Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas

EDMUND T. ROLLS

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

SUMMARY

The ways in which information about faces is represented and stored in the temporal lobe visual areas of primates, as shown by recordings from single neurons in macaques, are considered. Some neurons that respond primarily to faces are found in the cortex in the anterior part of the superior temporal sulcus (in which neurons are especially likely to be tuned to facial expression and to face movement involved in gesture), and in the TE areas more ventrally forming the inferior temporal gyrus (in which neurons are more likely to have responses related to the identity of faces). 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 rather than 'grandmother cell' encoding is used. It is argued that this type of tuning is a delicate compromise between very fine tuning, which has the advantage of low interference in neuronal network operations but the disadvantage of losing the useful properties (such as generalization, completion and graceful degradation) of storage in neuronal networks, and broad tuning, which has the advantage of allowing these properties of neuronal networks to be realized but the disadvantage of leading to interference between the different memories stored in an associative network. There is evidence that the responses of some of these neurons are altered by experience so that new stimuli become incorporated in the network. 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 that is built is object based, which is suitable for recognition and as an input to associative memory, and that another is viewer centred, which is appropriate for conveying information about gesture. Ways are considered in which such cortical representations might be built by competitive self-organization aided by backprojections in the multi-stage cortical porcessing hierarchy which has convergence from stage to stage.

1. INTRODUCTION

In this paper some of the ways in which information about faces is processed by the brain will be considered. First, empirical studies largely based on recordings of the activity of single neurons in the brain will be considered. The recordings in these studies are made mainly in non-human primates because the temporal lobe, in which this processing occurs, is much more developed than in non-primates, and because the findings are directly relevant to understanding the effects of damage to the temporal lobes in clinical patients. Second, some of the implications of these studies for the computations being done by neuronal networks underlying visual recognition will be considered.

2. THE STAGES IN VISUAL INFORMATION PROCESSING AT WHICH NEURONS RESPOND SELECTIVELY TO FACES

Visual pathways project by a number of cortico-cortical stages from the primary visual cortex, the striate cortex, until they reach the temporal lobe visual cortical areas (Seltzer & Pandya 1978; Maunsell & Newsome 1987; Baizer *et al.* 1991) in which some neurons that respond selectively to faces are found (Desimone & Gross 1979; Bruce *et al.* 1981; Desimone *et al.* 1984; Gross *et al.* 1985; Rolls 1981*a*, *b*, 1984; Perrett *et al.* 1982; Rolls 1991*a*; 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, and in addition there are a set of different areas in the cortex in the superior temporal sulcus (Seltzer & Pandya 1978; Baylis *et al.* 1987) (see figure 1). Of these last areas, TP0 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 & Pandya 1978).

