Exp Brain Res (1997) 114:149–162

RESEARCH ARTICLE

Edmund T. Rolls · Alessandro Treves
Martin J. Tovee

The representational capacity of the distributed encoding of information provided by populations of neurons in primate temporal visual cortex

Received: 12 April 1996 / Accepted: 19 August 1996

Abstract It has been shown that it is possible to read, from the firing rates of just a small population of neurons, the code that is used in the macaque temporal lobe visual cortex to distinguish between different faces being looked at. To analyse the information provided by populations of single neurons in the primate temporal cortical visual areas, the responses of a population of 14 neurons to 20 visual stimuli were analysed in a macaque performing a visual fixation task. The population of neurons analysed responded primarily to faces, and the stimuli utilised were all human and monkey faces. Each neuron had its own response profile to the different members of the stimulus set. The mean response of each neuron to each stimulus in the set was calculated from a fraction of the ten trials of data available for every stimulus. From the remaining data, it was possible to calculate, for any population response vector, the relative likelihoods that it had been elicited by each of the stimuli in the set. By comparison with the stimuli actually shown, the mean percentage correct identification was computed and also the mean information about the stimuli, in bits, that the population of neurons carried on a single trial. When the decoding algorithm used for this calculation approximated an optimal, Bayesian estimate of the relative likelihoods, the percentage correct increased from 14% correct (chance was 5% correct) with one neuron to 67% with 14 neurons. The information conveyed by the population of neurons increased approximately linearly from 0.33 bits with one neuron to 2.77 bits with 14 neurons. This leads to the important conclusion that the number of stimuli that can be encoded by a population of neurons in this part of the visual system increases approximately exponentially as the number of cells in the sample increases (in that the log of the number of stimuli increases almost linearly). This is in contrast to a local encoding scheme (of "grandmother" cells), in which the number of stimuli encoded increases linearly with the number of cells in the sample. Thus one of the potentially important properties of distributed representations, an exponential increase in the number of stimuli that can be represented, has been demonstrated in the brain with this population of neurons. When the algorithm used for estimating stimulus likelihood was as simple as could be easily implemented by neurons receiving the population's output (based on just the dot product between the population response vector and each mean response vector), it was still found that the 14-neuron population produced 66% correct guesses and conveyed 2.30 bits of information, or 83% of the information that could be extracted with the nearly optimal procedure. It was also shown that, although there was some redundancy in the representation (with each neuron contributing to the information carried by the whole population 60% of the information it carried alone, rather than 100%), this is due to the fact that the number of stimuli in the set was limited (it was 20). The data are consistent with minimal redundancy for sufficiently large and diverse sets of stimuli. The implication for brain connectivity of the distributed encoding scheme, which was demonstrated here in the case of faces, is that a neuron can receive a great deal of information about what is encoded by a large population of neurons if it is able to receive its inputs from a random subset of these neurons, even of limited numbers (e.g. hundreds).

E. T. Rolls (✉) · A. Treves[1] · M. J. Tovee[2]
University of Oxford, Department of Experimental Psychology, South Parks Road, Oxford, OX1 3UD, UK;
Fax: +44-1865-310447, e-mail: edmund.rolls@psy.ox.ac.uk

Present addresses:
[1] S.I.S.S.A. – Cognitive Neuroscience, via Beirut 2–4, I-34103 Trieste, Italy
[2] Department of Psychology, University of Newcastle, Ridley Building, Newcastle upon Tyne, NE1 7RU, UK

Introduction

The visual pathways project by a number of cortico-cortical stages from the primary visual cortex until they reach the temporal lobe visual cortical areas (Seltzer and Pandya 1978; Maunsell and Newsome 1987; Baizer et al. 1991; Rolls 1991, 1992a). Neurons with different types

