Brain Res.* **32,** 203–218.

Hasselmo, M. E., Rolls, E. T., Baylis, G. C. and Nalwa, V. (1989b) Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey. *Expl Brain Res.* **75,** 417–429.

Hawken, M. J. and Parker, A. J. (1987) Spatial properties of the monkey striate cortex. *Proc. R. Soc. London [B]* **231,** 251–288.

Hertz, J., Krogh, A. and Palmer, R. G. (1991) *Introduction to the Theory of Neural Computation*. Addison-Wesley: Wokingham, U.K.

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 and McClelland (1986), Ch. 4.

Hornak, J., Rolls, E. T. and Wade, D. (1996) Face and voice expression identification and their association with emotional and behavioural changes in patients with frontal lobe damage. *Neuropsychologia* **34,** 247–261.

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

Humphreys, G. W. and Bruce, V. (1989) *Visual Cognition*. Erlbaum: Hove, U.K.

Koenderink, J. J. and van Doorn, A. J. (1979) The internal representation of solid shape with respect to vision. *Biol. Cybernet.* **32,** 211–216.

Kovacs, G., Vogels, R. and Orban, G. A. (1995) Cortical correlate of pattern backward masking. *Proc. Natn. Acad. Sci.* **92,** 5587–5591.

Leonard, C. M., Rolls, E. T., Wilson, F. A. W. and Baylis, G. C. (1985) Neurons in the amygdala of the monkey with responses selective for faces. *Behav. Brain Res.* **15,** 159–176.

Linsker, E. (1986) From basic network principles to neural architecture. *Proc. natn. Acad. Sci. U.S.A.*, **83,** 7508–7512, 8390–8394, 8779–8783.

Logothetis, N. K., Pauls, J., Bülthoff, H. H. and Poggio, T. (1994) View-dependent object recognition by monkeys. *Curr. Biol.* **4,** 401–414.

Marr, D. (1982) *Vision*. W. H. Freeman: San Francisco.

Maunsell, J. H. R. and Newsome, W. T. (1987) Visual processing in monkey extrastriate cortex. *Ann. Rev. Neurosci.* **10,** 363–401.

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

Miller, E. K. and Desimone, R. (1994) Parallel neuronal mechanisms for short-term memory. *Science* **263,** 520–522.

Miyashita, Y. (1993) Inferior temporal cortex: where visual perception meets memory. *Ann. Rev. Neurosci.* **16,** 245–263.

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

Montague, R., Gally, J. and Edelman, G. (1991) Spatial signalling in the development and function of neural connections. *Cerebr. Cort.* **1,** 199–220.

Nass, M. M. and Cooper, L. N. (1975) A theory for the development of feature detecting cells in visual cortex. *Biol. Cybernet.* **19,** 1–18.

Oja, E. (1982) A simplified neuron model as a principal component analyzer. *J. Math. Biol.* **15,** 267–273.

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

Perrett, D. I., Rolls, E. T. and Caan, W. (1982) Visual neurons responsive to faces in the monkey temporal cortex. *Expl Brain Res.* **47,** 329–342.

Perrett, D. I., Smith, P. A. J., Mistlin, A. J., Chitty, A. J., Head, A. S., Potter, D. D., Broennimann, R., Milner, A. D. and Jeeves, M. A. (1985a) Visual analysis of body movements by neurons in the temporal cortex of the macaque monkey: a preliminary report. *Behav. Brain Res.* **16,** 153–170.

Perrett, D. I., Smith, P. A. J., Potter, D. D., Mistlin, A. J., Head, A. S., Milner, D. and Jeeves, M. A. (1985b) Visual cells in temporal cortex sensitive to face view and gaze direction. *Proc. R. Soc.* **223B,** 293–317.

Perrett, D. I., Mistlin, A. J. and Chitty, A. J. (1987) Visual neurons responsive to faces. *Trends Neurosci.* **10,** 358–364.

Perrett, D. I., Hietanen, J. K., Oram, M. W. and Benson, P. J. (1992)

Organisation and functions of cells responsive to faces in the temporal cortex. *Phil. Trans. R. Soc. London [B]* **335,** 23–30.

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

Poggio, T. and Girosi, F. (1990a) Regularization algorithms for learning that are equivalent to multilayer networks. *Science* **247,** 978–982.

Poggio, T. and Girosi, F. (1990b) Networks for approximation and learning. *Proc. IEEE* **78,** 1481–1497.

Rhodes, P. (1992) The open time of the NMDA channel facilitates the self-organisation of invariant object responses in cortex. *Soc. Neurosci. Abstr.* **18,** 740.

Rolls, E. T. (1984) Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces. *Human Neurobiol.* **3,** 209–222.

Rolls, E. T. (1989a) Functions of neuronal networks in the hippocampus and neocortex in memory. In: *Neural Models of Plasticity: Experimental and Theoretical Approaches*, Ch. 13, pp. 240–265. Eds J. H. Byrne and W. O. Berry. Academic Press: San Diego.

