Behavioural Processes, 33 (1994) 113–138

BEPROC 00575

Brain mechanisms for invariant visual recognition and learning

Edmund T. Rolls *

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

(Accepted 10 June 1994)

Abstract

Mechanisms by which the brain could perform invariant recognition of objects including faces are addressed neurophysiologically, and then a computational model of how this could occur is described. Some neurons that respond primarily to faces are found in the macaque cortex in the anterior part of the superior temporal sulcus (in which region neurons are especially likely to be tuned to facial expression, and to face movement involved in gesture). They are also found more ventrally in the TE areas which form the inferior temporal gyrus. Here the neurons are more likely to have responses related to the identity of faces. These areas project on to the amygdala and orbitofrontal cortex, in which face-selective neurons are also found. Quantitative studies of the responses of the neurons that respond differently to the faces of different individuals show that information about the identity of the individual is represented by the responses of a population of neurons, that is, ensemble encoding is used. The rather distributed encoding (within the class faces) about identity in these sensory cortical regions has the advantages of maximising the information in the representation useful for discrimination between stimuli, generalisation, and graceful degradation. In contrast, the more sparse representations in structures such as the hippocampus may be useful to maximise the number of different memories stored. There is evidence that the responses of some of these neurons are altered by experience so that new stimuli become incorporated in the network, in only a few seconds of experience with a new stimulus. It is shown that the representation that is built in temporal cortical areas shows considerable invariance for size, contrast, spatial frequency and translation. Thus the representation is in a form which is particularly useful for storage and as an output from the visual system. It is also shown that one of the representations which is built is view-invariant, which is suitable for recognition and as an input to associative memory. Another is

* Corresponding author. Fax: +44 (865) 310447; E-mail: erolls@psy.ox.ac.uk.

SSDI 0376-6357(94)00040-9

viewer-centred, which is appropriate for conveying information about gesture. It is shown that these computational processes operate rapidly, in that in a backward masking paradigm, 20–40 ms of neuronal activity in a cortical area is sufficient to support face recognition. In a clinical application of these findings, it is shown that humans with ventral frontal lobe damage have in some cases impairments in face and voice expression identification. These impairments are correlated with and may contribute to the problems some of these patients have in emotional and social behaviour. To help provide an understanding of how the invariant recognition described could be performed by the brain, a neuronal network model of processing in the ventral visual system is described. The model uses a multistage feed-forward architecture, and is able to learn invariant representations of objects including faces by use of a Hebbian synaptic modification rule which incorporates a short memory trace (0.5 s) of preceding activity to enable the network to learn the properties of objects which are spatio-temporally invariant over this time scale.

Key words: Inferior temporal visual cortex; Amygdala; Orbitofrontal cortex; Face recognition; Face expression identification; Competitive neuronal network

Introduction

This paper draws together evidence on how information about visual stimuli is represented in the temporal cortical visual areas and the brain areas to which these are connected; on how these representations are formed; and on how learning about these representations occurs. 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, about the utility of the different representations found, and about how learning occurs in the brain regions which receive information from the temporal cortical visual areas. The recordings described are made mainly in non-human primates, firstly because the temporal lobe, in which this processing occurs, is much more developed than in non-primates, and secondly because the findings are relevant to understanding the effects of brain damage in patients, as will be shown. In this paper, particular attention will be paid to 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; because of the importance of face recognition and expression identification in the primate social behaviour; and because of the application of understanding this neural system to understanding the effects of damage to this system in humans on behaviour.

Neuronal responses found in different temporal lobe cortex visual areas

Visual pathways project by 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) in which some neurons which respond selectively to faces are found (Desimone and Gross, 1979; Bruce et al., 1981; Desimone et al., 1984; Gross et al., 1985; Rolls, 1981a, b, 1984, 1991, 1992b, c; Perrett, Rolls and

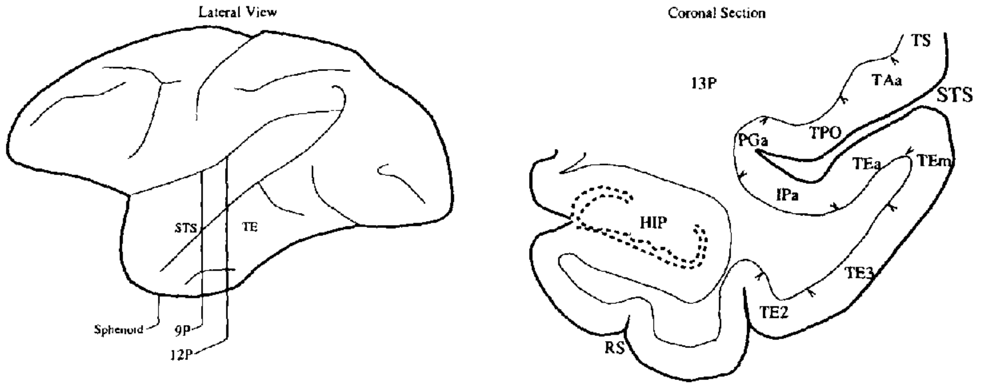

Fig. 1. Lateral view of the macaquue brain (left) and coronal section (right) showing the different architectonic areas (e.g. TEm, TPO) in and bordering the anterior part of the superior temporal sulcus (STS) of the macaque (see text). The coronal section is through the temporal lobe 133 mm P (posterior)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

Caan, 1982; Desimone, 1991). The inferior temporal visual cortex, area TE, is divided on the basis of cytoarchitecture, myeloarchitecture, and afferent input into areas TEa, TEm, TE3, TE2 and TE1. In addition there is a set of different areas in the cortex in the superior temporal sulcus (Seltzer and Pandya, 1978; Baylis, Rolls and Leonard, 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).

In order to investigate the information processing being performed by these parts of the temporal lobe cortex, the activity of single neurons was analysed in each of these areas in a sample of more than 2600 neurons in the rhesus macaque monkey during the presentation of simple and complex visual stimuli such as sine wave gratings, three-dimensional objects, and faces; and auditory and somatosensory stimuli (Baylis, Rolls and Leonard, 1987). Considerable specialization of function was found. 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 to the other temporal cortical areas in which they comprise 4–10%. The stimuli which activate other cells in these TE regions include simple visual patterns such as gratings, and combinations of simple stimulus features (Gross et al, 1985; Tanaka et al., 1990). Although face-selective neurons are thus found in the highest proportion in areas TPO within the superior temporal sulcus and TEa and TEm on the ventral lip of the sulcus, their extent is great in the anteroposterior direction (they are found in corresponding regions within the anterior half of the sulcus), and they are present in smaller proportions in many other temporal cortical areas (e.g. TE3, TE2 and TE1) (Baylis, Rolls and Leonard, 1987). Due to the fact that face-selective neurons have a wide distribution, it might be expected that only large lesions, or lesions that interrupt outputs of these visual areas, would produce readily apparent face-processing deficits. Further, as described below, neurons with responses related to facial expression, movement, and gesture are more likely to be found in the cortex in the superior temporal sulcus, whereas neurons with activity related to facial identity are more likely to be found in the TE areas (see also Hasselmo, Rolls and Baylis, 1989). These neurophysiological findings suggest that the appropriate tests for the effects of STS lesions will include tests of facial expression, movement, and gesture, whereas facial identity is more likely to be affected by TE lesions.