of sensitivity to visual stimuli tend to be found in different parts of these temporal cortical areas (Baylis et al. 1987). In some areas neurons respond to stimulus properties such as shape, orientation, texture and colour (Baylis et al. 1987; Tanaka et al. 1991), and in other areas, especially areas in the cortex in the superior temporal sulcus, up to 20% of the neurons with visual responses have selectivity for faces (Desimone and Gross 1979; Bruce et al. 1981; Rolls 1981a, b, 1984, 1992a, b; Perrett et al. 1982; Desimone et al. 1984; Gross et al. 1985; Desimone 1991). Some of the temporal cortical areas provide a representation of objects and faces that is relatively invariant with respect to retinal position, size, rotation and even view (Rolls 1994, 1995), and such invariant representations form appropriate inputs to associative neuronal networks in structures to which the temporal cortical areas project such as the hippocampus and amygdala (see, e.g. Rolls 1992a–c; Treves and Rolls 1994). Consistent with this, lesions of the inferior temporal visual cortex impair the ability of monkeys to respond to objects irrespective of changes in size, lighting and viewing angle (Weiskrantz and Saunders 1984).

A fundamental issue then arises of how the information about objects and faces is represented by the activity of temporal cortical neurons. Important questions are: how selective and "information-bearing" (Suga 1989) the neurons are for different classes of stimulus such as face compared with non-face; how selective or information-bearing the neurons are for individual items within a class; whether the neurons use "local" or "grandmother" cell encoding, with strong or even great selectivity of a single neuron for a particular object in the environment (Barlow 1972), or fully distributed representations in which all the neurons participate (Hinton et al. 1986; Churchland and Sejnowski 1992), or sparse representations in which the distributed encoding is not fully distributed (Rolls and Treves 1990; Treves and Rolls 1991). In a series of previous investigations, we have shown that single neurons in the temporal lobe visual cortex tuned to faces do not respond to only one face in a set of faces, but instead typically respond to several members of the set, with each cell having its own characteristic firing rate response profile to the different members of the set (see Rolls 1984, 1992a; Baylis et al. 1985; Rolls and Tovee 1995). The representation provided by these faces may be described as sparsely distributed, and not as local (Rolls and Tovee 1995).

Distributed representations, in which many of the neurons that participate in the representation of each stimulus or event (see, e.g. Hinton et al. 1986) have a number of advantages over local or grandmother cell encoding, for which there is strong or even great selectivity of the neuron for a particular environmental stimulus (Barlow 1972). The advantages of distributed representations include generalisation as the nature of the input changes and graceful degradation or fault tolerance if the network in which the representation is present is incomplete or damaged. If the distributed representation is not fully distributed (with, e.g. half the neurons active for any one stimulus), but is a sparse distributed representation, then this allows large numbers of representations to be stored and retrieved in associative neural networks (Rolls and Treves 1990; Treves and Rolls 1991). Another potential advantage of distributed representations is that large numbers of different stimuli or events can be encoded. Consider the number of stimuli that can be encoded by a population of C neurons without noise. If local encoding is used and the representation is binary (e.g. the neuron is either active or not), then C different representations can be encoded (one different neuron is "on" for each stimulus). If (fully) distributed encoding is used, then 2^C different representations can be encoded (2^C is the number of different combinations of C binary variables). The fundamental question addressed in this paper is the extent to which the brain can utilise the potential advantage of distributed representations to encode a very large (exponentially large) number of different stimuli in a population of neurons. The potential advantage will only be usefully realised to the extent that: each member of the population of neurons has different responses to each stimulus in a set of stimuli (with, e.g. different combinations of neurons firing to each stimulus); and the responses of a neuron on a given trial are not too noisy, that is, the standard deviation of the responses of a neuron to the same stimulus on different trials must not be too great, and the responses to different stimuli must be reliably different to each other. Evidence on this issue can thus only be obtained by examining the response properties of real neurons in the brain, and this is what is described in this paper. We analysed the responses of face-selective neurons to 20 different faces, obtaining at least ten trials of data to each stimulus (presented in random order). We were able to repeat this experiment for 14 different face-selective neurons and then analyse the information about which of the 20 stimuli had been presented.