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

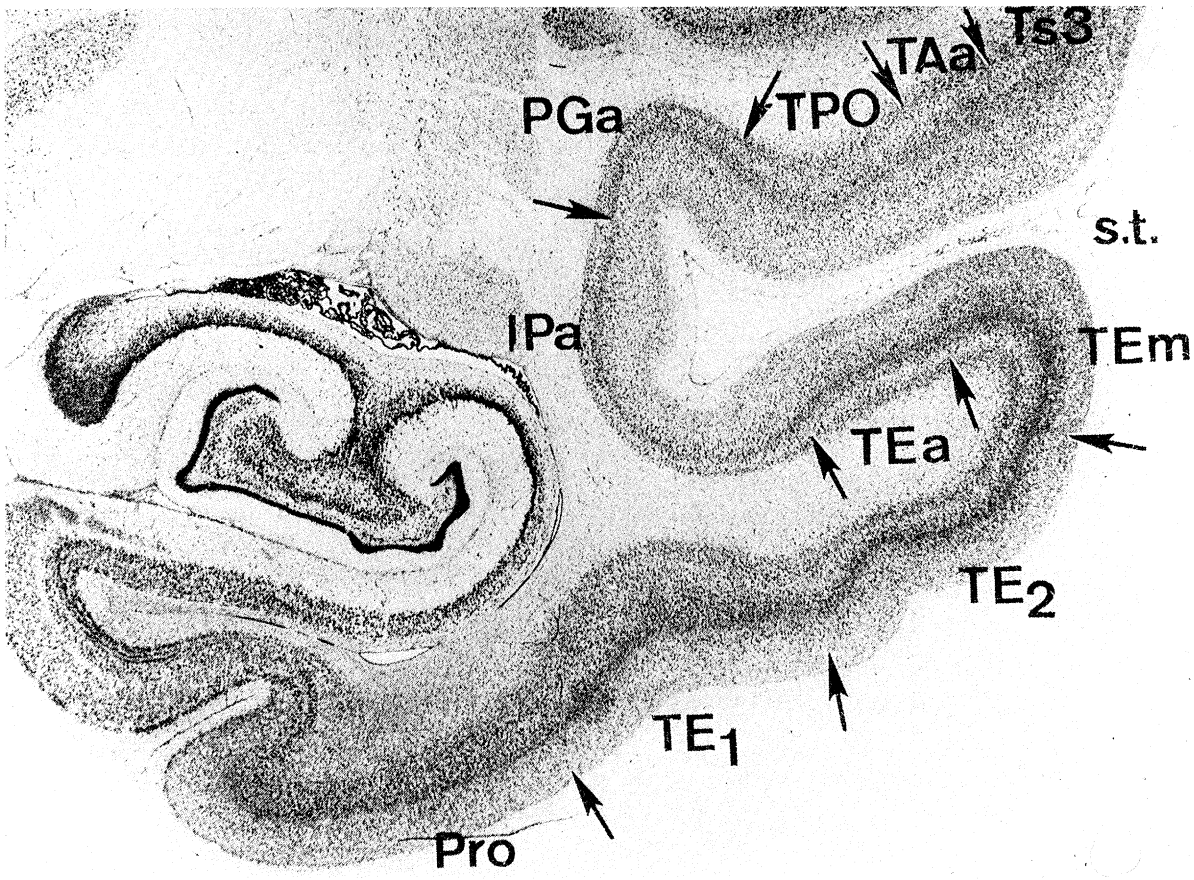

Figure 1. Coronal section showing the different architectonic areas in and bordering the anterior part of the superior temporal sulcus (s.t.) of the macaque (see text). Areas TE1, TE2 and TEm form part of the inferior temporal gyrus.

of simple and complex visual stimuli such as sine wave gratings, three-dimensional objects and faces, and also auditory and somatosensory stimuli (Baylis *et al.* 1987). Considerable specialization of function was found. For example, areas TPO, PGa and IPa are multimodal, with neurons which respond to visual, auditory and 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; 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 proportions of different types of neurons found in these areas were shown statistically to be significantly different.) 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 regions corresponding with 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). Because of the wide distribution of these neurons, 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 *et al.* 1989). These neuro-physiological 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.

3. THE SELECTIVITY OF ONE POPULATION OF NEURONS FOR FACES

The neurons described in our studies as having responses selective for faces do so 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 three-dimensional objects (see Rolls 1984, 1991*a*; Baylis *et al.* 1987). (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 figure 6 of Rolls (1991*a*), and the criteria for classification as face selective are elaborated further therein.) The responses to faces are excitatory, sustained and are time-locked to the stimulus presentation with a latency of between 80 ms 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 colour, size, distance and contrast (see below). These findings suggest 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 & 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). The differing degrees of selectivity of these neurons for faces, and the different ways in which the neurons are sensitive to faces, are consistent with the hypotheses described below, that these neurons acquire their particular pattern of selectivity by a process of competitive self-organization guided by the statistics of the differing visual stimuli actually present in the environment of primates.

4. 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-organization 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.

