Neuron, Vol. 27, 205–218, August, 2000, Copyright ©2000 by Cell Press

Functions of the Primate Temporal Lobe Cortical Visual Areas in Invariant Visual Object and Face Recognition

Review

Edmund T. Rolls*
University of Oxford
Department of Experimental Psychology
South Parks Road
Oxford, OX1 3UD
United Kingdom

There is now good evidence that neural systems in temporal cortical visual areas process information about faces. Because a large number of neurons are devoted to this class of stimuli, these systems have proved amenable to experimental analysis. Face recognition and the identification of face expression are important in primate social behavior, and analysis of the neural systems involved is important for understanding the effects of damage to these systems in humans. Damage to these or related systems can lead to prosopagnosia, an impairment in recognizing individuals from the sight of their faces, or to difficulty in identifying the expression on a face. It turns out that the temporal cortical visual areas also have similar neuronal populations that code for objects, and study of both sets of neurons is helping to unravel the enormous computational problem of invariant visual object recognition. The neurophysiological recordings are made mainly in the macaque, a species of nonhuman primate, first because the temporal lobe, in which this processing occurs, is much more developed than in nonprimates; and second because the findings are relevant to understanding the effects of brain damage in patients.

Neuronal Responses Found in Different Temporal Lobe Cortical Visual Areas

While recording in the temporal lobe cortical visual areas of macaques, Charles Gross and colleagues found some neurons that appeared to respond best to complex visual stimuli such as faces (Desimone and Gross, 1979; Bruce et al., 1981; see also Desimone, 1991). It was soon found that while some of these neurons could respond to parts of faces, other neurons required several parts of the face to be present in the correct spatial arrangement, and that many of these neurons did not just respond to any face that was shown but responded differently to different faces (Perrett et al., 1982; Desimone et al., 1984; Rolls, 1984; Gross et al., 1985). By responding differently to different faces, these neurons potentially encode information useful for identifying individual faces. It also appears that there is some specialization of function of different temporal cortical visual areas, as described next.

The visual pathways project from the primary visual cortex to the temporal lobe visual cortical areas by a number of intervening cortical stages (Seltzer and Pandya, 1978; Maunsell and Newsome, 1987; Baizer et al., 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 et al., 1987) (see Figure 1). Of these latter areas, TPO receives inputs from temporal, parietal, and occipital cortex; PGa and IPa from parietal and temporal cortex; and TS and TAa primarily from auditory areas (Seltzer and Pandya, 1978).

There is considerable specialization of function in these architectonically defined areas (Baylis et al., 1987). Areas TPO, PGa, and IPa are multimodal, with neurons that respond to visual, auditory, and/or somatosensory inputs. The more ventral areas in the inferior temporal gyrus (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, where they comprise ~20% of the visual neurons responsive to stationary stimuli, in contrast to the other temporal cortical areas, where they comprise 4%–10%. The stimuli that 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). 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. Moreover, 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 below and Hasselmo et al., 1989a).

The Selectivity of One Population of Neurons for Faces

Neurons with responses selective for faces respond 2–20 times more to faces than to a wide range of gratings, simple geometrical stimuli, or complex 3D objects (see Rolls, 1984, 1992b; Baylis et al., 1985, 1987). The responses to faces are excitatory, with firing rates often reaching 100 spikes/s, and sustained, and they have typical latencies of 80–100 ms. The neurons are typically unresponsive to auditory or tactile stimuli and to the sight of arousing or aversive stimuli. These findings indicate that explanations in terms of arousal, emotional or motor reactions, and simple visual feature sensitivity 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

*E-mail: edmund.rolls@psy.ox.ac.uk

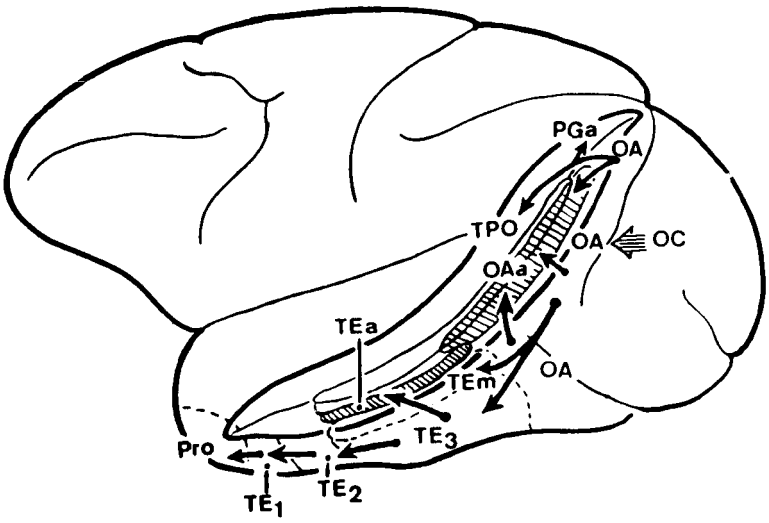

Figure 1. Lateral View of the Macaque Brain

Lateral view of the macaque brain (left) 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).

as edges and bars or to complex nonface stimuli (see also Gross et al., 1985).

These neurons are specialized to provide information about faces in that they provide much more information (on average 0.4 bits) about which (of 20) face stimuli are being seen than about which (of 20) nonface stimuli are being seen (on average 0.07 bits) (Rolls and Tovee, 1995a; Rolls et al., 1997). These information theoretic procedures provide an objective and quantitative way to show what is "represented" by a particular population of neurons.

The Selectivity of These Neurons for Individual Face Features or for Combinations of Face Features

Masking out or presenting parts of the face (e.g., eyes, mouth, or hair) in isolation reveals 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; Rolls et al., 1994a). 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 shows that the combinations formed can reflect not just the features present but also their spatial arrangement. This provides a way in which binding can be implemented in neural networks (see Elliffe et al., 2000b). Further evidence that neurons in these regions respond to combinations of features in the correct spatial configuration was found by Tanaka et al. (e.g., 1990), using combinations of features that are used by comparable neurons to define objects.

Distributed Encoding of Face and Object 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 different profiles of responsiveness to the stimuli provides the representation. A grandmother cell representation is a code which is very sparse, in that each neuron responds to only one object or stimulus. A very large number of neurons would be required, since each neuron responds to only one stimulus. This encoding is described as local, in that all the information that a particular object is present is carried by one neuron. In contrast, ensemble encoding is described as distributed, in that the information that a particular stimulus was shown is distributed across a population of neurons. Many more stimuli can potentially be represented by a distributed code, as each object is represented by a combination of different neurons firing, and this type of code can have many other advantages, as described below. The actual representation found is distributed. Baylis et al. (1985) showed this with the responses of temporal cortical neurons that typically responded to several members of a set of 5 faces, with each neuron having a different profile of responses to each face. In a more recent study using 23 faces and 45 nonface natural images, a distributed representation was found again (Rolls and Tovee, 1995a), with the average sparseness being 0.65. (The sparseness of the representation provided by a neuron can be defined as

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

where r_s is the mean firing rate of the neuron to stimulus s in the set of S stimuli [see Rolls and Treves, 1998]. If the neurons are binary [either firing or not to a given stimulus], then a would be 0.5 if the neuron responded to 50% of the stimuli, and 0.1 if a neuron responded to 10% of the stimuli.) If the spontaneous firing rate was subtracted from the firing rate of the neuron to each stimulus, so that the changes of firing rate, i.e., the *active 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.