The crucial feature of distributed representations examined here is that they have the potential, if different representations are provided by different cells, for a very large representational capacity over a cell population. This large-capacity situation is attained when the information coded by a population of cells increases linearly, or close to linearly, with the size of the population, in which case the number of stimuli coded grows exponentially with population size. This is in contrast to local representations, in which each stimulus is allocated one or a set of neurons to represent it; and thus the number of stimuli coded grows only linearly with the size of the population or, in terms of information, the information conveyed by the response of a population grows, on average, only logarithmically with population size. We emphasize that the potential advantage of distributed representations is realised only if different neurons code for different things: if the representations provided by several cells in a population were strongly correlated (i.e. largely the same), there would be no strong increase in representational capacity with population size, no matter how distributed the representations. Large populations would just provide more redundancy.

It is therefore very important to extend previous quantitative analyses based on the responses of single cells (or pairs of cells; Gawne and Richmond 1993) and to address directly the question of how much information is conveyed by the responses of populations of cells. This is the goal of the present analysis, which considers, at the population level, the responses of 14 face-selective temporal cortical cells that had been previously analysed at the single-cell level (Rolls et al. 1995). In carrying out the information-theoretic analysis, great care was devoted to extracting information measures, to monitoring the values obtained as they vary as a function of cell population size, and to taking ceiling effects into proper account. It is to these three factors that we ascribe the difference between our results and those of a previous study in inferior temporal cortex with similar goals (Gochin et al. 1994). We note from the outset two potential limitations of the data used here, to be discussed later. First, the cells were not recorded simultaneously, which prevented our analysis from detecting potential effects stemming from correlations between neurons in their trial-to-trial variability (see Gawne and Richmond 1993; Gochin et al. 1994). Second, the number of trials per stimulus available for each cell was low (ten), which again made it vital to use novel techniques, developed in order to allow correction for limited sampling, when extracting accurate information measures (see Optican et al. 1991; Treves and Panzeri 1995; Panzeri and Treves 1996).

This investigation is one of a series (Rolls 1992a, 1994, 1995; Rolls et al. 1994; Hornak et al. 1996) designed to investigate the normal functions of the temporal lobe visual cortical areas and how damage to these brain regions may underlie the perceptual and related deficits found in patients with disruption of function of these and connected regions. The neurons described here with responses that occur mainly in faces, but that within that class convey information about which face has been seen (see Tovee et al. 1993, 1994; Rolls and Tovee 1995; Tovee and Rolls 1995), form a useful population of neurons for this kind of investigation, for neurons of this type can frequently be found in the temporal cortical areas, so that sufficient data can be obtained in repeated tracks for analyses such as those described here.

Materials and methods

Neurophysiology

The responses of single neurons in the temporal cortical visual areas were measured to a set of 68 visual stimuli in macaques performing a visual fixation task. The stimuli included 20 monkey and human faces (S=20). The neurons were selected to meet the previously used criteria of face selectivity by responding more than twice as much to the optimal face as to the optimal non-face stimulus in the set (Rolls 1984, 1992a-c). The responses of each neuron to the same set of 20 faces provided the set of neuronal responses for the analyses described here. Ten trials for each stimulus were available. The set of stimuli were shown once in random order, then a second time in a new random sequence, etc. The neurons were not recorded simultaneously, but were recorded from the same brain region. The neurophysiological protocol was designed to provide data for the investigations described here and for measurement of the sparseness of the representation (see Rolls and Tovee 1995, where the neurophysiological methods are described in more detail). The recordings were made in two rhesus macaques (*Macaca mulatta*), but the 14 neurons described in this paper were all recorded from the first macaque, partly because we wished to ensure that whatever information was shown to be encoded by the set of neurons included in the study was present in an individual animal. The sites at which the neurons were recorded are shown by Rolls and Tovee (1995), and the majority were in the cortex in the anterior part of the superior temporal sulcus.

Data analysis and decoding algorithms

Response quantification

From the response of each neuron c to each stimulus in the set, we extracted a single mean firing rate (r_c, in spikes per second), calculated from the number of spikes recorded between 100 and 600 ms after the presentation of the stimulus. Because most of the information about which stimulus is shown is made evident by measuring the firing rate of the neuron, and temporal encoding adds relatively little additional information for this population of neurons (Tovee et al. 1993; Tovee and Rolls 1995), the analyses described here were based on the information available from the firing rate, and the period in which this was measured was the post-stimulus period 100–600 ms with respect to the onset of the visual stimulus, as most of the information about which stimulus was seen is available in this period (Tovee et al. 1994; Tovee and Rolls 1995). For comparison, we repeated all the analyses for the much shorter analysis period of 100–150 ms post-stimulus.