5. 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, or whether instead the firing of a group or ensemble of cells, each with somewhat different responsiveness, provides the representation (Barlow 1972). 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. To quantify how finely these neurons were tuned to the faces of particular individuals, a measure derived from information theory, the breadth of tuning metric developed by Smith & Travers (1979), was calculated. 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†. The breadth of tuning of the majority of the neurons analysed 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 five different faces) (Baylis *et al.* 1985). Such evidence shows that only across a population or ensemble of such cells is information conveyed which would be useful in making different behavioural responses to different faces. This information about which individual was being seen was very significant, as shown by the result 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). Expressed quantitatively, n neurons could be considered as an n-dimensional vector capable of indicating a unique region in an n-dimensional space if the neurons have independent responses, or as indicating a unique region in a space of somewhat less than n dimensions if the neuronal responses are somewhat dependent, with each neuron tuned so that its response decreases smoothly as the input stimulus becomes less similar to that which elicits the maximal response from the neuron (cf. Poggio 1990).

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, in that anatomically there do not appear to be a whole further set of visual processing areas present in the brain; from the temporal lobe visual

† $H = -k \sum_{i=1,n} p_i \log p_i$,

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.

areas such as those described, outputs 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 unique) coding still being present (Leonard *et al.* 1985; Rolls 1991*c*). The amygdala in turn projects to another structure that may be important in other behavioural responses to faces, the ventral striatum, and comparable neurons have also been found in the ventral striatum (Rolls & Williams 1987).

Given this empirical evidence for ensemble encoding, that is for distributed representation, in these parts of the visual system, we can ask what the advantages are of this type of coding. It has been suggested (Rolls 1987, 1989*a*, *b*, 1991*a*; Rolls & Treves 1990) that the advantages 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. 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, 1989*a*, *b*). The type of tuning found in at least some temporal lobe cortical visual areas is thus suggested to be a delicate compromise between very fine tuning, which has the advantage of low interference in neuronal network operations and a high number of stored memories but the disadvantage of losing the emergent properties of storage in neuronal networks, and broad tuning, which has the advantage of allowing the emergent properties of neuronal networks to be realized but of leading to interference between the different memories stored in the network (see Rolls 1987, 1989*a*, *b*; Rolls & Treves 1990).

The finding that some neurons in these temporal cortical regions are tuned to reflect using ensemble encoding the differences between the faces of different individuals is consistent with the hypotheses described below that competitive self-organization is important in the development of the responses shown by these neurons.

6. THE DEVELOPMENT OF SPECIFICITY OF THE NEURONAL RESPONSES

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 has 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 the 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 that 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). 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 categorization being done by self-organizing competitive neuronal networks, as described below and elsewhere (Rolls 1989*a*–*c*).

7. A NEURONAL REPRESENTATION OF FACES SHOWING INVARIANCE

One of the major problems that 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 & Baylis 1986). It was shown that the majority of these neurons had responses that 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 a three-dimensional to a two-dimensional 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 in direction. The point in 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 & 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 eight 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) band-pass 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, so that these neurons do not have linear properties (Rolls *et al.* 1987). The lack of linearity of these neurons, and their responsiveness to a wide range of spatial frequencies, suggest 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 when these cortical regions are operating normally 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 (Azzopardi & Rolls 1989). It was found that in many cases 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 3–20 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 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 that 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.

8. A VIEW-INDEPENDENT (OR OBJECT-BASED) REPRESENTATION OF VISUAL INFORMATION

The result that the responses of these neurons show considerable invariance over changes in the size, contrast and retinal translation of face stimuli is evidence that the encoding is not in retinal coordinates. Alternatives are that it is in viewer-centered or object-centered coordinates.

Some neurons with responses selective for faces only respond if the face is moving (Perrett *et al.* 1985). It is noted that encoding of gestures is likely to be important in the social behaviour of primates. We took advantage of the fact that these neurons respond to moving faces to investigate whether the encoding of faces by these neurons is viewer centred or object based (Hasselmo *et al.* 1986*a*, 1989). For ten neurons it has been shown that the neuron responds to particular movements that can only be described in object-based coordinates. 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 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 & Baylis 1986).