The selectivity of one population of neurons for faces

The neurons described in our studies as having responses selective for faces are selective in that they respond 2–20 times more (and statistically significantly more) to faces than to a wide range of gratings, simple geometrical stimuli, or complex 3-D objects (see Rolls, 1984; Baylis et al., 1985, 1987; Rolls, 1992b). (In fact, the majority of the neurons in the cortex in the superior temporal sulcus classified as showing responses selective for faces responded much more specifically than this. For half of these neurons, their response to the most effective face was more than five times as large as to the most effective non-face stimulus, and for 25% of these neurons, the ratio was greater than 10:1. The degree of selectivity shown by different neurons studied is illustrated in Fig. 6 of Rolls, 1992c and by Baylis, Rolls and Leonard, 1985, and the criteria for classification as face-selective are elaborated further by Rolls, 1992c.) The responses to faces are excitatory, sustained and are time-locked to the stimulus presentation with a latency of between 80 and 160 ms. The cells are typically unresponsive to auditory or tactile stimuli and to the sight of arousing or aversive stimuli. The magnitude of the responses of the cells is relatively constant despite transformations such as rotation so that the face is inverted or horizontal, and alterations of color, size, distance and contrast (see below). These findings indicate that explanations in terms of arousal, emotional or motor reactions, and simple visual feature sensitivity or receptive fields, are insufficient to account for the selective responses to faces and face features observed in this population of neurons (Perrett et al., 1982; Baylis et al., 1985; Rolls and Baylis, 1986). Observations consistent with these findings have been published by Desimone et al. (1984), who described a similar population of neurons located primarily in the cortex in the superior temporal sulcus which responded to faces but not to simpler stimuli such as edges and bars or to complex non-face stimuli (see also Gross et al., 1985).

In a recent study, further evidence has been obtained that these neurons are tuned to provide information about which face has been seen, but not about which non-face has been seen (Rolls and Tovee, 1994c). In this study a wide range of different faces (23) and non-face images (45) of real-world scenes was used. This enabled the function of this brain region to be analysed when it was processing natural scenes. The information available about which stimulus had been shown was measured quantitatively using information theory. This analysis showed that the responses of these neurons contained 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). Multidimensional scaling to produce a stimulus space represented by this population of neurons showed that the different faces were well separated in the space created, whereas the different non-face stimuli were grouped together. The information analyses and multidimensional scaling thus provided evidence that what was made explicit in the responses of these neurons was information about which face had been seen. Information about which non-face stimulus had been seen was not made explicit in these neuronal responses. These procedures provide an objective and quantitative way to show what is 'represented' by a particular population of neurons.

The selectivity of these neurons for whole faces or for parts of faces

Masking out or presenting parts of the face (e.g. eyes, mouth, or hair) in isolation reveal that different cells respond to different features or subsets of features. For some cells, responses to the normal organization of cut-out or line-drawn facial features are significantly larger than to images in which the same facial features are jumbled (Perrett et al., 1982). These findings are consistent with the hypotheses developed below that by competitive self-organisation some neurons in these regions respond to parts of faces by responding to combinations of simpler visual properties received from earlier stages of visual processing, and that other neurons respond to combinations of parts of faces and thus respond only to whole faces. Moreover, the finding that for some of these latter neurons the parts must be in the correct spatial configuration show that the combinations formed can reflect not just the features present, but also their spatial arrangement.

Ensemble encoding of facial identity

An important question for understanding brain function is whether a particular object (or face) is represented in the brain by the firing of one or a few gnostic (or 'grandmother') cells (Barlow, 1972), or whether instead the firing of a group or ensemble of cells each with somewhat different responsiveness provides the representation. We have investigated whether the face-selective neurons encode information which could be used to distinguish between faces and, if so, whether gnostic or ensemble encoding is used. We have found that in many cases (77% of one sample), these neurons are sensitive to differences between faces (as shown by analyses of variance) (Baylis et al., 1985). However, each neuron does not respond only to one face. Instead, each neuron has a different relative response to each of the members of a set of faces. This evidence from the neuronal responses about which individual was being seen was very significant, as shown by the finding that the number of standard deviations which separated the response to the most effective from that to the least effective face in the set (a measure analogous to detectability, d', in signal detection theory) was for many neurons greater than 1.0 (Baylis et al., 1985). A recent advance has been to quantify this evidence available in the neuronal responses about which face has been seen by using an information theoretic approach. This shows that these neurons convey on average approximately 0.4 bits of information about which face in a set of 20 faces has been seen (Tovee and Rolls, 1994; cf. Tovee et al, 1993). In an extension of these analyses, we are finding that from 10 neurons, the information available about which face has been seen is approximately 2 bits. (If this could be extrapolated, which is not yet known, then 40 such neurons could convey up to 8 bits of information, or sufficient to identify 256 faces. These of course represent maximum numbers, and the actual numbers of faces that could be identified would be less than this, depending on the extent to which the responses of these neurons are not independent.)

Having shown that these neurons do represent information that could support identification, we can now consider how finely or broadly tuned these neurons are to a set of faces, that is whether the representation is grandmother-cell like, sparse, or very distributed, and then the factors that determine the utility of the representation found. One way in which the fineness of tuning of these neurons has been quantified is with the breadth of tuning metric developed by Smith and Travers (1979). This is a coefficient of entropy (H) for each cell which ranges from 0.0, representing total specificity to one stimulus, to 1.0 which indicates an equal response to the different stimuli [1]. The breadth of tuning of the majority of the neurons analyzed was in the range 0.8–1.0. It was thus clear from this and other quantitative measures of the tuning of these face-responsive neurons that they did not respond only to the face of one individual, but that instead typically each neuron responded to a number of faces in the stimulus set (which included 5 different faces) (Baylis, Rolls and Leonard, 1985).

Another 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*, where

$$a = \frac{\langle \eta \rangle^2}{\langle \eta^2 \rangle}$$

and $\langle \cdot \rangle$ denotes an average over the statistical distribution characterizing the firing rate η of a cell to the set of input stimuli; or

$$a = \left(\sum_{i=1,n} r_i/n \right)^2 / \sum_{i=1,n} (r_{ii}^2/n)$$

where r_i is the firing rate to the i'th stimulus in the set of n stimuli.