The distributed nature of the representation can be further understood by the finding that the firing rate

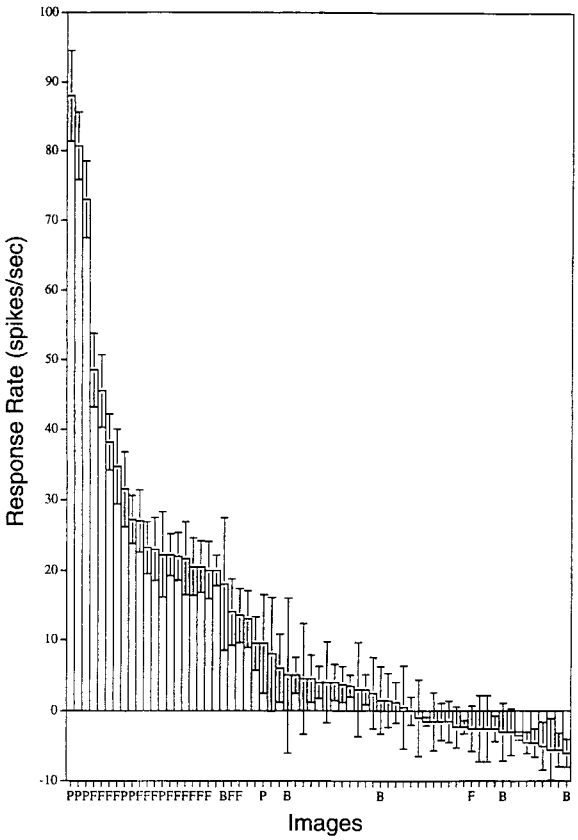

Figure 2. Firing Rate Distribution of a Single Neuron in the Temporal Visual Cortex to a Set of 23 Face (F) and 45 Nonface Images of Natural Scenes

The firing rate to each of the 68 stimuli is shown. The neuron does not respond to just one of the 68 stimuli. Instead, it responds to a small proportion of stimuli with high rates, to more stimuli with intermediate rates, and to many stimuli with almost no change of firing. This is typical of the distributed representations found in temporal cortical visual areas. After Rolls and Tovee (1995a). P, face profile; B, body part.

distribution of single neurons when a wide range of natural visual stimuli are being viewed is approximately exponentially distributed, with rather few stimuli producing high firing rates, and increasingly large numbers of stimuli producing lower and lower firing rates (Rolls and Tovee, 1995a; Baddeley et al., 1997; Treves et al., 1999) (see Figure 2). This is a clear answer to whether these neurons are grandmother cells: they are not, in the sense that each neuron has a graded set of responses to the different members of a set of stimuli, with the prototypical distribution similar to that of the neuron illustrated in Figure 2. On the other hand, each neuron does respond very much more to some stimuli than to many others and in this sense is tuned to some stimuli. The sparseness of such an exponential distribution of firing rates is 0.5. It has been shown that the distribution may arise from the threshold nonlinearity of neurons combined with short-term variability in the responses of neurons (Treves et al., 1999). The distributed properties of the code used are further revealed by applying information theory (see Shannon, 1948; MacKay and McCullough, 1952; Eckhorn and Popel, 1974; Rolls and Treves, 1998 [Appendix 2]) to analyze how information is represented by a population of these neurons. The information required to identify which of S equiprobable stimuli were shown is $\log_2 S$ bits. (Thus, one bit is required to specify which of two stimuli was shown, two bits to specify which of four stimuli was shown, three bits to specify which of eight stimuli was shown, etc.) If the encoding was local (or grandmother cell–like), the number of stimuli encoded by a population of neurons would be expected to rise approximately linearly with the number of neurons in the population. In contrast, with distributed encoding, provided that the neuronal responses are sufficiently independent and reliable (not too noisy), the number of stimuli encodable by the population of neurons might be expected to rise exponentially as the number of neurons in the sample of the population was increased. The information about which of 20 equiprobable faces had been shown that was available from the responses of different numbers of these neurons is shown in Figure 3. First, it is clear (Figure 3) that the information rises approximately linearly, and the number of stimuli encoded thus rises approximately exponentially, as the number of cells in the sample increases (Abbott et al., 1996; Rolls et al., 1997; see also Rolls and Treves, 1998). This direct neurophysiological evidence thus demonstrates that the encoding is distributed, and the responses are sufficiently independent and reliable, such that the representational capacity increases exponentially with the number of neurons in the ensemble. The consequence of this is that large numbers of stimuli, and fine discriminations between them, can be represented without having to measure the activity of an enormous number of neurons. (It has been shown that the main reason why the information tends to asymptote, as shown in Figure 3, as the number of neurons in the sample increases is just that the ceiling is being approached of how much information is required to discriminate between the sets of stimuli, which with 20 stimuli is $\log_2 20 = 4.32$ bits [Abbott et al., 1996].) Second, it is clear that some information is available from the responses of just one neuron—on average, $\sim$0.34 bits. Thus, knowing the activity of just one neuron in the population does provide some evidence about which stimulus was present, even when the activity of all the other neurons is not known. This indicates that information is made explicit in the firing of individual neurons in a way that will allow neuronally plausible decoding, in which a receiving neuron simply uses each of its synaptic strengths to weight the input activity being received from each afferent axon and sums the result over all inputs (see below).

It has recently been shown that there are neurons in the inferior temporal visual cortex that encode view-invariant representations of objects, and for these neurons the same type of representation is found, namely distributed encoding with independent information conveyed by different neurons (Booth and Rolls, 1998).

The analyses just described were obtained with neurons that were not simultaneously recorded, but similar results have now been obtained with simultaneously recorded neurons—that is, the information about which stimulus was shown increases approximately linearly with the number of neurons, showing that the neurons convey information that is nearly independent (Panzeri