Rolls, E. T. (1989b) The representation and storage of information in neuronal networks in the primate cerebral cortex and hippocampus. In: *The Computing Neuron*, Ch. 8, pp. 125–159. Eds R. Durbin, C. Miall and G. Mitchison. Addison-Wesley: Wokingham, U.K.

Rolls, E. T. (1989c) Functions of neuronal networks in the hippocampus and cerebral cortex in memory. In: *Models of Brain Function*, pp. 15–33. Ed. R. M. J. Coterill. Cambridge University Press: Cambridge, U.K.

Rolls, E. T. (1990) A theory of emotion, and its application to understanding the neural basis of emotion. *Cogn. Emotion* **4,** 161–190.

Rolls, E. T. (1991) Neural organisation of higher visual functions. *Curr. Opin. Neurobiol.* **1,** 274–278.

Rolls, E. T. (1992a) Neurophysiology and functions of the primate amygdala. In: *The Amygdala*, Ch. 5, pp. 143–165. Ed. J. P. Aggleton. Wiley-Liss: New York.

Rolls, E. T. (1992b) Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Phil. Trans. R. Soc.* **335,** 11–21.

Rolls, E. T. (1994) Brain mechanisms for invariant visual recognition and learning. *Behav. Proc.* **33,** 113–138.

Rolls, E. T. (1995a) A theory of emotion and consciousness, and its application to understanding the neural basis of emotion. In: *The Cognitive Neurosciences*, Ch. 72, pp. 1091–1106. Ed. M. S. Gazzaniga. MIT Press: Cambridge, MA.

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

Rolls, E. T. (1996a) Roles of long term potentiation and long term depression in neuronal network operations in the brain. In: *Cortical Plasticity: LTP and LTD*, Ch. 11, pp. 223–250. Eds M. S. Fazeli and G. L. Collingridge. Bios: Oxford, U.K.

Rolls, E. T. (1996b) A neurophysiological and computational approach to the functions of the temporal lobe cortical visual areas in invariant object recognition. In: *Computational and Biological Mechanisms of Visual Coding*, Eds L. Harris and M. Jenkin. Cambridge University Press: Cambridge, U.K.

Rolls, E. T., Baylis, G. C. and Leonard, C. M. (1985) Role of low and high spatial frequencies in the face-selective responses of neurons in the cortex in the superior temporal sulcus. *Vision Res.* **25,** 1021–1035.

Rolls, E. T. and Baylis, G. C. (1986) Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. *Expl Brain Res.* **65,** 38–48.

Rolls, E. T., Baylis, G. C. and Hasselmo, M. E. (1987) The responses of neurons in the cortex in the superior temporal sulcus of the monkey to band-pass spatial frequency filtered faces. *Vis. Res.* **27,** 311–326.

Rolls, E. T., Baylis, G. C., Hasselmo, M. E. and Nalwa, V. (1989) The effect of learning on the face-selective responses of neurons in the cortex in the superior temporal sulcus of the monkey. *Expl Brain Res.* **76,** 153–164.

Rolls, E. T. and Treves, A. (1990) The relative advantages of sparse versus distributed encoding for associative neuronal networks in the brain. *Network* **1,** 407–421.

Rolls, E. T., Tovee, M. J. and Ramachandran, V. S. (1993) Visual learning reflected in the responses of neurons in the temporal visual cortex of the macaque. *Soc. Neurosci. Abstr.* **19,** 27.

Rolls, E. T. and Tovee, M. J. (1994) Processing speed in the cerebral cortex, and the neurophysiology of visual backward masking. *Proc. R. Soc. B* **257,** 9–15.

Rolls, E. T., Tovee, M. J., Purcell, D. G., Stewart, A. L. and Azzopardi, P. (1994) The responses of neurons in the temporal cortex of primates, and face identification and detection. *Expl Brain Res.* **101,** 474–484.

Rolls, E. T. and Tovee, M. J. (1995a) Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *J. Neurophysiol.* **73,** 713–726.

Rolls, E. T. and Tovee, M. J. (1995b) The responses of single neurons in the temporal visual cortical areas of the macaque when more than one stimulus is present in the visual field. *Expl Brain Res.* **103,** 409–420.

Rolls, E. T., Treves, A. and Tovee, M. J. (1996a) The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. *Expl Brain Res.* (in press).

Rolls, E. T., Booth, M. C. A. and Treves, A. (1996b) View-invariant representations of objects in the inferior temporal visual cortex. *Soc. Neurosci. Abstr.* **22.**

Rolls, E. T., Tovee, M. and Treves, A. (1997) Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. *J. Comput. Neurosci.* (in press).

Rolls, E. T. and Treves, A. (1997) *Neural Networks and Brain Function*. Oxford University Press: Oxford.

Seltzer, B. and Pandya, D. N. (1978) Afferent cortical connections and architectonics of the superior temporal sulcus and surrounding cortex in the rhesus monkey. *Brain Res.* **149,** 1–24.