The sparseness has a maximum value of 1.0, and a minimum value close to zero (1/n, if a neuron responded to only one of the n stimuli in a set of stimuli). (To interpret this measure, if the neurons had binary firing distributions – firing or not –, then if a neuron responded to 50% of the stimuli, the sparseness of its representation would be 0.5, and if it responded to 10% of the stimuli, the sparseness of its representation would be 0.1.) 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, 1994c). If the spontaneous firing rate was subtracted from the firing rate of the neuron to each stimulus, so that the responses of the neurons were used in the sparseness calculation, then the 'response sparseness' had a lower value, with a mean of 0.33 for the population of neurons to the set of 68 stimuli. If the response sparseness for these neurons was calculated over the set of 23 faces, the value was 0.60 (Rolls and Tovee, 1994c).

These findings show first that these cells do represent information that would be useful in face recognition and identification. Second, the results show that these cells do not use local or 'grandmother cell' encoding of face identity, but that instead information about

[1] $H = -k\Sigma_{i=1,n} p_i \log p_i$ where H = breadth of responsiveness, k = scaling constant (set so that H = 1.0 when the neuron responds equally well to all stimuli in the set of size n), p_i = the response to stimulus i expressed as a proportion of the total response to all the stimuli in the set.

which face is seen is represented over an ensemble of neurons, with each neuron responding to a subset of stimuli. The subsets are different for different neurons. Third, the results show that over a large set of stimuli (68), the representation, expressed as response sparseness, was quite distributed, with an average value of 0.33. (Of course, the exact value of this depends on the set of stimuli over which the sparseness is measured, and would be very low if few faces but many non-faces were present in the set of stimuli for these neurons. It is for this reason that a set of natural images was used in this study. Future studies with sets of images chosen to reflect fully the natural world of the monkeys will be useful.) On the other hand, it was interesting that the response sparseness within the set of stimuli about which these neurons conveyed information (faces) indicated a rather distributed representation (c.f. Young and Yamane, 1992), with a mean value of 0.60.

These findings address theoretical issues which are fundamental to understanding brain function. Important aspects of the ensemble rather than local encoding used by these neurons may arise from the fact that the outputs of these cortical areas reach structures such as the amygdala which are involved with making associations between visual representations of objects and with primary reinforcing stimuli such as the taste of food using pattern association neuronal networks; and structures such as the hippocampus which are involved in recognition and episodic memory using auto-association neuronal networks (Rolls, 1987, 1989a,b, 1992a; Rolls and Treves, 1990; Treves and Rolls, 1994). The advantages of distributed representations for the inputs to such association neuronal networks include pattern completion, generalization, and graceful degradation (fault tolerance) (or graceful performance with an imperfect specification of the connectivity during development) (see Rolls, 1987, 1989a,b). Another advantage of the rather distributed encoding about face identity found in this population of neurons is that in so far as the response sparseness is close to a value which implies that each neuron responds to half the face stimuli in a set, this may enable the maximum information about a set of stimuli to be provided by a population of neurons, provided of course that they do not have the same profile of responsiveness to the set of stimuli. Such a 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). However, a representation as distributed as this would not be appropriate for a memory system, in which the aim is to store a large number of memories. In an associative memory with neurons with continuously variable firing rates, such as the autoassociative memory believed to be implemented in the hippocampus (Rolls, 1989), we have shown (Treves, 1990; Treves and Rolls, 1991, 1994) that the maximum number p_{max} of firing patterns that can be (individually) retrieved is proportional to the number C^{RC} of (associatively) modifiable recurrent collateral synapses per cell, by a factor that increases roughly with the inverse of the sparseness a of the neuronal representation. Approximately,

$$p_{max} \simeq \frac{C^{RC}}{a \ln(1/a)} k$$

where k is a factor that depends weakly on the detailed structure of the rate distribution, on the connectivity pattern, etc., but is roughly in the order of 0.2–0.3. Thus in memory systems one parameter limiting the number of memories that can be stored and retrieved correctly is the number of modifiable synapses received by each neuron (in the order of 5000–15 000), and another is the sparseness of the representation, a. (Similar considerations apply to pattern associators, Rolls and Treves, 1990, such as are suggested to be implemented in the amygdala and orbitofrontal cortex, and in cortico-cortical Backprojections, Rolls, 1989a,b, 1992a,c; Treves and Rolls, 1994). Thus in memory systems optimal performance requires a sparse representation, and this is consistent with values for sparseness in the hippocampus which are in the order of 0.01–0.04 (see Rolls and Tovee, 1994c; Treves and Rolls, 1994). It is therefore proposed that these fundamentally different constraints account for the different sparsenesses of representations found in the high order sensory cortices such as the temporal cortical areas described here, and in memory systems such as the hippocampus. In the sensory cortex, a relatively distributed representation may be used in order to optimize discriminative ability. In memory systems, much more sparse representations may be used in order to maximize the number of memories that can be stored.

It may be noted that it is unlikely that there are further processing areas beyond those described where ensemble coding changes into grandmother cell encoding. Anatomically, there does not appear to be a whole further set of visual processing areas present in the brain; and outputs from the temporal lobe visual areas such as those described, are taken to limbic and related regions such as the amygdala and via the entorhinal cortex the hippocampus. Indeed, tracing this pathway onwards, we have found a population of neurons with face-selective responses in the amygdala, and in the majority of these neurons, different responses occur to different faces, with ensemble (not local) coding still being present (Leonard et al., 1985; Rolls, 1992a). The amygdala in turn projects to another structure which may be important in other behavioural responses to faces, the ventral striatum, and comparable neurons have also been found in the ventral striatum (Williams et al., 1993).

Invariance in the neuronal representation of stimuli

One of the major problems which must be solved by a visual system 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, angle of view, etc. To investigate whether these neurons in the temporal lobe visual cortex are at a stage of processing where such invariance is being represented in the responses of neurons, the effect of such transforms of the visual image on the responses of the neurons was investigated.