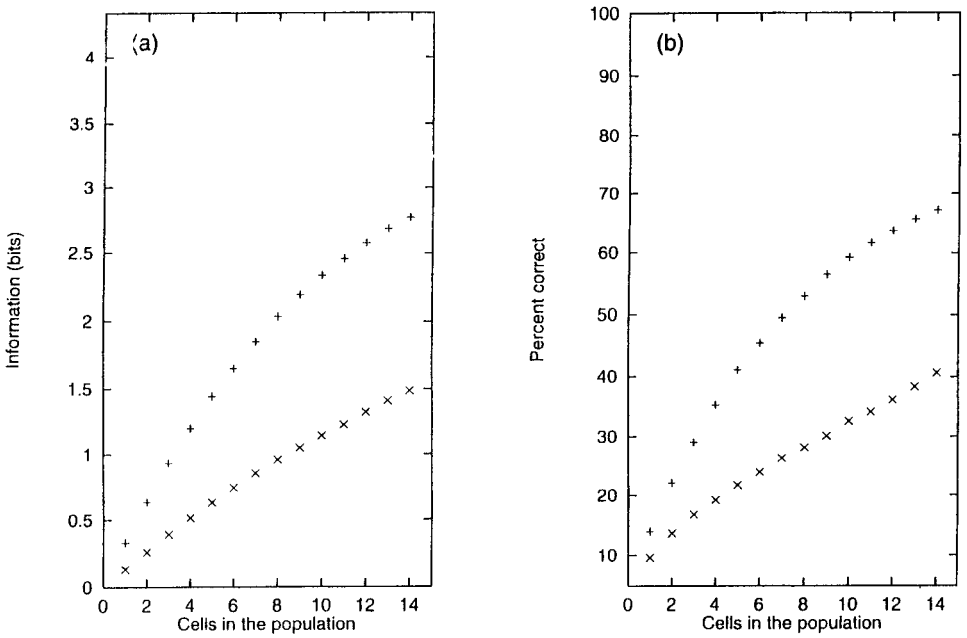

Figure 3. The Average Information from Different Numbers of Inferior Temporal Cortex Neurons about which of 20 Faces Had Been Shown

(a) The values for the average information available in the responses of different numbers of these neurons on each trial, about which of a set of 20 face stimuli has been shown. The decoding method was DP (diamonds) or PE (crosses), and the effects obtained with cross-validation procedures utilizing 50% of the trials as test trials are shown. The remainder of the trials in the cross-validation procedure were used as training trials. The full line indicates the amount of information expected from populations of increasing size, when assuming random correlations within the constraint given by the ceiling (the information in the stimulus set; I = 4.32 bits).

(b) The percent correct for the corresponding data to those shown in Figure 3a. After Rolls et al. (1997).

et al., 1999b; Rolls et al., 1999, Soc. Neurosci., abstract). (Consistently, Gawne and Richmond [1993] showed that even adjacent pairs of neurons recorded simultaneously from the same electrode carried information that was ~80% independent.) Panzeri et al. (1999b) developed a method for measuring the information in the relative time of firing of simultaneously recorded neurons, which might be significant if the neurons became synchronized to some but not other stimuli in a set, as postulated by Singer and colleagues (e.g., Engel et al., 1992). We found that for the set of inferior temporal cortex neurons currently available, almost all the information was available in the firing rates of the cells, and almost no information was available about which static image was shown in the relative time of firing of different simultaneously recorded neurons (Panzeri et al., 1999b; Rolls et al., 1999, Soc. Neurosci., abstract). Consistently, there were no significant cross-correlations between the spikes of these simultaneously recorded inferior temporal cortex neurons. Thus, the evidence is that most of the information is available in the firing rates of the neurons and not in synchronization for representations of faces and objects in the inferior temporal visual cortex (and this is also the case for space in the hippocampus and for odors in the orbitofrontal cortex; see Rolls et al., 1996, 1998).

It is unlikely that there are further processing areas beyond those described where ensemble coding changes into grandmother cell (local) 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, the orbitofrontal cortex, and—via the entorhinal cortex—the hippocampus, where associations between the visual stimuli and other sensory representations are formed (see Rolls and Treves, 1998; Rolls, 1999). Indeed, tracing this pathway onward, we have found a population of neurons with face-selective responses in the amygdala (Leonard et al., 1985; Rolls, 1992a, 2000) and orbitofrontal cortex (Booth et al., 1998, Soc. Neurosci., abstract), and in the majority of these neurons different responses occur to different faces, with ensemble (not local) coding still being present. The amygdala in turn projects to another structure that may be important in other behavioral responses to faces, the ventral striatum, and comparable neurons have also been found in the ventral striatum (Williams et al., 1993).

Advantages for Brain Processing of the Distributed Representation of Objects and Faces

The advantages of the distributed encoding actually found are as follows. (These advantages do not apply to local, that is grandmother cell, encoding schemes, nor to all distributed encoding schemes [see Rolls and Treves, 1998].)

Exponentially High Coding Capacity

This property arises from two factors: (1) the encoding is sufficiently close to independent by the different neurons (i.e., factorial), and (2) the encoding is sufficiently distributed. Part of the biological significance of the exponential encoding capacity is that a receiving neuron

or neurons can obtain information about which one of a very large number of stimuli is present by receiving the activity of relatively small numbers of inputs (in the order of hundreds) from each of the neuronal populations from which it receives. In particular, the characteristics of the actual visual cells described here indicate that the activity of 15 neurons could encode 192 face stimuli (at 50% accuracy), 20 neurons could encode 768 stimuli, 25 neurons could encode 3,072 stimuli, 30 neurons could encode 12,288 stimuli, and 35 neurons could encode 49,152 stimuli (Abbott et al., 1996; the values are for the optimal decoding case). Given that most neurons receive a limited number of synaptic contacts, in the order of several thousand, this type of encoding is ideal. (The capacity of the distributed representations was calculated from ensembles of neurons, with each already shown to provide information about faces. If inferior temporal cortex neurons were chosen at random, 20 times as many neurons would be needed in the sample if face-selective neurons comprised 5% of the population. This brings the number of inputs to each neuron required from an ensemble of sending neurons up to a reasonable number, given the several thousand synapses typically received by each neuron.) This type of encoding (in contrast with local encoding) would enable, for example, neurons in the amygdala and orbitofrontal cortex to learn associations of visual stimuli with reinforcers such as the taste of food when each neuron received a reasonable number, perhaps in the order of hundreds, of inputs from the visually responsive neurons in the temporal cortical visual areas, which specify which visual stimulus or object is being seen (see Rolls, 1990, 1992a, 1992b; Rolls and Treves, 1998). This type of representation is also appropriate for interfacing to the hippocampal system, to allow an episodic memory to be formed that, for example, a particular visual object was seen in a particular place in the environment (Treves and Rolls, 1994; Rolls, 1996; Rolls and Treves, 1998). It is useful to realize that although the sensory representation may have exponential encoding capacity, this does not mean that the associative networks in brain regions such as the amygdala, orbitofrontal cortex, and hippocampus that receive the information can store such large numbers of different patterns. Indeed, there are strict limitations on the number of memories that associative networks can store (Rolls and Treves, 1990, 1998; Treves and Rolls, 1991). The particular value of the exponential encoding capacity of sensory representations is that very fine discriminations can be made, as there is much information in the representation, and that the representation can be decoded if the activity of even a limited number of neurons in the representation is known.

One of the underlying themes here is the neural representation of objects. How would one know that one had found a neuronal representation of objects in the brain? The criterion suggested (Rolls and Treves, 1998) is that when receiving neurons can identify (with neuronally plausible decoding, such as the synaptically weighted sum of inputs described above) the object or stimulus that is present (from a large set of stimuli, thousands or more) from a realistic number of sending neurons, say in the order of 100, then the sending neurons provide a useful representation of the object.