Cross-validation

In general, the analyses we then performed involved constructing pseudosimultaneous population response vectors (**r**), occurring in what were labelled as "test" trials (**r** is a vector with one element, or component, for each of the C cells considered). Each response vector was compared with the mean population response vector to each stimulus, as derived from a different set of "training" data, in order to estimate, by means of one of several decoding algorithms described below, the relative probabilities [$P(s'|\mathbf{r})$] that the response **r** had been elicited by any one stimulus s' in the set. Summing over different test trial responses to the same stimulus s, we could extract the probability that by presenting stimulus s the neuronal response would be interpreted as having been elicited by stimulus s', and from that the resulting measures of percentage correct identification and of the information decoded from the responses. Separating the test from the training data is called cross-validation, the details of which follow.

In part of the analyses the conventional cross-validation procedure was used of allocating a proportion (1-x) of the ten trials available for each cell for each stimulus as training data, to compute the mean response by that cell to that stimulus. Then 10x test trial population responses to each stimulus were constructed by randomly selecting, cell by cell, one from the remaining number of trials. No trial was used twice. In this procedure, each trial was used either for training or for testing. Different values for x were tried, but the most reliable results were obtained by using a different procedure, which allows effective use of all available data both for training and as test trials. In this second procedure, only one of the ten trials was used for testing, the remaining nine for training, allowing better decoding, as shown under Results. The resulting probability that s is decoded as s' is, however, averaged over all choices of test trials, thus alleviating finite sampling problems more effectively than with the first procedure. Finally, we also compared the results with those obtained in the absence of cross-validation, i.e. when all trials were used both as test and as training trials.

Algorithms for likelihood estimation

Several different decoding algorithms were used for estimating the likelihood of each stimulus from the recorded response. In the final analysis reported here, two are selected. The first algorithm, the "probability estimator" (PE), tries to reconstruct the correct Bayesian probabilities from the data, extracting from the data itself as much information as is possible by any decoding procedure. The second algorithm, based on a simple "dot product" (DP), tries to emulate the processing that could be performed by neurons receiving the output of the neuronal population recorded, thus extracting that portion of the information theoretically available that could be extracted with simple neurophysiologically plausible operations by receiving neurons.

The PE algorithm extracts $P(s'|\mathbf{r})$ from an estimate of the probability $P(\mathbf{r},s')$ of a stimulus-response pair, by normalizing so that $\Sigma_{s'} P(s'|\mathbf{r})=1$ (see Foldiak 1993). The probability $P(\mathbf{r},s')$ is estimated for this purpose as $P(s')\Pi_c P(r_c|s')$, where r_c is the firing rate of cell c. Finally, $P(r_c|s')$ is derived from the responses of cell c in the training trials. Those are fitted with a Gaussian distribution, whose amplitude at r_c gives $P(r_c|s')$, except when $r_c=0$, in which case $P(r_c|s')$ is the best estimate of the fraction of training trials yielding zero firing.

The DP algorithm computes the normalized DPs between the current firing vector $\mathbf{r}$ on a test trial and each of the mean firing rate response vectors in the training trials for each stimulus s'. (The normalized DP is the dot or inner product of two vectors divided by the product of the length of each vector. The length of each vector is the square root of the sum of the squares.) Thus, what is computed are the cosines of the angles of the test vector of cell rates with, in turn for each stimulus, the mean response vector to that stimulus. The highest DP indicates the most likely stimulus that was presented, and this is taken as the best guess for the percentage correct measures. For the information measures, it is desirable to have a graded set of probabilities for which of the different stimuli was shown, and these were obtained from the DPs as follows. The S DP values were cut at a threshold equal to their own mean plus 1 SD, and the remaining non-zero ones were normalized to sum to 1. It is clear that in this case each operation could be performed by an elementary neuronal circuit (the DP by a weighted sum of excitatory inputs, the thresholding by activity-dependent inhibitory subtraction, and the normalization by divisive inhibition). The resulting relative probabilities are cruder estimates than those obtained with the PE algorithm, and a precise quantitative assessment of the price paid for using a simpler and neurophysiologically plausible algorithm can be derived from a comparison of the amounts of information extracted in both cases.