To investigate whether the responses of these neurons show some of the perceptual properties of recognition including tolerance to isomorphic transforms (i.e. in which the shape is constant), the effects of alteration of the size and contrast of an effective face stimulus on the responses of these neurons were analysed quantitatively in macaque monkeys (Rolls and Baylis, 1986). It was shown that the majority of these neurons had responses which were relatively invariant with respect to the size of the stimulus. The median size change tolerated with a response of greater than half the maximal response was 12 times. Also, the neurons typically responded to a face when the information in it had been reduced from 3D to a 2D representation in grey on a monitor, with a response which was on average 0.5 of that to a real face. (This reduction in amplitude does not by itself mean that the point in multidimensional space represented by the ensemble of neurons has moved. The point represented by a facial identity ensemble will move only to the extent that the responses of neurons in the facial identity ensemble are affected differently by this transform. The original data are shown in Rolls and Baylis, 1986.) Another transform over which recognition is relatively invariant is spatial frequency. For example, a face can be identified when it is blurred (when it contains only low spatial frequencies), and when it is high-pass spatial frequency filtered (when it looks like a line drawing). It has been shown that if the face images to which these neurons respond are low-pass filtered in the spatial frequency domain (so that they are blurred), then many of the neurons still respond when the images contain frequencies only up to 8 cycles per face. Similarly, the neurons still respond to high-pass filtered images (with only high spatial frequency edge information) when frequencies down to only 8 cycles per face are included (Rolls et al., 1985). Face recognition shows similar invariance with respect to spatial frequency (see Rolls et al., 1985). Further analysis of these neurons with narrow (octave) bandpass spatial frequency filtered face stimuli shows that the responses of these neurons to an unfiltered face can not be predicted from a linear combination of their responses to the narrow band stimuli (Rolls et al., 1987). This lack of linearity of these neurons, and their responsiveness to a wide range of spatial frequencies, indicate that in at least this part of the primate visual system recognition does not occur using Fourier analysis of the spatial frequency components of images.

To investigate whether neurons in the inferior temporal visual cortex and cortex in the anterior part of the superior temporal sulcus operate with translation invariance in the awake behaving primate, their responses were measured during a visual fixation (blink) task in which stimuli could be placed in different parts of the receptive field (Tovee, Rolls and Azzopardi, 1994). It was found that in most cases the responses of the neurons were little affected by which part of the face was fixated, and that the neurons responded (with a greater than half-maximal response) even when the monkey fixated 2–5 degrees beyond the edge of a face which subtended 8–17 degrees at the retina. Moreover, the stimulus selectivity between faces was maintained this far eccentric within the receptive field. These results held even across the visual midline. It was also shown that these neurons code for identity and not fixation position, in that there was approximately six times more information in the responses of these neurons about which face had been seen than about where the monkey fixated on the face. It is concluded that at least some of these neurons in the temporal lobe visual areas do have considerable translation invariance so that this is a computation which must be performed in the visual system. Ways in which the translation and size invariant representations shown to be present in the brain by these studies could be built are considered below. It is clearly important that translation 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.

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 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 of rhesus macaques performing a visual fixation task. We found that the response of neurons to an effective face centred 8.5 degrees 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/s. On the other hand, the average response to a fixated non-effective face with an effective face in the periphery was 22 spikes/s. (These firing rates reflected the fact that in this population of neurons, the mean response for an effective face was 49 spikes/s with the face at the fovea, and 35 spikes/s with the face 8.5 degrees 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, 1994b). 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 (c.f. Ballard, 1993). These findings are a first step towards understanding how the visual system functions in a normal environment.

A view-independent representation of visual information

For recognizing and learning about objects (including faces), it is important that an output of the visual system should be not only translation and size invariant, but also relatively view invariant. In an investigation of whether there are such neurons, we found that some temporal cortical neurons reliably responded differently to the faces of two different individuals independently of viewing angle (Hasselmo, Rolls, Baylis and Nalwa, 1989), although in most cases (16/18 neurons) the response was not perfectly view-independent. Mixed together in the same cortical regions there are neurons with view-dependent responses (e.g. Hasselmo et al., 1989; Rolls and Tovee, 1994c). Such neurons might respond for example to a view of a profile of a monkey but not to a full-face view of the same monkey (Perrett et al., 1985b). These findings, of view-dependent, partially view independent, and view independent representations in the same cortical regions are consistent with the hypothesis discussed below that view-independent representations are being built in these regions by associating together neurons that respond to different views of the same individual. These findings also provide evidence that the outputs of the visual system are likely to include representations of what is being seen, in a view independent way that would be useful for object recognition and for learning associations about objects; and in a view-based way that would be useful in social interactions to determine whether another individual is looking at one, and for selecting details of motor responses, for which the orientation of the object with respect to the viewer is required.

Further evidence that some neurons in the temporal cortical visual areas have object-based rather than view-based responses comes from a study of a population of neurons that responds to moving faces (Hasselmo, Rolls, Baylis and Nalwa, 1989). For example, four neurons responded vigorously to a head undergoing ventral flexion, irrespective of whether the view of the head was full face, of either profile, or even of the back of the head. These different views could only be specified as equivalent in object-based coordinates. Further, for all of the 10 neurons that were tested in this way, the movement specificity was maintained across inversion, responding for example to ventral flexion of the head irrespective of whether the head was upright or inverted. In this procedure, retinally encoded or viewer-centered movement vectors are reversed, but the object-based description remains the same. It was of interest that the neurons tested generalized across different heads performing the same movements.

Also consistent with object-based encoding is the finding of a small number of neurons which respond to images of faces of a given absolute size, irrespective of the retinal image size (Rolls and Baylis, 1986).

Different neural systems are specialized for recognition and for face expression decoding

To investigate whether there are neurons in the cortex in the anterior part of the superior temporal sulcus of the macaque monkey which could provide information about facial expression (Rolls, 1981b, 1984, 1986a,b, 1990b), neurons were tested with facial stimuli which included examples of the same individual monkey with different facial expressions (Hasselmo et al, 1986b; Hasselmo, Rolls and Baylis, 1989). The responses of 45 neurons with responses selective for faces were measured to a set of 3 individual monkey faces with three expressions for each monkey, as well as to human expressions. Of these neurons, 15 showed response differences to different identities independently of expression, and 9 neurons showed responses which depended on expression but were independent of identity, as measured by a two-way ANOVA. Multidimensional scaling confirmed this result, by showing that for the first set of neurons the faces of different individuals but not expressions were well separated in the space, whereas for the second group of neurons, different expressions but not the faces of different individuals were well separated in the space. The neurons responsive to expression were found primarily in the cortex in the superior temporal sulcus, while the neurons responsive to identity were found in the inferior temporal gyrus. These results show that there are some neurons in this region the responses of which could be useful in providing information about facial expression, of potential use in social interactions (Rolls, 1981b, 1984, 1986a,b, 1990b). Damage to this population may contribute to the deficits in social and emotional behavior which are part of the Kluver-Bucy syndrome produced by temporal lobe damage in monkeys (see Rolls, 1981b, 1984, 1986a,b, 1990b, 1991a,c; Leonard et al., 1985).