The properties of the representation of faces, of objects (Booth and Rolls, 1998), and of olfactory and taste stimuli have been evident when the readout of the information was done by measuring the *firing rate* of the neurons, typically over a 20, 50, or 500 ms period. Thus, at least where objects are represented in the visual, olfactory, and taste systems (e.g., individual faces, odors, and tastes), information can be read rather efficiently by the receiving neurons without taking into account any aspects of the possible temporal synchronization between neurons (Engel et al., 1992; Rolls et al., 1997; Panzeri et al., 1999b) or temporal encoding within a spike train (Tovee et al., 1993). Thus, rate coding carries an enormous amount of information. The challenge for those who believe that synchronization is important in neuronal coding is to quantify how much extra information it may add (see Panzeri et al., 1999b). In addition, correlations between the firing of different neurons to a set of stimuli do not appear to have any major impact in decreasing what is encoded by small ensembles of neurons in the situation in which there are many objects to be coded for—that is, in which the stimulus space is high dimensional (Panzeri et al., 1999b).

Ease with which the Code Can Be Read by Receiving Neurons

For a code to be plausible, it is a requirement that neurons should be able to read the code. This is why when we have estimated the information from populations of neurons, we have used in addition to a probability estimating measure (PE, optimal, in the Bayesian sense) also a dot product (DP) measure, which is a way of specifying that all that is required of decoding neurons would be the property of adding up postsynaptic potentials produced through each synapse as a result of the activity of each incoming axon (Abbott et al., 1996; Rolls et al., 1997) (see Figure 4). It was found that with such a neuronally plausible decoding algorithm (the DP algorithm), the same generic results were obtained, with only a 40% reduction of information compared to the more efficient (PE) algorithm. This is an indication that the brain could utilize the exponentially increasing capacity for encoding stimuli as the number of neurons in the population increases. For example, by using the representation provided by the neurons described here as the input to an associative or autoassociative memory, which computes effectively the dot product on each neuron between the input vector and the synaptic weight vector, most of the information available would in fact be extracted (see Rolls and Treves, 1990, 1998; Treves and Rolls, 1991). (The 40% reduction with DP as compared to optimal decoding is a minor cost of using this simple type of decoding. The important point is that the number of stimuli still rises exponentially with the number of neurons, so that the number of stimuli that can be represented by a given sample of these neurons is still very large.)

The fundamental point being made here is that the actual code being used in these brain areas can be read efficiently with the simplest type of decoding that neurons are thought to use, taking a sum of the input firings to the neuron with each weighted by the synaptic weight connecting the input to the neuron. It is from the fact that a weighted sum of the firings can be used to read the code about which stimulus was shown that the

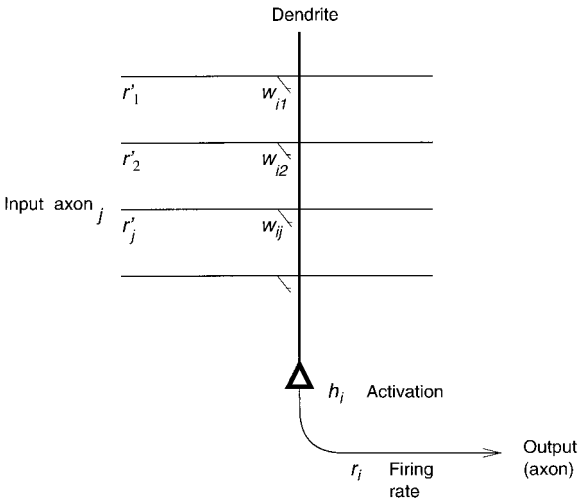

Figure 4. Neuronal Activation and the DP as Used in Models of Associative Memory

When a stimulus is present on the input axons, the total activation h_i of a neuron i is the sum of all the activations produced through each strengthened synapse w_{ij} by each active axon r'_j. We can express this as

$$h_i = \Sigma_j r'_j w_{ij},$$

where Σ_j indicates that the sum is over the C input axons (or connections) indexed by j. The synapse w_{ij} is the j'th synapse (arising from axon j) onto neuron i. The multiplicative form indicates that activation should be produced by an axon only if it is firing, and only if it is connected to the dendrite by a strengthened synapse. The sum of all such activations expresses the idea that summation (of synaptic currents in real neurons) occurs along the length of the dendrite, to produce activation at the cell body, where the activation h_i is converted into firing r_i by a function that includes a threshold. Calculation of the neuronal activation by multiplying the vector of input firings by the vector of synaptic weights is an inner or dot product of two vectors that measures the similarity of the two vectors. It is this computation of similarity (very close to the correlation) between the two vectors that enables neurons to show the interesting properties of generalization, completion, graceful degradation, and resistance to noise, provided that the input representations r′ are distributed (see Rolls and Treves, 1998).

interesting properties of generalization, etc., described next arise. (It may be emphasized that not all distributed codes have this property, so that the finding being discussed is important. If, for example, binary encoding of numbers as in computers were being used [with the bits representing, from the lowest, 1, 2, 4, 8, 16, 32, 64, etc.), then the code would not provide for generalization, completion, etc.)

Generalization, Completion, Graceful Degradation, and Higher Resistance to Noise

Because the decoding of a distributed representation involves assessing the activity of a whole population of neurons, and computing a DP or correlation between the set (or vector) of inputs and the synaptic weights (see Figure 4), a distributed representation provides more resistance to variation in individual components than does a local encoding scheme, and this provides for higher resistance to noise (Panzeri et al., 1996) and for graceful (in that it is gradual) degradation of performance when synapses or input axons are lost. The DP decoding coupled with the type of representation actually found also enables the receiving neuron(s) to generalize to similar stimuli (where the similarity is measured by the number of inputs that correspond) and to complete an incomplete pattern when a network with recurrent collateral connections is used (see Figure 4; Rolls and Treves, 1998). Although these are well-known properties of some neural networks (Willshaw et al., 1969; Kohonen, 1977, 1989; Hopfield, 1982; Bishop, 1995), the important point being made here is that the encoding being used by the inferior temporal visual cortex (and by the olfactory cortex and hippocampus; see Rolls et al., 1996, 1998) is of a type that allows these properties to arise when simple neuronally plausible DP decoding is being used by the receiving neurons.

A point being made here is that the sparse distributed representation found in the inferior temporal cortex as an output stage of the visual system has all the properties required for an input to pattern associators in receiving structures such as the orbitofrontal cortex and amygdala, and autoassociators in structures such as the hippocampus, as the sparseness of the representation allows these associative networks to operate with high memory capacity, and the distributed nature of the representation allows for the properties of generalization, completion, etc., just described. The argument here is completely different from that of Olshausen and Field (1997), who suggested that the code in the primary visual cortex, V1, was sparse because of the sparse feature structure of visual images.

Speed of Readout of the Information

The information available in a distributed representation can be decoded by an analyzer more quickly than can the information from a local representation, given comparable firing rates. Within a fraction of an interspike interval, with a distributed representation, much information can be extracted (Treves, 1993; Rolls et al., 1997; Treves et al., 1997; Panzeri et al., 1999a). In effect, spikes from many different neurons can contribute to calculating the angle between a neuronal population and a synaptic weight vector within an interspike interval. With local encoding, the speed of information readout depends on the exact model considered, but if the rate of firing needs to be taken into account, this will necessarily take time, because of the time needed for several spikes to accumulate in order to estimate the firing rate.