Note that no attempt was made to optimize the DP algorithm. The PE algorithm had a rather fixed structure, too, in which the only "free" choice was that of a convenient distribution with which to fit $P(r_c|s')$. The truncated Gaussian was then chosen over a Poisson distribution (with an additional weight at $r_c=0$), because it produced higher values for both percentage correct and information (this does not necessarily hold for other cell populations; our unpublished observations). In contrast, the neural network decoding procedure developed by Hertz et al. (1992) is optimized extensively, albeit only across the parameters describing a fixed class of neural network decoders.

Procedures for extracting information measures

Probability and frequency tables

Having estimated the relative probabilities that the test trial response had been elicited by any one stimulus, the stimulus that turned out to be most likely, i.e. that which had the highest (estimated) probability, was defined to be the predicted stimulus, s^P. The fraction of times that the predicted stimulus s^P was the same as the actual stimulus s is directly a measure of the percentage correct for a given data set. In parallel, the estimated relative probabilities (normalized to 1) were averaged over all test trials for all stimuli, to generate a table $P^R_N(s,s')$ describing the relative probability of each pair of actual stimulus s and posited stimulus s'. We also generated a second (frequency) table $P^F_N(s,s^P)$ from the fraction of times an actual stimulus s elicited a response that led to a predicted (most likely) stimulus s^P. The difference between the table P^R_N and the table P^F_N can be appreciated by noting that each vector comprising a pseudosimultaneous trial contributes to P^R_N a set of numbers (one for each possible s') whose sum is 1, while to P^F_N it contributes a single 1 for s^P and zeroes for all other stimuli. Obviously each contribution was normalized by dividing, in both cases, by the total number N of (test) trials available ($N=10\times x\times 20$ for the conventional cross-validation procedure, and $N=10\times 20$ for the more efficient procedure with one test trial and the remaining trials used for training).

Information measures

From any probability table $P(s,\mathbf{r})$ embodying a relationship between the variable s (here, the stimulus) and $\mathbf{r}$ (here, the response rate vector), one can extract the mutual information

$$I(s,r) = \Sigma_s \Sigma_r P(s,r) \log_2 \frac{P(s,r)}{P(s)P(r)}$$

When the probability table has to be estimated as the frequency table of a limited data sample, however, it becomes crucial to evaluate the effects of limited sampling on the information estimate. When $\mathbf{r}$ is a multidimensional quantity (a vector, $\mathbf{r}$), as it necessarily is if it represents the firing rate of several cells, the minimum number of trials required to sample sufficiently the response space grows exponentially with the dimensionality of that space, i.e. the number of cells considered (Treves and Panzeri 1995). This rules out, in our case, any attempt to evaluate directly the quantity $I(s,\mathbf{r})$. A standard procedure is then to derive from the original frequency table of stimuli and responses an auxiliary table, of stimuli and additional variables, spanning a limited set, which are derived from the responses by any arbitrary algorithm. These additional variables can be chosen, in particular, to coincide with the stimuli themselves, which comprise the minimum set with the potential still for full correlation, or maximal information. In general, though, the information content of the auxiliary table will be less than that of the original table, by an amount that depends on the severity of the manipulation performed. Two types of auxiliary tables were derived here, called P^R_N and P^F_N.

In deriving P^F_N, each response is used to predict its stimulus. While s^P spans only S values compared with the very large number of possible (multidimensional) rate responses, the auxiliary table is otherwise unregularized, in that each trial of a limited total number produces a relatively large "bump" in $P^F_N(s,s^P)$. The result of this is that a raw estimate of $I(s,s^P)$ [which can be denoted as $I_N(s,s^P)$ to point out that it is obtained from a total of N trials] can still be very inaccurate, in particular, overestimated. Sophisticated methods have been devised (Panzeri and Treves 1996) to correct raw information estimates for limited sampling, by subtracting out the mean of the error. These methods are safely applicable when the subtracted term $[I_N(s,s^P)]-I(s,s^P)$ is smaller than approximately 1 bit. With the present data (and only a handful of trials per stimulus) the subtracted term turns out to be large when few cells are considered and to become sufficiently small only when more than about ten cells are included (the reason for this is just that more cells produce more accurate predictions and therefore more concentrated tables). The conclusion is that the (corrected) estimate of $I(s,s^P)$ is reliable only when most of the 14 cells in the total population are considered, which makes it impossible to discuss effectively how $I(s,s^P)$ depends on C, the number of cells.