Further evidence supporting the hypothesis that some of the neurons in this region use object-based descriptions is that their selectivity between the faces of different individuals is maintained across anisomorphic transforms of the stimulus. For example, some neurons reliably responded differently to the faces of two different individuals independently of viewing angle. However, in most cases (16 out of 18 neurons), although the identity of the face was reflected in the neuronal response, the response was not perfectly view independent, and viewing angle also influenced the response (Hasselmo *et al.* 1989). Now, mixed together in the same cortical regions there are neurons with view-dependent responses. 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. These findings, of view-dependent, partly view-independent, and view-independent or object-based 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.

9. DIFFERENT NEURONS 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 1981*b*, 1984, 1986*a*, *b*, 1990*b*), neurons were tested with facial stimuli which included examples of the same individual monkey with different facial expressions (Hasselmo *et al.* 1986*b*; Hasselmo *et al.* 1989). The responses of 45 neurons with responses selective for faces were measured to a set of three 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 nine neurons showed responses which depended on expression but were independent of identity, as measured by a two-way ANOVA. The neurons responsive to expression were found primarily in the cortex in the superior temporal sulcus, but 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 1981*b*, 1984, 1986*a*, *b*, 1990*b*). Damage to this population may contribute to the deficits in social and emotional behaviour which are part of the Kluver-Bucy syndrome produced by temporal lobe damage in monkeys (see Rolls 1981*b*, 1984, 1986*a*, *b*, 1990*b*, 1991*a*, *c*; Leonard *et al.* 1985).

Another 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. 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). 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 the viewer's 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 *et al.* 1989*a*), head and face movement involved in, for example, gesture (Hasselmo *et al.* 1989*b*) and eye gaze (Perrett *et al.* 1985*b*), 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 *et al.* 1989*a*) and with view-independent responses (Hasselmo *et al.* 1989*b*) 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 to make the correct social or emotional response.

10. POSSIBLE COMPUTATIONAL MECHANISMS IN THE VISUAL CORTEX FOR OBJECT RECOGNITION

The neurophysiological findings described above (see also Rolls 1990*a*, 1991*a*), and wider considerations on the possible computational properties of the cerebral cortex (Rolls 1989*a*, *b*), 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 recognize 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 suclus, 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, for example, 8 degrees in V4, 20 degrees in TEO, and 50 degrees in inferior temporal cortex (Boussaoud *et al.* 1991)) (see figure 2). Such zones of convergence would overlap continuously with each other (see figure 2). This connectivity would be part of the architecture by which translation invariant representations are computed. Each layer is considered to act partly as a set of local self-organizing competitive neuronal networks with overlapping inputs. (The region within which competition would be implemented would depend on the spatial properties of inhibitory interneurons, and might operate over distances of 1–2 mm in the cortex.) These competitive nets operate by a single set of forward inputs leading to (typically nonlinear, e.g. sigmoid) activation of output neurons; of competition between the output

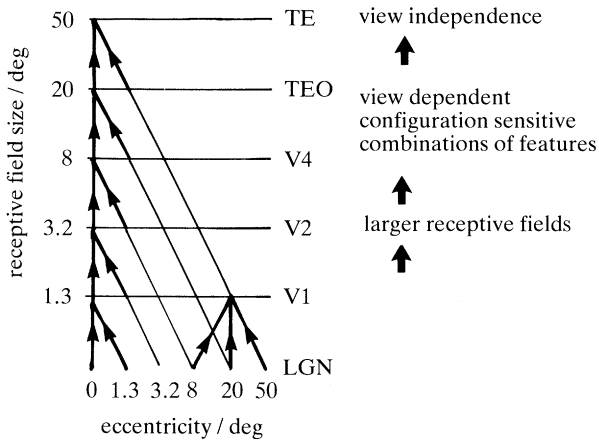

Figure 2. Schematic diagram showing convergence achieved by the forward projections in the visual system, and the types of representation that may be built by competitive networks operating at each stage of the system from the primary visual cortex (V1) to the inferior temporal visual cortex (area TE) (see text). LGN, lateral geniculate nucleas. Area TE0 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).