Invariance in the Neuronal Representation of Stimuli

One of the major problems that must be solved by a visual system is the building of a representation of visual information that allows recognition to occur relatively independently of size, contrast, spatial frequency, position on the retina, angle of view, etc. This is required so that if the receiving regions such as the amygdala, orbitofrontal cortex, and hippocampus learn about one view, position, or size of the object, the animal generalizes correctly to other positions, views, and sizes of the object. The majority of face-selective neurons in the inferior temporal cortex have responses that are relatively invariant with respect to the size of the stimulus (Rolls and Baylis, 1986). 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 gray on a monitor, with a response that was on average 0.5 of that to a real face. 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). 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 eight 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) bandpass spatial frequency–filtered face stimuli shows that the responses of these neurons to an unfiltered face cannot 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.

Inferior temporal visual cortex neurons also often show considerable translation (shift) invariance, not only under anesthesia (see Gross et al., 1985), but also in the awake, behaving primate (Tovee et al., 1994). In most cases the responses of the neurons were little affected by which part of the face was fixated, and the neurons responded (with a greater than half-maximal response) even when the monkey fixated 2°–5° beyond the edge of a face that subtended 8°–17° at the retina. Moreover, the stimulus selectivity between faces was maintained this far eccentrically within the receptive field.

Until recently, research on translation invariance considered the case in which there is only one object in the visual field. What happens in a cluttered, natural 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 (cf. Sato, 1989)? If so, the output of the visual system might be confusing for structures which receive inputs from the temporal cortical visual areas. In an investigation of this, it was found that the mean firing rate across all cells to a fixated effective face with a noneffective face in the parafovea (centered 8.5° from the fovea) was 34 spikes/s. On the other hand, the average response to a fixated noneffective face with an effective face in the periphery was 22 spikes/s (Rolls and Tovee, 1995b). 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 noneffective face, even when there are other parafoveal stimuli effective for the neuron. Thus, the neurons provide information biased toward what is present at the fovea and not equally about what is present anywhere in the visual field. This makes the interface to action simpler, in that what is at the fovea can be interpreted (e.g., by an associative memory) partly independently of the surroundings, and choices and actions can be directed if appropriate to what is at the fovea (cf. Ballard, 1993). These findings are a step toward understanding how the visual system functions in a normal environment (see also Gallant et al., 1998; Stringer and Rolls, 2000).

A View-Independent Representation of Faces and Objects

Some temporal cortical neurons reliably responded differently to the faces of two different individuals independently of viewing angle, although in most cases (16/18 neurons) the response was not perfectly view independent (Hasselmo et al., 1989b). 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 (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.

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 et al., 1989b). 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, the movement specificity was maintained across inversion, with neurons 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.

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

Neurons with view-invariant responses of objects seen naturally by macaques have also been described recently (Booth and Rolls, 1998). The stimuli were presented for 0.5 s on a color video monitor while the monkey performed a visual fixation task. The stimuli were images of ten real plastic objects that had been in the monkey's cage for several weeks, to enable him to build view-invariant representations of the objects. Control stimuli were views of objects that had never been seen as real objects. The neurons analyzed were in the TE cortex, in and close to the ventral lip of the anterior part of the superior temporal sulcus. Many neurons were found that responded to some views of some objects. However, for a smaller number of neurons, the responses occurred to only a subset of the objects (using ensemble encoding), irrespective of the viewing angle.

Further evidence consistent with these findings is that some studies have shown that the responses of some visual neurons in the inferior temporal cortex do not depend on the presence or absence of critical features for maximal activation (e.g., Perrett et al., 1982; see Tanaka, 1993, 1996). For example, Mikami et al. (1994) have shown that some TE cells respond to partial views of the same laboratory instrument(s), even when these partial views contain different features. In a different approach, Logothetis et al. (1994) have reported that in monkeys extensively trained (over thousands of trials) to treat different views of computer-generated wire frame "objects" as the same, a small population of neurons in the inferior temporal cortex did respond to different views of the same wire frame object (see also Logothetis and Sheinberg, 1996). However, extensive training is not necessary for invariant representations to be formed, and indeed no explicit training in invariant object recognition was given in the experiment by Booth and Rolls (1998), as Rolls' hypothesis (1992b) is that view-invariant representations can be learned by associating the different views of objects as they are moved and inspected naturally in a period that may be on the order of a few seconds.

Different Neural Systems Are Specialized for Face Recognition and for Face Expression Decoding

Some neurons respond to face identity, and others to face expression (Hasselmo et al., 1989a). 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. Information about facial expression is of potential use in social interactions. Indeed, damage to this population may contribute to the deficits in emotional and social behavior such as tameness and social withdrawal that are part of the Kluver-Bucy syndrome produced by temporal lobe damage in monkeys (see Rolls, 1984, 1999; Leonard et al., 1985).

A further way in which some of these neurons in the cortex in the superior temporal sulcus may be involved in social interactions is that some of them respond to gestures, e.g., to a face undergoing ventral flexion (Perrett et al., 1985a; Hasselmo et al., 1989a). 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., 1989a). These two movements (eye and eyelid lowering) often occur together when a monkey is breaking social contact with another. It is also important when decoding facial expression to retain some information about the direction of the head 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, head and body gesture (Hasselmo et al., 1989b), and eye gaze (Perrett et al., 1985b) representations in some of these cortical regions where face expression is represented is consistent with this requirement. In contrast, the TE areas (more ventral, mainly in the macaque inferior temporal gyrus), in which neurons tuned to face identity (Hasselmo et al., 1989a) and with view-independent responses (Hasselmo et al., 1989b) 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.

As we have seen, 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, 1990, 1992a, 1992b, 1999). 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 respond to facial and body gesture (Brothers et al., 1990). Rolls et al. (2000) have found a number of face-responsive neurons in the orbitofrontal cortex, and they are also present in adjacent prefrontal cortical areas (Wilson et al., 1993; O'Scalaidhe et al., 1999).

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 that 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 behavioral problems such as disinhibited or socially inappropriate behavior (Hornak et al., 1996). A group of patients with lesions outside this brain region, without these behavioral problems, was unimpaired on the expression identification tests. 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 correctly identifying facial (and vocal) expression and in learning associations involving such visual stimuli (Rolls, 1990, 1999; Rolls et al., 1994b; Hornak et al., 1996).

Neuroimaging data, while not being able to address the details of what is encoded in a brain area or how it is encoded, do provide evidence consistent with the neurophysiology that there are different face-processing systems in the human brain. For example, Kanwisher et al. (1997) and Ishai et al. (1999) have shown activation by faces of an area in the fusiform gyrus; Hoffman and Haxby (2000) have shown that distinct areas are activated by eye gaze and face identity; Dolan et al. (1997) have shown that a fusiform gyrus area becomes activated after humans learn to identify faces in complex scenes; and the amygdala (Morris et al., 1996) and orbitofrontal cortex (Blair et al., 1999) may become activated particularly by certain facial expressions.