P^R_N, on the other hand, can be conceived of as being more *regularized* than P^F_N, because each trial contributes not a relatively large bump to just one bin (s^P), but smaller additions to several bins (s'). The consequence is that the distortion in the information estimate due to limited sampling (small N) is smaller, and the subtraction of a suitable correction term $[I_N(s,s')]-I(s,s')$ is enough to produce accurate corrected estimates of information. The correc-

tion term to be used differs from that appropriate to correct $I(s,s^{\mathrm{P}})$; it takes the form:

$$\langle I_N(s,s')\rangle - I(s,s') \approx \frac{1}{2N\log_{(2)}}\Sigma_s P(s)\Sigma_{s'}\left[\frac{Q^{R}_{N}(s,s')}{P^{R}_{N}(s,s')} - \frac{P^{R}_{N}(s,s')}{P(s)}\right] - \frac{1}{2N\log_{(2)}}\Sigma_{s'}\left[\frac{Q^{R}_{N}(s')}{P^{R}_{N}(s')} - P^{R}_{N}(s')\right]$$

where $Q^{\mathrm{R}}{}_{N}(s,s')$ is the table obtained analogously to $P^{\mathrm{R}}{}_{N}(s,s')$, but averaging over all test trials $P^{2}(s'|\mathrm{r})$ instead of $P(s'|\mathrm{r})$, and where care has to be taken in performing the sums over s', to avoid including stimuli posited to have zero probability. For a derivation of this and other correction terms and for a fuller discussion of the difference between various information estimates, we refer to Panzeri and Treves (1996), where the advantages of this correction procedure over the earlier regularization procedure of Kjaer et al. (1994) are also described (see also the explicit comparison in Golomb et al. 1996). Here it is sufficient to note that $I(s,s')$ (as best estimated with the present correction procedure) will in any case tend to a "true" value that, being based on a regularized probability distribution, is less than the value (unmeasurable except with few cells) attained by $I(s,\mathrm{r})$. The same applies to $I(s,s^{\mathrm{P}})$. $I(s,s')$ is the quantity that can be measured more accurately for any number of cells with the present data, and comparisons across data sets should only be performed using the same quantities.

Averaging

To generate the results quoted in the paper, i.e. percentage correct and information as a function of number of cells in the population, means were taken over ten different partitions between test and training trials for any set of cells, and over a large number of different sets of C cells randomly selected from the total population of 14 cells.

Results

Comparison between measures obtained with different procedures

The values for the mean information, $I(s,s')$, available in the responses of different numbers of these neurons in each trial, about which of a set of 20 face stimuli have been shown, are displayed in Fig. 1a. The PE algorithm was used for estimating the relative probability of posited stimuli s', and different cross-validation procedures were used. The same data produced the percentage correct predictions reported in Fig. 1b. It can be seen that, whatever the procedure, both the information and the percentage correct rise initially linearly with population size from their baseline level (which is zero for the information and $1/S$=0.05 for the percentage correct) and then tend to slow down as the population gets close to including all 14 cells. This essentially linear rise in information as the number of cells in the sample is increased is the first major result described in this paper. In addition, both graphs show a small dependence on the cross-validation procedure adopted, the details of which are considered in the next paragraph.