A further way in which some of these neurons may be involved in social interactions is that some of them respond to gestures, e.g. to a face undergoing ventral flexion, as described above and by Perrett et. al. (1985a). The interpretation of these neurons as being useful for social interactions is that in some cases these neurons respond not only to ventral head flexion, but also to the eyes lowering and the eyelids closing (Hasselmo et al., 1989). Now these two movements (head lowering and eyelid lowering) often occur together when a monkey is breaking social contact with another, e.g. after a challenge, and the information being conveyed by such a neuron could thus reflect the presence of this social gesture. That the same neuron could respond to such different, but normally co-occurrent, visual inputs could be accounted for by the Hebbian competitive self-organization described below. It may also be noted that it is important when decoding facial expression not to move entirely into the object-based domain (in which the description would be in terms of the object itself, and would not contain information about the position and orientation of the object relative to the observer), but to retain some information about the head direction of the face stimulus being seen relative to the observer, for this is very important in determining whether a threat is being made in your direction. The presence of view-dependent representations in some of these cortical regions is consistent with this requirement. Indeed, it may be suggested that the cortex in the superior temporal sulcus, in which neurons are found with responses related to facial expression (Hasselmo, Rolls and Baylis, 1989), head and face movement involved in for example gesture (Hasselmo, Rolls, Baylis and Nalwa, 1989), and eye gaze (Perrett et al., 1985b), may be more related to face expression decoding; whereas the TE areas (more ventral, mainly in the macaque inferior temporal gyrus), in which neurons tuned to face identity (Hasselmo, Rolls and Baylis, 1989) and with view-independent responses (Hasselmo, Rolls, Baylis and Nalwa, 1989) are more likely to be found, may be more related to an object-based representation of identity. Of course, for appropriate social and emotional responses, both types of subsystem would be important, for it is necessary to know both the direction of a social gesture, and the identity of the individual, in order to make the correct social or emotional response.

Outputs from the temporal cortical visual areas reach the amygdala and the orbitofrontal cortex, and evidence is accumulating that these brain areas are involved in social and emotional responses to faces (Rolls, 1990a,b, 1992a–c, 1994). For example, lesions of the amygdala in monkeys disrupt social and emotional responses to faces, and we have identified a population of neurons with face-selective responses in the primate amygdala (Leonard et al., 1985), some of which may respond to facial and body gesture (Brothers et al., 1990). We (observations of E.T. Rolls and H.D. Critchley), and Wilson et al., 1993, have also found a small number of face-responsive neurons in the orbitofrontal cortex, and also in the ventral striatum, which receives projections from the amygdala and orbitofrontal cortex (Williams, Rolls, Leonard and Stern, 1993).

We have applied this research to the study of humans with frontal lobe damage, to try to develop a better understanding of the social and emotional changes which may occur in these patients. Impairments in the identification of facial and vocal emotional expression were demonstrated in a group of patients with ventral frontal lobe damage who had behavioural problems such as disinhibited or socially inappropriate behaviour (Hornak et al., 1994). A group of patients with lesions outside this brain region, without these behavioural problems, was unimpaired on the expression identification tests. The impairments shown by the frontal patients on these expression identification tests could occur independently of perceptual difficulties. Face expression identification was severely impaired in some patients whose recognition of the identity of faces was normal. Severe impairments on the vocal expression test (which consisted of non-verbal emotional sounds) were found in patients who produced excellent imitations of the sounds they could not identify, and whose identification of environmental sounds was also normal.

These findings suggest that some of the social and emotional problems associated with ventral frontal lobe or amygdala damage may be related to a difficulty in identifying correctly facial (and vocal) expression (Hornak et al., 1994). The question then arises of what functions are performed by the orbitofrontal cortex and amygdala with the face-related outputs they receive from the temporal cortical visual areas. The hypothesis has been developed that these regions are important in emotional and social behaviour because of their role in reward-related learning (Rolls, 1986a,b, 1990b, 1994). The amygdala is especially involved in learning associations between visual stimuli and primary (unlearned) rewards and punishments such as food taste and touch, and the orbitofrontal cortex is especially involved in the rapid reversal (i.e. adjustment or relearning) of such stimulus reinforcement associations. According to this hypothesis, the importance of projecting face-related information to the amygdala and orbitofrontal cortex is so that they can learn associations between faces, using information about both face identity and facial expression, and rewards and punishments. Now it is particularly in primate social behaviour that rapid relearning about individuals, identified by their face, and depending on their facial expression, must occur very rapidly and flexibly, to keep up with the continually changing social exchanges between different individuals and groups of individuals. It is crucial to be able to remember recent reinforcement associations of different individuals, and to be able to continually adjust these. It is suggested that these factors have led to the very major development of the orbitofrontal cortex in primates, to receive appropriate inputs (about identity from faces, and about facial expression), and to provide a very rapid and flexible learning mechanism for the current reinforcement associations of these inputs. Consistent with this, the same patients that are impaired in face expression identification are also impaired on stimulus-reinforcement relearning tasks such as visual discrimination reversal and extinction (Rolls et al., 1994). Moreover, this learning impairment is highly correlated with the social and behavioural changes found in these patients (Rolls et al., 1994).

Learning of new representations in the temporal cortical visual areas

Given the fundamental importance of a computation which results in relatively finely tuned neurons which across ensembles but not individually specify objects including individual faces in the environment, we have investigated 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, we 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 for example that the population would make small adjustments in the responsiveness of its individual neurons, so that neurons would acquire filter properties which would enable the population as a whole to discriminate between the faces actually seen. We thus 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. First, 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 s 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 (1–2) 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 was 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. 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 categorisation being performed by self-organizing competitive neuronal networks, as described below and elsewhere (Rolls, 1989a–c).

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 ten 0.5s presentations. It was found that in a number of cases, if the neuron happened to be responsive to that face, that when the binarized version of the same face was shown next, the neurons responded to it (Rolls, Tovee and Ramachandran, 1993). 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) of experience with a visual stimulus.

Such rapid learning of representations of new objects 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, for some of these neurons respond differently to recently seen stimuli in short term visual memory tasks (Baylis and Rolls, 1987; Miller and Desimone, 1994), and neurons in a more ventral cortical area respond during the delay in a short term memory task (Miyashita, 1993).

The speed of processing in the temporal cortical visual areas

Given that there is a whole sequence of visual cortical processing stages including V1, V2, V4, and the posterior inferior temporal cortex to reach the anterior temporal cortical areas, and that the response latencies of neurons in V1 are about 40–50 ms, and in the anterior inferior temporal cortical areas approximately 80–100 ms, each stage may need to perform processing for only 15–30 ms 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). This seems to imply very fast computation by each cortical area, and therefore to place constraints on the type of processing performed in each area that is necessary for final object identification. We note that rapid identification of visual stimuli is important in social and many other situations, and that there must be strong selective pressure for rapid identification. For these reasons, we have investigated the speed of processing quantitatively, as follows.