Learning of New Representations in the Temporal Cortical Visual Areas

To investigate the idea 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, the responses of inferior temporal cortex face-selective neurons have been analyzed while a set of new faces were shown. 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 (one to two) presentations of the set (Rolls et al., 1989). If in a different experiment a single novel face was introduced when the responses of a neuron to a set of familiar faces were being recorded, the responses to the set of familiar faces were not disrupted, while the responses to the novel face became stable within a few presentations. Alteration of the tuning of individual neurons in this way may result in a good discrimination over the population as a whole of the faces known to the monkey. This evidence is consistent with the categorization being performed by self-organizing competitive neuronal networks, as described below and elsewhere (Rolls and Treves, 1998).

Further evidence that these neurons can learn new representations very rapidly comes from an experiment in which binarized black-and-white images of faces that blended with the background were used. These did not activate face-selective neurons. Full gray-scale images of the same photographs were then shown for ten 0.5 s presentations. In a number of cases, if the neuron happened to be responsive to that face, when the binarized version of the same face was shown next, the neurons responded to it (Tovee et al., 1996). This is a direct parallel to the same phenomenon that is observed psychophysically, and provides dramatic evidence that these neurons are influenced by only a very few seconds (in this case, 5 s) of experience with a visual stimulus. We have shown a neural correlate of this effect using similar stimuli and a similar paradigm in a positron emission tomography (PET) neuroimaging study in humans, with a region showing an effect of the learning found for faces in the right temporal lobe and for objects in the left temporal lobe (Dolan et al., 1997).

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. In addition, some of these neurons may be involved in a short-term memory for whether a particular familiar visual stimulus (such as a face) has been seen recently. The evidence for this is that some of these neurons respond differently to recently seen stimuli in short-term visual memory tasks (Baylis and Rolls, 1987; Miller and Desimone, 1994; Xiang and Brown, 1998), and neurons in a more ventral cortical area respond during the delay in a match-to-sample task with a delay between the sample stimulus and the to-be-matched stimulus (Miyashita, 1993; Renart et al., 2000).

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 are ~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). In a first approach to the very rapid processing apparently being performed by each cortical stage, we measured the information available in short temporal epochs of the responses of temporal cortical face-selective neurons about which face had been seen. 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 poststimulus period of 100–120 ms. Moreover, 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). A comparable result is also found with a much larger set of stimuli: with 20 faces, 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, 1995). This analysis shows that there is considerable information about which stimulus has been seen in short time epochs of the responses of temporal cortex visual neurons.

The next type of experiment shows that very short periods of firing are *sufficient* for a cortical stage to perform its computation. This experiment used a backward masking paradigm, in which 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). The mask stimulus impairs or masks the perception of the test stimulus. When there is no mask, inferior temporal cortex neurons respond to a 16 ms presentation of the test stimulus for 200–300 ms, far longer than the presentation time (Rolls and Tovee, 1994). This reflects the operation of a short-term memory system implemented in the cortical circuitry. If the pattern mask followed the onset of the test face stimulus by 20 ms (a stimulus onset asynchrony of 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, 1994; Rolls et al., 1999). Under these conditions (a test-to-mask stimulus onset asynchrony of 20 ms), human observers looking at the same displays could just identify which of six faces was shown (Rolls et al., 1994a).

These results provide evidence that a cortical area can perform the computation necessary for the recognition of a visual stimulus in 20–30 ms (although it is true that the stimuli did not look very clear), and emphasizes just how rapidly cortical circuitry can operate. Although this speed of operation does seem fast for a network with recurrent connections (mediated, e.g., by recurrent collateral connections between pyramidal cells or by inhibitory interneurons), recent analyses of integrate-and-fire networks with biophysically modeled neurons that integrate their inputs and have spontaneous activity to keep the neurons close to the firing threshold show that such networks can settle very rapidly (Treves, 1993; Treves et al., 1997; Rolls and Treves, 1998). This approach has been extended to multilayer networks such

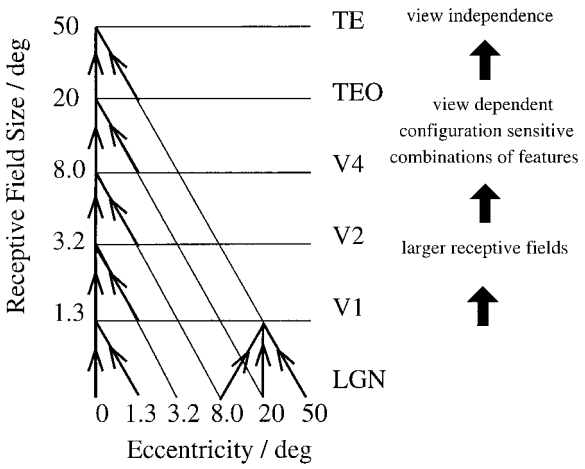

Figure 5. Receptive Field Size and Processing in the Primate Ventral Visual System

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).

as those found in the visual system, and again very rapid propagation (in 50–60 ms) of information through such a four-layer network with recurrent collaterals operating at each stage has been found (Panzeri et al., 2000).

Possible Computational Mechanisms in the Visual Cortex for Object Recognition

The neurophysiological findings described above, and wider considerations on the possible computational properties of the cerebral cortex (Rolls, 1989, 1992b; Rolls and Treves, 1998), lead to the following outline working hypotheses on object (including face) recognition by visual cortical mechanisms.

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). There is convergence from each small part of a region to the succeeding region (or layer or stage in the hierarchy) in such a way that the receptive field sizes of neurons (e.g., 0.5°–1° near the fovea in V1) become larger by a factor of ~2.5 with each succeeding stage (and the typical parafoveal receptive field sizes found would not be inconsistent with the calculated approximations of, e.g., 8° in V4, 20° in TEO, and 50° in inferior temporal cortex [Boussaoud et al., 1991]) (see Figure 5). Such zones of convergence would overlap continuously with each other (see Figure 5). This connectivity would be part of the architecture by which translation-invariant representations are computed.

Each stage in the hierarchy is considered to act partly as a set of local self-organizing competitive neuronal networks with overlapping inputs. The operation of competitive networks is described by Kohonen (1989) and Rolls and Treves (1998). They use competition implemented by lateral inhibition and associative modification of active inputs onto output neurons that are left firing after the competition. Competitive networks can be thought of as building feature analyzers, in that each neuron in a competitive network uses associative synaptic modification to learn to respond to a set or combination of coactive inputs to the neuron, which might represent a visual feature (see Wallis and Rolls, 1997; Rolls and Treves, 1998; Rolls and Milward, 2000).