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 1989*c*). (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 nonlinear function of the output activation which mimics the operation of the NMDA receptors in learning (see Rolls 1989*c*, 1991*a*).) 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 categorizers. In relation to visual information processing, they would remove redundancy from the input representation, and would develop low entropy representations of the information (cf. Barlow 1985; Barlow *et al.* 1989). I suggest competitive nets for this rather than the decorrelation nets suggested by Barlow & Foldiak (1989) because of biological plausibility. (Hebb-modifiable forward excitatory connections, with competitive inhibition mediated by the relatively small proportion of cortical inhibitory neurons with relatively small numbers of synapses, within each stage, is more plausible than postulating a large number of inhibitory connections which directly link cortical neurons and can be modified by a rule which results in an increase in the strength of the inhibitory synapses between coactive cortical pyramidal cells.) 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 1989*a–c*), and has the advantages that the output neurons learn better to distribute themselves between the input patterns (cf. Bennett 1990), and that the sparse representations formed (which provide 'coarse coding') have utility in maximizing the number of memories that can be stored when, towards the end of the visual system, the visual representation of objects is interfaced to associative memory (Rolls 1989*a*, *b*; Rolls & 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 & Girosi (1990*a*,*b*). 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 back projections (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 generalize smoothly from the examples with which it has learned (cf. Poggio & Girosi 1990*a*, *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 utilizing competitive learning to detect regularities in inputs when real objects are translated in the physical world. The hypothesis is that because objects have continuous properties in the world, an object at one place on the retina might activate feature analysers at the next stage of cortical processing, and when the object was translated to a nearby position, because this would occur in a short period (e.g. 0.5 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, but the 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. 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, and then scale invariance to 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. To avoid the combinatorial explosion it is proposed, following Feldman (1985), that low-order combinations of inputs would be 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*b*); 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) 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 that 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 face neurons require the face features to be in the correct spatial configuration, and not jumbled.) 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 (cf. von der Malsburg 1990). By this stage of processing a view-dependent representation of objects suitable for view-dependent processes such as behavioural responses to face expression and gesture would be available.

It is suggested that view-independent representations could be formed by the same type of computation, operating to combine a limited set of views of objects. The plausibility of providing view-independent recognition of objects by combining a set of different views of objects has been proposed by a number of investigators (Koenderink & Van Doorn 1979; Poggio & Edelman 1990). Consistent with the suggestion that the view-independent representations are formed by combining view-dependent representations in the primate visual system is 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 they could therefore receive) (Hasselmo *et al.* 1989; Perrett *et al.* 1987). This solution to 'object-based' representations is very different from that traditionally proposed for some artificial vision systems, in which the coordinates in three-dimensional space of objects are stored in a database, and general-purpose algorithms operate on these to do transforms such as translation, rotation, and scale change in three-dimensional space. In the present, much more limited but more biologically plausible, scheme the representation would be suitable for recognition of an object, and for linking associative memories to objects, but would be less good for making actions in three-dimensional space to particular parts of, or inside, objects, as the three-dimensional coordinates of each part of the object would not be explicitly available. It is 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 (cf. Ballard 1990).

The computational processes proposed above operate by an unsupervised learning mechanism, which utilizes regularities in the physical environment to enable representations with low entropy to be built. In some cases it may be advantageous to utilize some form of mild teaching input to the visual system, to enable it to learn, for example, that rather similar visual inputs have very different consequences in the world, so that different representations of them should be built. In other cases it might be helpful to bring representations together, if they have identical consequences, to use storage capacity efficiently. It is proposed elsewhere (Rolls 1989*a*, *b*) that the back projections from each adjacent cortical region in the hierarchy (and from the amygdala and hippocampus to higher regions of the visual system) have 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 back-projecting neurons to the principal (pyramidal) neurons of the competitive networks in the preceding stages (Rolls 1989*a*, *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 generalize 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, and 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, to allow recognition of many individual faces when all share a number of common properties.