Simmen, M. W., Rolls, E. T. and Treves, A. (1996) On the dynamics of a network of spiking neurons. In *Computations and Neuronal Systems: Proceedings of CNS95*, Eds F. H. Eekman and J. M. Bower. Kluwer: Boston.

Snedecor, G. W. and Cochran, W. G. (1989) *Statistical Methods*, 8th edn. Iowa State University Press: Ames, IA.

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

Tanaka, K., Saito, C., Fukada, Y. and Moriya, M. (1990) Integration of form, texture, and color information in the inferotemporal cortex of the macaque. In: *Vision, Memory and the Temporal Lobe*, Ch. 10, pp. 101–109. Eds E. Iwai and M. Mishkin. Elsevier: New York.

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

Tarr, M. J. and Pinker, S. (1989) Mental rotation and orientation-dependence in shape recognition. *Cognit. Psychol.* **21,** 233–282.

Thorpe, S. J. and Imbert, M. (1989) Biological constraints on connectionist models. In: *Connectionism in Perspective*, pp. 63–92. Eds R. Pfeifer, Z. Schreter and F. Fogelman-Soulie. Elsevier: Amsterdam.

Tovee, M. J., Rolls, E. T., Treves, A. and Bellis, R. P. (1993) Information encoding and the responses of single neurons in the primate temporal visual cortex. *J. Neurophysiol.* **70,** 640–654.

Tovee, M. J., Rolls, E. T. and Azzopardi, P. (1994) Translation invariance and the responses of neurons in the temporal visual cortical areas of primates. *J. Neurophysiol.* **72,** 1049–1060.

Tovee, M. J. and Rolls, E. T. (1995) Information encoding in short firing rate epochs by single neurons in the primate temporal visual cortex. *Vis. Cognit.* **2,** 35–58.

Tovee, M. J., Rolls, E. T. and Ramachandran, V. S. (1996) Visual learning in neurons of the primate temporal visual cortex. *NeuroReport* **7,** 2757–2760.

Treves, A. (1993) Mean-field analysis of neuronal spike dynamics. *Network* **4,** 259–284.

Treves, A. and Rolls, E. T. (1994) A computational analysis of the role of the hippocampus in memory. *Hippocampus* **4,** 374–391.

Turvey, M. T. (1973) On the peripheral and central processes in vision: inferences from an information processing analysis of masking with patterned stimuli. *Psychol. Rev.* **80,** 1–52.

Ungerleider, L. G. and Mishkin, M. (1982) Two cortical visual systems. In: *Analysis of Visual Behaviour*, pp. 549–586. Eds D. J. Ingle, M. A. Goodale and R. J. W. Mansfield. MIT Press: Cambridge, MA.

Ungerleider, L. G. and Haxby, J. V. (1994) "What" and "Where" in the human brain. *Curr. Opin. Neurobiol.* **4,** 157–165.
Van Essen, D., Anderson, C. H. and 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 [Reprinted in Anderson and Rosenfeld, 1988]. *Kybernetik* **14,** 85–100.
von der Malsburg, C. (1981) *The Correlation Theory of Brain Function*. Technical report 81-2, Department of Neurobiology, Max-Planck-Institute for Biophysical Chemistry, Göttingen.
von der Malsburg, C. and Schneider, W. (1986) A neural cocktail-party processor. *Biol. Cybernet.* **54,** 29–40.
von der Malsburg, C. (1990) A neural architecture for the representation of scenes. In: *Brain Organization and Memory: Cells, Systems and Circuits*, Ch. 18, pp. 356–372. Eds J. L. McGaugh, N. M. Weinberger and G. Lynch. Oxford University Press: New York.
Wallis, G. (1994) *Neural Mechanisms Underlying Processing in the Visual Areas of the Occipital and Temporal Lobes*. Doctoral thesis, Department of Experimental Psychology, Oxford University, U.K.
Wallis, G. (1996a) Optimal unsupervised learning in invariant object recognition *Neural Computation* (in press).
Wallis, G. (1996b) Using spatio-temporal correlations to learn invariant object recognition. *Neural Networks* (in press).
Wallis, G. (1996c) Temporal order in human object recognition. *J. Biol. Syst.* (in press).
Wallis, G., Rolls, E. T. and Foldiak, P. (1993) Learning invariant responses to the natural transformations of objects. *Intl Joint Conf. Neural Networks* **2,** 1087–1090.
Yamane, S., Kaji, S. and Kawano, K. (1988) What facial features activate face neurons in the inferotemporal cortex of the monkey?. *Expl Brain Res.* **73,** 209–214.

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 and Cochran (1989) describe 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 normalizes 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:

$$\text{Discrimination Factor} = F_C \frac{(N_S - 1)S_S^2 + N_S(N_L - 1)S_e^2}{(N_L - 1)S_L^2 + N_L(N_S - 1)S_e^2}$$

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
$S_e^2 = S_{error}^2$ Sample error	F_C Correction factor

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