In a first approach, we measured the information available in short temporal epochs of the responses of temporal cortical face-selective neurons about which face had been seen. We found that if a period of the firing rate of 50 ms was taken, then this contained 84.4% of the information available in a much longer period of 400 ms about which of four faces had been seen. If the epoch was as little as 20 ms, the information was 65% of that available from the firing rate in the 400 ms period (Tovee et al., 1993). These high information yields were obtained with the short epochs taken near the start of the neuronal response, for example in the post-stimulus period 100--120 ms. Moreover, we were able to show 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 was stable over the whole response period of the neurons (Tovee et al., 1993). We were able to extend this finding to the case when a much larger stimulus set, of 20 faces, was used. Again, we found that the information available in short (e.g. 50 ms) epochs was a considerable proportion (e.g. 65%) of that available in a 400 ms long firing rate analysis period (Tovee and Rolls, 1994). 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 how long 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 ms) 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 could 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 of neuronal activity, we analysed the effects of backward masking with a pattern mask on the responses of single neurons to faces (Rolls and Tovee, 1994a). 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 ms presentation of the test stimulus for 200–300 ms, far longer than the presentation time. It is suggested that this reflects the operation of a short term memory system implemented in cortical circuitry, the importance of which 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 ms, face-selective neurons in the inferior temporal cortex of macaques responded for a period of 20–30 ms before their firing was interrupted by the mask (Rolls and Tovee, 1994a). We went on to show that under these conditions (a test-mask stimulus onset asynchrony of 20 ms), human observers looking at the same displays could just identify which of 6 faces was shown (Rolls, Tovee, Purcell, Stewart and Azzopardi, 1994).

These results provide evidence that a cortical area can perform the computation necessary for the recognition of a visual stimulus in 20–30 ms, and provide a fundamental constraint which must be accounted for in any theory of cortical computation. The results emphasise just how rapidly cortical circuitry can operate. This rapidity of operation has obvious adaptive value, and allows the rapid behavioral 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, shows that such networks can settle very rapidly (Treves, Rolls and Tovee, 1994).

Possible computational mechanisms in the visual cortex for object recognition

The neurophysiological findings described above (see also Rolls, 1990a, 1991), and wider considerations on the possible computational properties of the cerebral cortex (Rolls, 1989a,b, 1992b), lead to the following outline working hypotheses on object recognition by visual cortical mechanisms. 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 recognise the faces of many different individuals, that is 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 degree 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 degrees in V4, 20 degrees in TEO, and 50 degrees 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

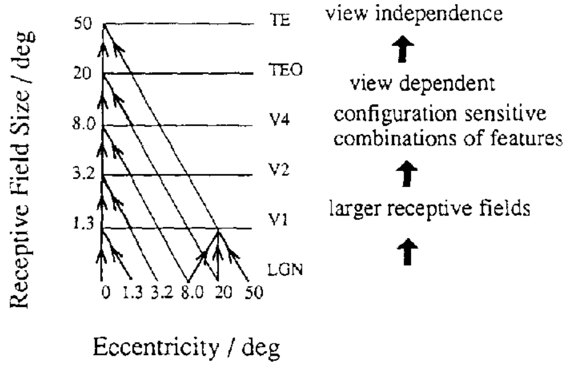

Fig. 2. Schematic diagram showing convergence achieved by the forward projections in the visual system, and the types of representation that may be built by competitive networks operating at each stage of the system from the primary visual cortex (V1) to the inferior temporal visual cortex (area TE) (see text). LGN, lateral geniculate nucleus. 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).

representations are computed. Each layer is considered to act partly as a set of local self-organising 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 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). (A biologically plausible form of this learning rule that operates well in such networks is

$$\delta s_{rc} = k \cdot m_c(a_r - s_{rc})$$

where k is a constant, δs_{rc} is the change of synaptic weight, a_r is the firing rate of the r'th axon, and m_c is a non-linear function of the output activation which mimics the operation of the NMDA receptors in learning; see Rolls, 1989a–c). 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. These networks thus act as categorisers. In relation to visual information processing, they would remove redundancy from the input representation, and would develop low entropy representations of the information (c.f. Barlow, 1985; Barlow et al., 1989). Such competitive nets are biologically plausible, in that they utilise Hebb-modifiable forward excitatory connections, with competitive inhibition mediated by cortical inhibitory neurons. The competitive scheme I suggest 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–c). The scheme has the advantages that the output neurons learn better to distribute themselves between the input patterns (c.f. Bennett, 1990), and that the sparse representations formed (which provide 'coarse coding') have utility in maximising 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,b; 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 I propose is learned differently, in that instead of using perhaps non-biologically plausible algorithms to optimally locate 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 below). 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 generalise smoothly from the examples with which it has learned (c.f. Poggio and Girosi, 1990a,b), to help the system to respond for example with the correct invariances as described below.

Translation invariance would be computed in such a system by utilising 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 analyzers 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 s), the membrane of the postsynaptic neuron would still be in its 'Hebb-modifiable' state (caused for example 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 s) 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 ms 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, 1994a; Rolls, Tovee et al., 1994). I also suggest that other invariances, for example 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 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, 1990, and with our data on face neurons, in that some faces 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) 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 (c.f. 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., 1989; Perrett et al., 1987). This solution to 'object-based' representations is very different from that traditionally proposed for artificial vision systems, in which the coordinates in 3D-space of objects are stored in a database, and general-purpose algorithms operate on these to perform transforms such as translation, rotation, and scale change in 3D space (e.g. Marr, 1982). In the present, much more limited but more biologically plausible scheme, the representation would be suitable for recognition of an object, and for linking associative memories to objects, but would be less good for making actions in 3D-space to particular parts of, or inside, objects, as the 3D coordinates of each part of the object would not be explicitly available. It is therefore proposed 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 (c.f. Ballard, 1990).

The computational processes proposed above operate by an unsupervised learning mechanism, which utilises regularities in the physical environment to enable representations with low entropy to be built. In some cases it may be advantageous to utilise 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,b) 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,b).

The computational processes outlined above use coarse 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 coarse coding and fine tuning 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 generalise correctly to solve for example 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.

A computational model of invariant visual object recognition

To test and clarify the hypotheses just described about how the visual system may operate to learn invariant object recognition, we have performed a simulation which implements many of the ideas just described, and is consistent and based on much of the neurophysiology summarized above. The network simulated can perform object, including face, recognition in a biologically plausible way, and after training shows for example translation and view invariance (Wallis, Rolls and Foldiak, 1993).