In the absence of cross-validation, the percentage correct and information rise to higher levels, which is just a spurious effect due to the use of the same data for both training and testing, or in other words to trials being compared with themselves in the extraction of relative probabilities ("overfitting"). The results with the cross-validation procedure using x=1 test trial and the remaining (9) trials as training trials, repeated in turn using a different test trial from the dataset each time and averaging the results (crosses in Fig. 1a), were more efficient than the conventional cross-validation procedure (in Fig. 1a: diamonds for x=3/10, and triangles for x=5/10) in that it resulted in a higher percentage correct. This is expected, because the mean percentage correct depends on the quality of the decoding, which is better if based on 9 training trials (for the efficient procedure) than if based on, respectively, $10\times(1-x)$=7 or 5 training trials. The information obtained with the one test trial proce-

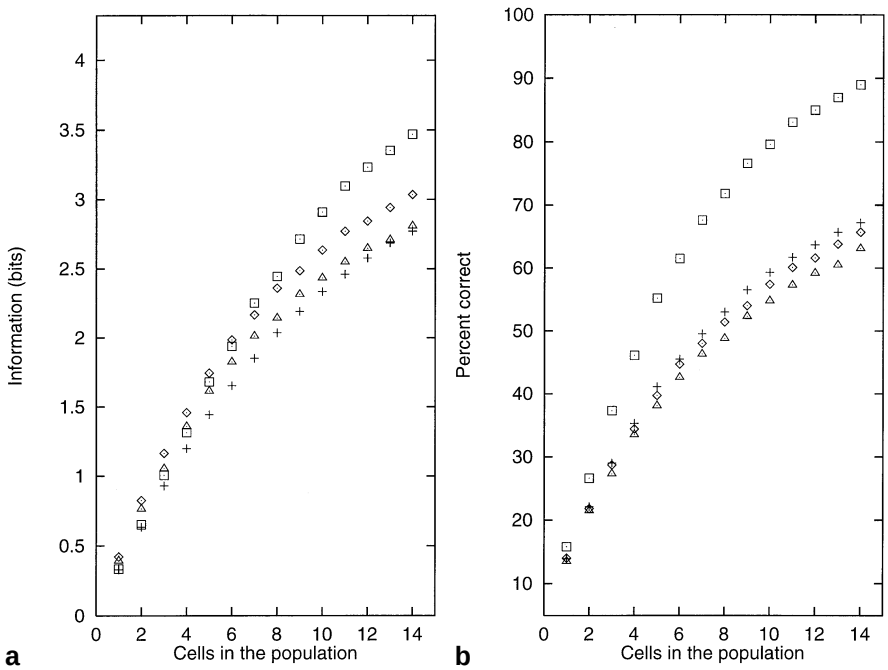

Fig. 1a The values for the mean information available in the responses of different numbers of these neurons in each trial, about which of a set of 20 face stimuli have been shown. The decoding method was probability estimation, and the effects were obtained with cross-validation procedures utilising 30% of the trials as test trials (*diamonds*) and 50% (*triangles*), with the remainder of the trials in the cross-validation procedure used as training trials, are compared with those obtained with the more efficient procedure explained in the text (+), and with those obtained without cross-validation (*squares*). **b** The percentage correct for the corresponding data to those shown in **a**

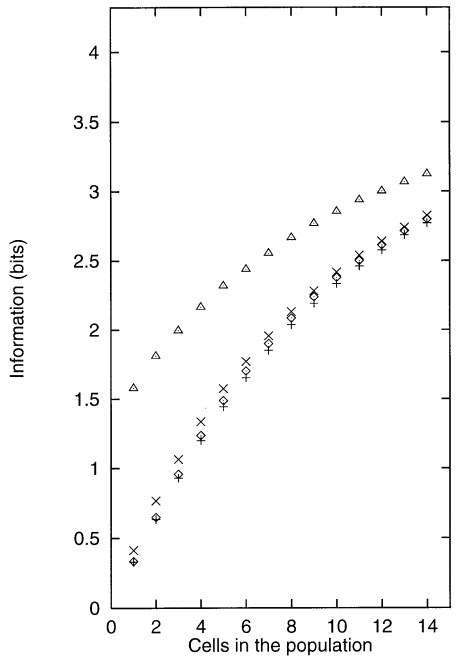

Fig. 2 For the efficient cross-validation procedure and probability estimation decoding, $I(s,s^P)$ after correction (×) and its corresponding raw, uncorrected, measure (*triangles*) is contrasted with $I(s,s')$ after correction (+) and its corresponding raw, uncorrected, measure (*diamonds*)