Increasing complexity of representations could also be built in such a multiple-layer hierarchy by similar competitive learning mechanisms. In order to avoid a combinatorial explosion, low-order combinations of inputs would be learned by each neuron. 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 (i.e., the line must terminate in the receptive field), to tongues flanked by inhibitory subregions, or to combinations of colors (see references cited by Rolls, 1991); in posterior inferior temporal cortex to stimuli that may require two or more simple features to be present (Tanaka et al., 1990); and in the temporal cortical face-processing areas only to images in which several features in a face (such as eyes, hair, and mouth) are present (see above and Yamane et al., 1988). 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 et al. [1990] and with our data on face neurons, in that some face neurons require the face features to be in the correct spatial configuration and not jumbled [Rolls et al., 1994a].) 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 that this mechanism would provide for the feature binding problem (Elliffe et al., 2000b).

Although hierarchical processing schemes have been investigated before (e.g., Fukushima, 1980, 1989, 1991), Rolls (1992b) suggested that translation, size, and view invariance could be computed in such a system by utilizing competitive learning that operates across short time scales to detect regularities in inputs when real objects are transforming in the physical world (Rolls, 1992a). The idea 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 is 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 associatively modifiable state, and the presynaptic afferents activated with the object in its new position would thus become strengthened on the still-activated postsynaptic neuron. The neuronal mechanisms that might implement this short-term temporal averaging in the modifiability are of interest, and include lasting effects of calcium entry as a result of the voltage-dependent activation of NMDA receptors, and continuing firing of the neuron implemented by recurrent collateral connections forming a short-term memory (see The Speed of Processing in the Temporal Cortical Visual Areas). The short temporal window (e.g., 0.5 s) of associative modifiability helps neurons to learn the statistics of objects transforming 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 translation invariance. I suggest that other invariances, for example size, spatial frequency, rotation, and view invariance, could be learned by similar mechanisms to those just described.

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; Rolls, 1992b; Logothetis et al., 1994; Ullman, 1996; Riesenhuber and Poggio, 1998). 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) (Perrett et al., 1987; Hasselmo et al., 1989b; Booth and Rolls, 1998).

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, VisNet, 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 et al., 1993; Wallis and Rolls, 1997; Rolls and Milward, 2000). The network can identify objects shown in a cluttered environment (for example, a natural scene) and can identify partially occluded objects (Stringer and Rolls, 2000). Parga and Rolls (1998) and Elliffe et al. (2000a) incorporated the associations between exemplars of the same object in the recurrent synapses of an autoassociative ("attractor") network, so that the techniques of statistical physics could be used to analyze the storage capacity of a system implementing invariant representations in this way. They showed that such networks did have an "object" phase in which the presentation of any exemplar (e.g., view) of an object would result in the same firing state as other exemplars of the same object, and that the number of different objects that could be stored is proportional to the number of synapses per neuron divided by the number of "views" of each object. Rolls and Milward (2000) explored the operation of the trace learning rule used in the VisNet architecture further, and showed that the rule operated especially well if the trace incorporated activity from previous presentations of the same object but no contribution from the current neuronal activity being produced by the current exemplar of the object. The explanation for this is that this temporally asymmetric rule (the presynaptic term from the current exemplar, and the trace from the preceding exemplars) encourages neurons to respond to the current exemplar in the same way as they did to previous exemplars. Elliffe et al. (2000b) examined the issue of spatial binding in this general class of hierarchical architecture and showed how, by forming high–spatial precision feature combination neurons early in processing, it is possible for later layers to maintain high precision for the relative spatial position of features within an object yet achieve invariance for the spatial position of the whole object.

These results with VisNet 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 (Stringer and Rolls, 2000).

Conclusions

Neurophysiological investigations of the inferior temporal cortex are revealing at least part of the way in which neuronal firing encodes information about faces and objects, and are showing that the representation implements several types of invariance. The representation found has clear utility for the receiving networks. These neurophysiological findings are stimulating the development of computational neuronal network models, which suggest that part of the cellular processing involves the operation of a modified associative learning rule with a short-term memory trace to help the system learn invariances from the statistical properties of the inputs it receives. It is a challenge to identify the cellular processes that could implement this short-term memory trace, and also the processes that might help to keep the total synaptic strength received by each neuron approximately constant, as is required for competitive networks (Rolls and Treves, 1998).

Acknowledgments

The author has worked on some of the investigations described here with P. Azzopardi, G. C. Baylis, M. Booth, P. Foldiak, M. E. Hasselmo, C. M. Leonard, T. J. Milward, D. I. Perrett, S. M. Stringer, M. J. Tovee, A. Treves, and G. Wallis, and their collaboration is sincerely acknowledged. Different parts of the research described were supported by the Medical Research Council, PG8513790 and PG9826105; by a Human Frontier Science Program grant; by an EC

Human Capital and Mobility grant; by the MRC Oxford Interdisciplinary Research Centre in Cognitive Neuroscience; and by the Oxford McDonnell-Pew Centre in Cognitive Neuroscience.

References

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

Baddeley, R.J., Abbott, L.F., Booth, M.J.A., Sengpiel, F., Freeman, T., Wakeman, E.A., and Rolls, E.T. (1997). Responses of neurons in primary and inferior temporal visual cortices to natural scenes. Proc. R. Soc. Lond. B Biol. Sci. *264,* 1775–1783.

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

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

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

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.

Bishop, C.M. (1995). Neural Networks for Pattern Recognition (Oxford: Clarendon Press).

Blair, R.J., Morris, J.S., Frith, C.D., Perrett, D.I., and Dolan, R.J. (1999). Dissociable neural responses to facial expressions of sadness and anger. Brain *122*, 883–893.

Booth, M.C.A., and Rolls, E.T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. Cereb. Cortex *8*, 510–523.

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

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. Neurophysiol. *46*, 369–384.

Desimone, R. (1991). Face-selective cells in the temporal cortex of monkeys. J. Cogn. 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.

Dolan, R.J., Fink, G.R., Rolls, E.T., Booth, M., Holmes, A., Frackowiak, R.S.J., and Friston, K.J. (1997). How the brain learns to see objects and faces in an impoverished context. Nature *389*, 596–599.

Eckhorn, R., and Popel, B. (1974). Rigorous and extended application of information theory to the afferent visual system of the cat. Biol. Cybern. *16*, 191–200.

Elliffe, M.C.M., Rolls, E.T., Parga, N., and Renart, A. (2000a). A recurrent model of transformation invariance by association. Neural Networks *13*, 225–237.

Elliffe, M.C.M., Rolls, E.T., and Stringer, S.M. (2000b). Invariant recognition of feature combinations in the visual system. Biol. Cybern., in press.

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

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

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

Fukushima, K. (1989). Analysis of the process of visual pattern recognition by the neocognitron. Neural Networks *2*, 413–420.

Fukushima, K. (1991). Neural networks for visual pattern recognition. IEEE Trans. E *74*, 179–190.

Gallant, J.L., Connor, C.E., and Van-Essen, D.C. (1998). Neural activity in areas V1, V2 and V4 during free viewing of natural scenes compared to controlled viewing. Neuroreport *9*, 85–90.

Gawne, T.J., and Richmond, B.J. (1993). How independent are the messages carried by adjacent inferior temporal cortical neurons? J. Neurosci. *13*, 2758–2771.

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. (1989a). The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. Behav. Brain Res. *32*, 203–218.