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. & Rolls, E.T. 1989 Translation invariance in the responses of neurons in the inferior temporal visual cortex of the macaque. *Soc. Neurosci. Abstr.* **15**, 120.

Baizer, J.S., Ungerleider, L.G. & 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* (ed. R. Eckmiller), pp. 229–236. Amsterdam: North-Holland.

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* (ed. D. Rose & V. G. Dobson), pp. 37–46. Chichester: Wiley.

Barlow, H.B. & Foldiak, P. 1989 Adaptation and decorrelation in the cortex. In *The Computing neuron* (ed. R. Durbin, C. Miall & G. Mitchison), pp. 54–72. Wokingham: Addison-Wesley.

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

Baylis, G.C., Rolls, E.T. & Leonard, C.M. 1985 Selectivity between faces in the responses of a population of neurons in the cortex in the superior temporal sulcus of the monkey. *Brain Res.* **342**, 91–102.

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

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

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

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

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

Desimone, R. & 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. & 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 Computat.* **3**, 194–200.

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

Hasselmo, M.E., Rolls, E.T. & Baylis, G.C. 1986*a* Object-centered encoding of faces by neurons in the cortex in the superior temporal sulcus of the monkey. *Society for Neuroscience Abstracts* **11**, 1369.

Hasselmo, M.E., Rolls, E.T. & Baylis, G.C. 1986*b* Selectivity between facial expressions in the responses of a population of neurons in the superior temporal sulcus of the monkey. *Neuroscience Letters* **S26**, S571.

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

Koenderink, J.J. & Van Doorn, A.J. 1979 The internal representation of solid shape with respect to vision. *Biol. Cyber.* **32**, 211–217.

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

Malsburg, C. von der 1990 A neural architecture for the representation of scenes. In *Brain organization and memory: cells, systems and circuits* (ed. J. L. McGaugh, N. M. Weinberger & G. Lynch), pp. 356–372. New York: Oxford University Press.

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

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

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

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

Poggio, T. 1990 A theory of how the brain might work. *Cold Spring Harb. Symp. Quant. Biol.* **55**, 899–910.

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

Poggio, T. & Girosi, F. 1990*a* Regularization algorithms

for learning that are equivalent to multilayer networks. *Science, Wash.* **247**, 978–982.

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

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

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

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

Rolls, E.T. 1986*a* A theory of emotion, and its application to understanding the neural basis of emotion. In *Emotions. Neural and chemical control* (ed. Y. Oomura), pp. 325–344. Tokyo and Karker, Basel: Japan Scientific Societies Press.

Rolls, E.T. 1986*b* Neural systems involved in emotion in primates. In *Emotion: theory, research, and experience* (ed. R. Plutchik & H. Kellerman), vol. 3. *Biological Foundations of Emotion*, pp. 125–143. New York: Academic Press.

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* (ed. J.-P. Changeux & M. Konishi), pp. 503–540. Chichester: Wiley.

Rolls, E.T. 1989*a* Functions of neuronal networks in the hippocampus and neocortex in memory. In *Neural models of plasticity: experimental and theoretical approaches* (ed. J. H. Byrne & W. O. Berry), pp. 240–265. San Diego: Academic Press.

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

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

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

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

Rolls, E.T. 1991*a* The processing of face information in the primate temporal lobe. In *Processing images of faces* (ed. V. Bruce & M. Burton). Norwood, New Jersey: Ablex.

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

Rolls, E.T. 1991*c* Neurophysiology and functions of the primate amygdala. In *The amygdala* (ed. J. P. Aggleton). New York: Wiley.