In the four layer network, the successive layers correspond approximately to V2, V4, the posterior temporal cortex, and the anterior temporal cortex. The forward connections to a cell in one layer are derived from a topologically corresponding region of the preceding layer, using a Gaussian distribution of connection probabilities to determine the exact neurons in the preceding layer to which connections are made. This schema is constrained to preclude the repeated connection of any cells. Each cell receives 50 connections from the 32 × 32 cells of the preceding layer, with a 67% probability that a connection comes from within 4 cells of the distribution centre. Figure 3 shows the general convergent network architecture used, and may be compared with Fig. 2. Within each layer, lateral inhibition between neurons has a radius of effect just greater than the radius of feedforward convergence just defined. The lateral inhibition is simulated via a linear local contrast enhancing filter active on each neuron. (Note that this differs from the global 'winner-take-all' paradigm implemented by Foldiak 1991). The cell activation is then passed through a

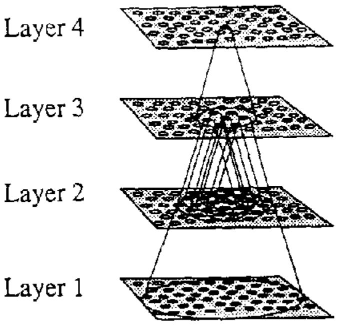

Fig. 3. Hierarchical network structure.

non-linear cell output activation function, which also produces contrast enhancement of the firing rates.

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. These response characteristics of neurons in the input layer are 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 (see Wallis et al., 1993). Four filter spatial frequencies (in the range 0.0625 to 0.25 pixels^{-1} over four octaves), each with one of four orientations (0° to 135°) were implemented. Cells of layer 1 receive a topologically consistent, localised, random selection of the filter responses in the input layer, under the constraint that each cell samples every filter spatial frequency and receives a constant number of inputs.

The synaptic learning rule used can be summarised as follows:

$$\delta s_{rc} = k \cdot m_c \cdot a_r$$

and

$$m_c^t = (1 - \eta) f_c^{(t)} + \eta m_c^{(t-1)}$$

where a_r is the rth input to the neuron, f_c is the output of the cth neuron, s_{rc} is the rth weight on the cth neuron, η governs the relative influence of the trace and the new input (typically 0.4–0.6), and $m_c^{(t)}$ represents the value of the cth cell's memory trace at time t. In the simulation the neuronal learning was bounded by normalisation of each cell's dendritic weight vector. An alternative, more biologically relevant implementation, using a local weight bounding operation, has in part been explored using a version of the Oja update rule (Oja 1982; Kohonen 1984). To train the network to produce a translation invariant representation, one stimulus was placed successively in a sequence of 7 positions across the input, then the next stimulus was placed successively in the same sequence of 7 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.

One test of the network used a set of three non-orthogonal stimuli, based upon probable 3-D edge cues (such as 'T, L and + ' shapes). During training these stimuli were chosen in random sequence to be swept across the 'retina' of the network, a total of 1000 times. In order to assess the characteristics of the cells within the net, a two-way analysis of variance was performed on the set of responses of each cell, with one factor being the stimulus type and the other the position of the stimulus on the 'retina'. A high F ratio for stimulus type (F_s), and low F ratio for stimulus position (F_p) would imply that a cell had learned a position invariant representation of the stimuli. The discrimination factor of a particular cell was then simply the ratio F_s/F_p (a factor useful for ranking at least the most invariant cells). To assess the utility of the trace learning rule, nets trained with the trace rule were compared with nets trained with standard Hebbian learning without a trace, and

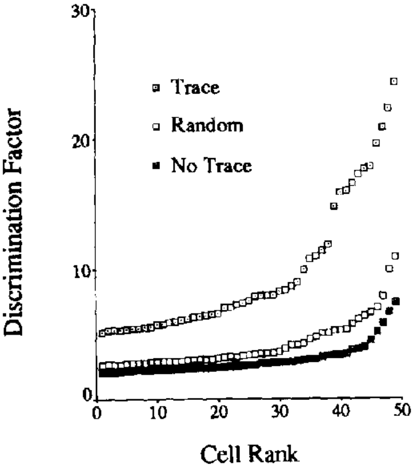

Fig. 4. Comparison of network discrimination when trained with the trace learning rule, with a Hebb rule (No trace), and when not trained (Random).

with untrained nets (with the initial random weights). The result of the simulations, illustrated in Fig. 4, show that networks trained with the trace learning rule do have neurons with much higher values of the discrimination factor. An example of the responses of one such cell are illustrated in Fig. 5. Similar position invariant encoding has been demonstrated for a stimulus set consisting of 8 faces. View invariant coding has also been demonstrated for a set of 5 faces each shown in 4 views.

These results show that the proposed learning mechanism and neural architecture can produce cells with responses selective for stimulus type with considerable position or view invariance. The ability of the network to be trained with natural scenes may also help to advance our understanding of encoding in the visual system.

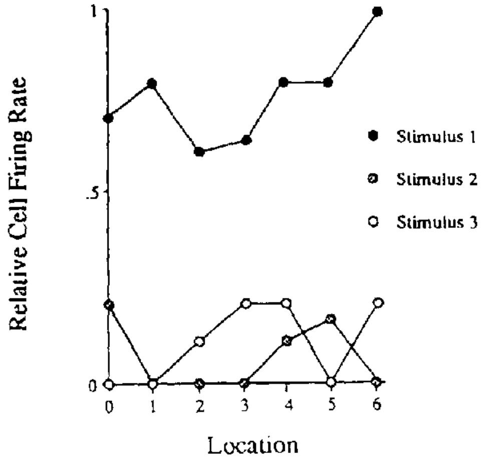

Fig. 5. The responses of a layer 4 cell in the simulation. The cell had a translation invariant response to stimulus 1.

Acknowledgements

The author has worked on some of the investigations described here with P. Azzopardi, G.C. Baylis, M. Hasselmo, C.M. Leonard, G. Littlewort, D.I. Perrett, R. Payne, M. Tovee and A. Treves, and their collaboration is sincerely acknowledged. This research was supported by the Medical Research Council, PG8513790.

References

Azzopardi, P. and Rolls, E.T., 1989. Translation invariance in the responses of neurons in the inferior temporal visual cortex of the macaque. Soc. Neurosci. Abs., 15: 120.

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, R. Eckmiller (Editor), Amsterdam, North-Holland: pp. 229–236.

Ballard, D.H., 1993. Subsymbolic modelling of hand-eye co-ordination. In: The Simulation of Human Intelligence, D.E. Broadbent (Editor), Ch. 3, pp. 71–102, 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, D. Rose and V.G. Dobson (Editors), Chichester, Wiley: pp. 37–46.

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., Rolls,E.T., 1987 Responses of neurons in the inferior temporal cortex in short term and serial recognition memory tasks. Experimental Brain Research 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.

Brothers, L., Ring, B. and Kling, A.S., 1990. Response of neurons in the macaque amygdala to complex social stimuli. Behav. Brain Res. 41: 199–213.

Bruce, C., Desimone, R. and Gross, C.G., 1981. Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. J. Neurophys., 46: 369–384.

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

Desimone, R. and Gross, C.G., 1979. Visual areas in the temporal lobe of the macaque. Brain Res., 178: 363–380.