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

Hoffman, E.A., and Haxby, J.V. (2000). Distinct representations of eye gaze and identity in the distributed neural system for face perception. Nat. Neurosci. *3*, 80–84.

Hopfield, J.J. (1982). Neurons with graded responses have collective properties like those of two-state neurons. Proc. Natl. Acad. Sci. USA *81*, 3088–3092.

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

Ishai, A., Ungerleider, L.G., Martin, A., Schouten, J.L., and Haxby, J.V. (1999). Distributed representation of objects in the human ventral visual pathway. Proc. Natl. Acad. Sci. USA *96*, 9379–9384.

Kanwisher, N., McDermott, J., and Chun, M.M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. *17*, 4302–4311.

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

Kohonen, T. (1977). Associative Memory: A System Theoretical Approach (New York: Springer).

Kohonen, T. (1989). Self-Organization and Associative Memory, Third Edition (Berlin: Springer-Verlag).

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., and Sheinberg, D.L. (1996). Visual object recognition. Annu. Rev. Neurosci. *19*, 577–621.

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

MacKay, D.M., and McCullough, W.S. (1952). The limiting information capacity of a neuronal link. Bull. Math. Biophys. *14*, 127–135.

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

Mikami, A., Nakamura, K., and Kubota, K. (1994). Neuronal responses to photographs in the superior temporal sulcus of the rhesus monkey. Behav. Brain Res. *60*, 1–13.

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. Annu. Rev. Neurosci. *16*, 245–263.

Morris, J.S., Fritch, C.D., Perrett, D.I., Rowland, D., Young, A.W., Calder, A.J., and Dolan, R.J. (1996). A differential neural response in the human amygdala to fearful and happy face expressions. Nature *383*, 812–815.

Olshausen, B.A., and Field, D.J. (1997). Sparse coding with an overcomplete basis set: a strategy employed by V1? Vision Res. *37*, 3311–3325.

O'Scalaidhe, S.P., Wilson, F.A.W., and Goldman-Rakic, P.S. (1999). Face-selective neurons during passive viewing and working memory performance of rhesus monkeys: evidence for intrinsic specialization of neuronal coding. Cereb. Cortex *9*, 459–475.

Panzeri, S., Biella, G., Rolls, E.T., Skaggs, W.E., and Treves, A. (1996). Speed, noise, information and the graded nature of neuronal responses. Network *7*, 365–370.

Panzeri, S., Treves, A., Schultz, S., and Rolls, E.T. (1999a). On decoding the responses of a population of neurons from short time epochs. Neural Comput. *11*, 1553–1577.

Panzeri, S., Schultz, S.R., Treves, A., and Rolls, E.T. (1999b). Correlations and the encoding of information in the nervous system. Proc. R. Soc. B Biol. Sci. *266*, 1001–1012.

Panzeri, S., Rolls, E.T., Battaglia, F., and Lavis, R. (2000). Speed of information retrieval in multilayer networks of integrate-and-fire neurons. Network, in press.

Parga, N., and Rolls, E.T. (1998). Transform invariant recognition by association in a recurrent network. Neural Comput. *10*, 1507–1525.

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. R. Soc. Lond. B. Biol. Sci. *223*, 293–317.

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

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

Renart, A., Parga, N., and Rolls, E.T. (2000). A recurrent model of the interaction between the prefrontal cortex and inferior temporal cortex in delay memory tasks. Adv. Neural Inform. Proc. Systems *12*, in press.

Riesenhuber, M., and Poggio, T. (1998). Just one view: invariances in inferotemporal cell tuning. Adv. Neural Inform. Proc. Systems *10*, 215–221.

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. (1989). 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, eds. (San Diego: Academic Press), pp. 240–265.

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

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

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

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

Rolls, E.T. (1996). A theory of hippocampal function in memory. Hippocampus *6*, 601–620.

Rolls, E.T. (1999). The Brain and Emotion (Oxford: Oxford University Press).

Rolls, E.T. (2000). Neurophysiology and functions of the primate amygdala, and the neural basis of emotion. In The Amygdala: A Functional Analysis. J.P. Aggleton, ed. (Oxford: Oxford University Press), pp. 447–478.

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., and Milward, T. (2000). A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures. Neural Comput. *11*, in press.

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

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

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

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., and Treves, A. (1998). Neural Networks and Brain Function (Oxford: Oxford University Press).

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

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

Rolls, E.T., Hornak, J., Wade, D., and McGrath, J. (1994b). Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. J. Neurol. Neurosurg. Psychiatry *57*, 1518–1524.

Rolls, E.T., Critchley, H.D., and Treves, A. (1996). The representation of olfactory information in the primate orbitofrontal cortex. J. Neurophysiol. *75*, 1982–1996.

Rolls, E.T., Treves, A., and Tovee, M.J. (1997). The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. Exp. Brain Res. *114*, 149–162.

Rolls, E.T., Treves, A., Robertson, R.G., Georges-François, P., and Panzeri, S. (1998). Information about spatial view in an ensemble of primate hippocampal cells. J. Neurophysiol. *79*, 1797–1813.

Rolls, E.T., Tovee, M.J., and Panzeri, S. (1999). The neurophysiology of backward visual masking: information analysis. J. Cogn. Neurosci. *11*, 335–346.

Sato, T. (1989). Interactions of visual stimuli in the receptive fields of inferior temporal neurons in macaque. Exp. Brain Res. *77*, 23–30.

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.

Shannon, C.E. (1948). A mathematical theory of communication. ATT Bell Labs. Tech. J. *27*, 379–428.

Stringer, S.M., and Rolls, E.T. (2000). Position invariant recognition

in the visual system with cluttered environments. Neural Networks *13*, 305–315.

Tanaka, K. (1993). Neuronal mechanisms of object recognition. Science *262*, 685–688.

Tanaka, K. (1996). Inferotemporal cortex and object vision. Annu. Rev. Neurosci. *19*, 109–139.

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, eds. (New York: Elsevier), pp. 101–109.

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

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

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

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

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

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

Treves, A., and Rolls, E.T. (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 *4*, 374–391.

Treves, A., Rolls, E.T., and Simmen, M. (1997). Time for retrieval in recurrent associative memories. Physica D *107*, 392–400.

Treves, A., Panzeri, S., Rolls, E.T., Booth, M., and Wakeman, E.A. (1999). Firing rate distributions and efficiency of information transmission of inferior temporal cortex neurons to natural visual stimuli. Neural Comput. *11*, 611–641.

Ullman, S. (1996). High-Level Vision: Object Recognition and Visual Cognition (Cambridge, MA: Bradford/MIT Press).

Wallis, G., and Rolls, E.T. (1997). Invariant face and object recognition in the visual system. Prog. Neurobiol. *51*, 167–194.

Wallis, G., Rolls, E.T., and Foldiak, P. (1993). Learning invariant responses to the natural transformations of objects. Int. Joint Conf. Neural Networks *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.

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

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

Xiang, J.Z., and Brown, M.W. (1998). Differential neuronal encoding of novelty, familiarity and recency in regions of the anterior temporal lobe. Neuropharmacology *37*, 657–676.

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.