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

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

Rolls, E.T. & Williams, G.V. 1987 Sensory and movement-related neuronal activity in different regions of the primate striatum. In *Basal ganglia and behavior: sensory aspects and motor functioning* (ed. J. S. Schneider & T. I. Lidsky), pp. 37–59. Bern: Hans Huber.

Rolls, E.T., Baylis, G.C. & 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. *Vision Res.* **27**, 311–326.

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

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

Seltzer, B. & 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. & Travers, J.B. 1979 A metric for the breadth of tuning of gustatory neurons. *Chem. Senses* **4**, 215–229.

Tanaka, K., Saito, C., Fukada, Y. & Moriya, M. 1990 Integration of form, texture, and color information in the inferotemporal cortex of the macaque. In *Vision, memory and the temporal lobe* (ed. E. Iwai & M. Mishkin), pp. 101–109. New York: Elsevier.

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

Discussion

A. Cowey (*Department of Experimental Psychology, University of Oxford, U.K.*). The neural network scheme Dr Rolls has presented includes back projections as an important component. In the scheme the back projections are directed to the points of origin of the forward projections. Anatomically there is clear evidence that back projections are directed not only to the origin of the forward projections but to more widespread areas, some of which are clearly different physiological compartments. Examples in visual cortical areas are given by Shipp & Zeki (1989). Might this demonstrated pattern of diffuse back projections be useful to the kind of model Dr Rolls has presented?

Reference

Zeki, S. & Shipp, S. 1989 Modular connections between areas V2 and V4 of macaque visual cortex. *Eur. J. Neurosci.* **1,** 494–504.

E. T. Rolls. There is indeed a body of evidence suggesting that back projections are more widespread than forward projections (see, for example, Amaral (1987)). A possible functional implication, related to the ideas I presented, of the fact that the forward projections are more concentrated may be that there must be sufficiently strong drive from them to activate neurons in the next cortical areas sufficiently to provide the degree of postsynaptic activation required to support synaptic modification. The back-projecting axons, however, do not need themselves to produce such strong postsynaptic activation of the neurons to which they connect, but only to provide the presynaptic term necessary for synaptic modification if other (forward) inputs have produced strong postsynaptic activation. (Discussion on this issue with A. Treves was helpful.) A further implication is that during recall (produced by back projections) neurons at earlier stages that were active during the original learning can be activated by the back-projecting axons, even if they are not on the forward path to the neurons from which they receive back projections.

Reference

Amaral, D.G. 1987 Memory anatomical organization of candidate brain regions. In *Handbook of neurophysiology. Section 1: the nervous system*, vol. 1, part 1, pp. 211–294. Washington, D.C.: American Physiological Society.

V. Bruce (*Department of Psychology, University of Nottingham, U.K.*) Dr Rolls distinguished 'features' from 'configurations' of features in the encoding of faces. What does he think are the 'features' extracted from faces and does he see these as corresponding to those for which we happen to have labels?

E. T. Rolls. The general thrust of the more theoretical part of my presentation is that the neurons between cortical visual area V1 and the configuration-sensitive face-responsive neurons in the temporal cortical visual areas could respond to low-order combinations of the outputs of V1, formed successively in V2 and V4. Now, given that the responses of at least some V1 neurons can be described as responses of elongated filters of different spatial scales, then one might expect that some neurons in higher areas would respond to combinations of these activity patterns produced in V1 by faces. For example, such a higher-order neuron might respond if two sets of neurons were active in V1, with the sets separated by a small vertical angle and having horizontal orientation sensitivity, as this combination of activity would be produced in V1 by a mouth. Such a neuron might then be described as responding well to the feature of a mouth, but could also be described as having learned as a result of visual experience to respond to frequent combinations of input activity patterns produced at early stages of the visual system by stimuli frequently encountered in the visual world.