Desimone, R., Albright, T.D., Gross, C.G. and Bruce, C., 1984. Stimulus-selective properties of inferior temporal neurons in the macaque. J. Neurosci., 4: 2051–2062.

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

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

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

Hasselmo, M.E., Rolls, E.T. and Baylis, G.C., 1986. Selectivity between facial expressions in the responses of a population of neurons in the superior temporal sulcus of the monkey. Neurosci. Lett., S26, S571.

Hasselmo, M.E., Rolls, E.T. and Baylis, G.C., 1989. 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., 1989. Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey. Exp. Brain Res., 75: 417–429.

Hornak, J., Rolls, E.T. and Wade, D., 1994. Face and voice expression identification and their association with emotional and behavioural changes in patients with frontal lobe damage.

Humphreys, G.W. and Bruce, V., 1989. Visual Cognition. Hove, Erlbaum.

Koenderink, J.J. and Van doorn, A.J., 1979. Biol. Cybern., 32: 211–217.

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.

Logothetis, N.K., Pauls, J., Bulthoff, H.H. and Poggio, T., 1994, Current Biology, 4: 401–414.

Malsburg, C. Von der, 1990. A neural architecture for the representation of scenes. In: Brain Organization and Memory: Cells, Systems and Circuits, J.L. McGaugh, N.M. Weinberger and G. Lynch (Editors), New York, Oxford University Press: Ch. 18, pp. 356–372.

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

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

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.

Perrett, D.I., Rolls, E.T. and Caan, W., 1982. Visual neurons responsive to faces in the monkey temporal cortex. Exp. 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. Roy. Soc., 223B: 293–317.

Perrett, D.I., Mistlin, A.J. and Chitty, A.J., 1987. Visual neurons responsive to faces. Trends in Neurosc., 10: 358–364.

Poggio, T., 1990. A theory of how the brain might work. Cold Spring Harbor Symposia in Quantitative Biology, 55: 899–910.

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.

Rolls, E.T., 1981a. Processing beyond the inferior temporal visual cortex related to feeding, learning, and striatal function. In: Brain Mechanisms of Sensation, Y. Katsuki, R. Norgren and M. Sato (Editors), New York, Wiley: Ch 16, pp. 241–269.

Rolls, E.T., 1981b. Responses of amygdaloid neurons in the primate. In: The Amygdaloid Complex, Y. Ben-Ari (Editor), Amsterdam, Elsevier: pp. 383–393.

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., 1986a. A theory of emotion, and its application to understanding the neural basis of emotion. In: Emotions. Neural and Chemical Control, Y. Oomura (Editor), Tokyo and Karker, Basel, Japan Scientific Societies Press: pp. 325–344.

Rolls, E.T., 1986b. Neural systems involved in emotion in primates. In: Emotion: Theory, Research, and Experience, R. Plutchik and H. Kellerman (Editors), Volume 3, Biological Foundations of Emotion, New York, Academic Press: Ch 5, pp. 125–143.

Rolls, E.T., 1987. Information representation, processing and storage in the brain: analysis at the single neuron level. In: The Neural and Molecular Bases of Learning, J.-P. Changeux and M. Konishi (Editors), Chichester, Wiley: pp. 503–540.

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

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

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

Rolls, E.T., 1990a. The representation of information in the temporal lobe visual cortical areas of macaques. In: Advanced Neural Computers, R. Eckmiller (Editor), Amsterdam, North-Holland: pp. 69–78.

Rolls, E.T., 1990b. A theory of emotion, and its application to understanding the neural basis of emotion. Cog. and Emot., 4: 161–190.

Rolls, E.T., 1991. Neural organisation of higher visual functions. Curr. Op. Neurobiol., 1: 274–278.

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

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

Rolls, E.T., 1992c. The processing of face information in the primate temporal lobe. In: Processing Images of Faces, V. Bruce and M. Burton (Editors), Ablex, Norwood, New Jersey: Ch 3, pp. 41–68.

Rolls, E.T., 1994. A theory of emotion and consciousness, and its application to understanding the neural basis of emotion. Ch. 72, pp. 1091–1106 in: The Cognitive Neurosciences, M.S. Gazzaniga (Editor), MIT press, Cambridge, MA.

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. Vis. 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. Exp. 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. Exp. 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. Abs., 19: 27.

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. Exp. Brain Res., in press.

Rolls, E.T. and Tovee, M.J., 1994a. Processing speed in the cerebral cortex, and the neurophysiology of visual backward masking. Proc. Roy. Soc. B., 257: 9--15.

Rolls, E.T. and Tovee, M.J., 1994b. 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. Exp. Brain Res., in press.

Rolls, E.T. and Tovee, M.J., 1994c. The sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. J. Neurophysiol., in press.

Rolls, E.T., Hornak, J., Wade, D. and McGrath, J., 1994. Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. J. Neurol. Neurosurg. Psychiat.

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.

Smith, D.V. and Travers, J.B., 1979. A metric for the breadth of tuning of gustatory neurons. Chem. Sens., 4: 215–229.

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, E. Iwai and M. Mishkin (Editors), New York, Elsevier: Ch 10, pp. 101–109.

Thorpe, S.J. and Imbert, M. 1989 Biological constraints on connectionist models. In: Connectionism in Perspective, edited by R. Pfeifer, Z. Schreter, and F. Fogelman-Soulie. Amsterdam: Elsevier, p. 63- 92.

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., in press.

Tovee, M.J. and Rolls, E.T., 1994. Information encoding in short firing rate epochs by single neurons in the primate temporal visual cortex. Visual Cognition, in press.

Treves, A. and Rolls, E.T., 1991. What determines the capacity of autoassociative memories in the brain? Network, 2: 371–397.

Treves, A. and Rolls, E.T., 1994. A computational analysis of the role of the hippocampus in memory. Hippocampus, in press.

Treves, A., Rolls, E.T. and Tovee, M.J., 1994. On the time required for recurrent processing in the brain. in preparation.

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

Wallis, G., Rolls, E.T. and Foldiak, P., 1993. Learning invariant responses to the natural transformations of objects. Int. Joint Conf. on Neural Net., 2: 1087–1090.

Williams, G.V., Rolls, E.T., Leonard, C.M. and Stern, C., 1993. Neuronal responses in the ventral striatum of the behaving macaque. Behav. Brain Res., 55: 243–252.

Wilson, F.A.W., O'Scalaidhe, S.P and Goldman-Rakic, P.S., 1993. Dissociation of object and spatial processing domains in primate preforontal cortex. Science 260: 1955--1958.

Yamane, S., Kaji, S. and Kawano, K., 1988. What facial features activate face neurons in the inferotemporal cortex of the monkey? Exp. Brain Res., 73: 209–214.

Young, M.P. and Yamane, S., 1992, Sparse population encoding of faces in the inferotemporal cortex. Science 256: 1327- 1331.