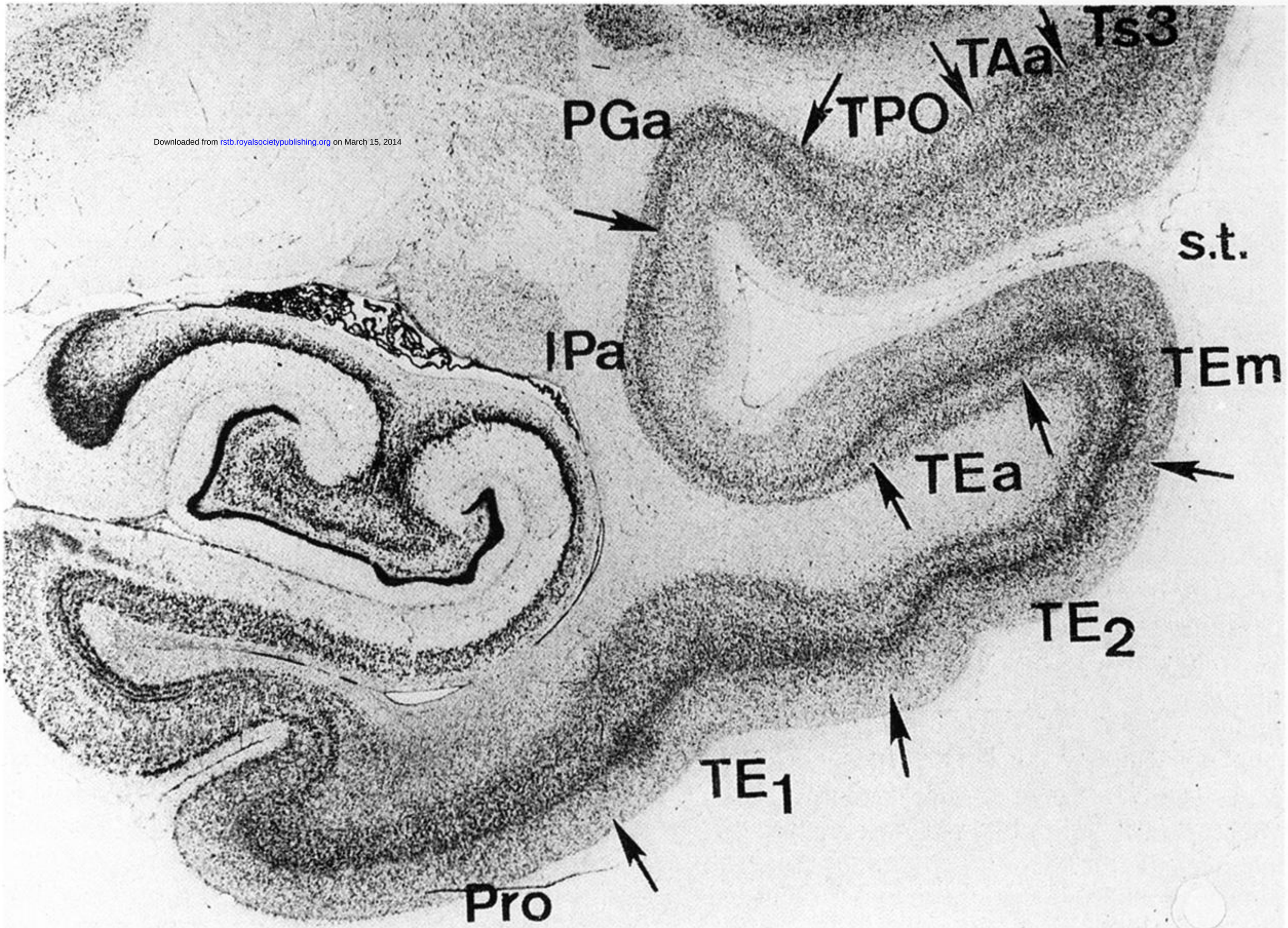

Figure 1. Coronal section showing the different architectonic areas in and bordering the anterior part of the superior temporal sulcus (s.t.) of the macaque (see text). Areas TE1, TE2 and TEm form part of the inferior temporal gyrus.