dure is a little lower than that from the conventional cross-validation procedure, but this is just due to undercorrection with effectively three or five test trials with the conventional procedure, so that this latter procedure actually is a slight overestimate of the information. The undercorrection is due to the fact that with very limited numbers of trials, such as three or five, the correction procedure gives a result which is biased upwards a little (see Panzeri and Treves 1996 for full analysis). The undercorrection applies particularly to the conventional procedure, for which there are three or five test trials and does not occur when there are as many as ten test trials, as there effectively are with the efficient procedure (one test trial repeated for ten different test trial choices). Conventional cross-validation is thus suboptimal with the relatively few trials (ten) available, because, with this number of trials, separating out a fixed subset to be used for training leaves so few trials for the testing set that the information estimates are slightly overestimated. The efficient cross-validation procedure in which one trial at a time is used as a test trial and the remaining (nine) trials are used for the training set appears, on the other hand, to afford both the best decoding and the least limited sampling bias and is therefore the one used from now on.

Measures of the information $I(s,s^P)$ were also taken, as mentioned under Materials and methods, in order to compare the effect of using the tables $P^R{}_N$ and $P^F{}_N$. Figure 2 shows, for the efficient cross-validation procedure, both $I(s,s^P)$ and $I(s,s')$ and their corresponding raw measures, before the correction term has been subtracted out. While the corrected measures do not differ by much, the raw measures for the information $I(s,s^P)$ based on the frequency table $P^F{}_N$ are very high, resulting in a correction term so large (1.16 bits for single cells) as to be really at the border of the region where subtracting it is enough to remove finite sampling biases (Panzeri and Treves 1996). Therefore $I(s,s^P)$ is somewhat less reliable a measure than $I(s,s')$, at least for small populations, which is also where they differ proportionally the most.

Having analysed the small dependence of the results on the cross-validation procedure and on the exact information quantity being measured, we focus for simplicity

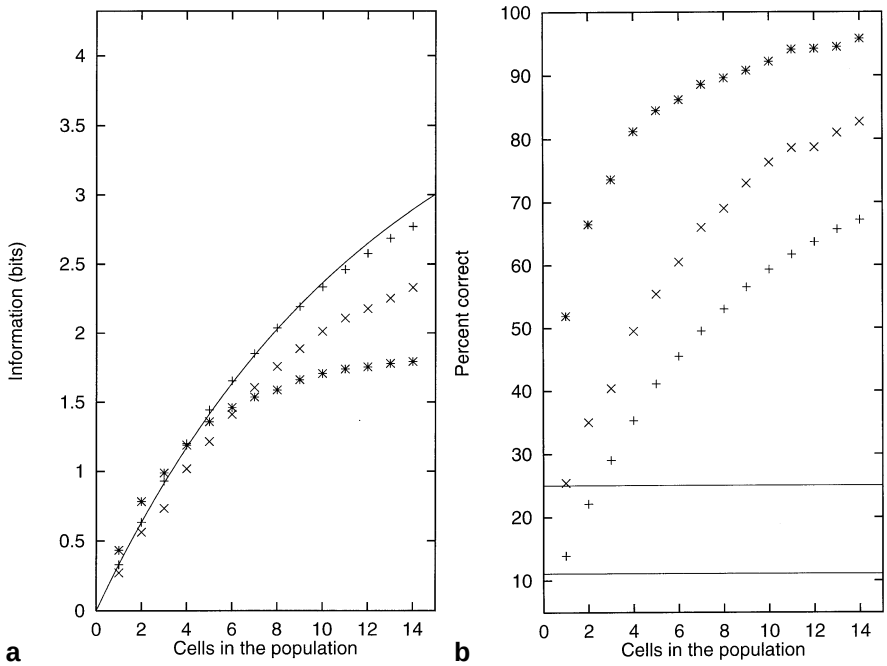

Fig. 3a The information values obtained for the full set of 20 stimuli (+) are compared with those obtained for two examples of reduced stimulus sets, comprising nine (×) and four (*stars*) stimuli. The *continuous line* is a fit to the information *C* cells provide about 20 stimuli, calculated as $(1-\Phi^C)\log_2(20)$, as explained in the text. **b** The percentage correct for the corresponding data to those shown in **a**

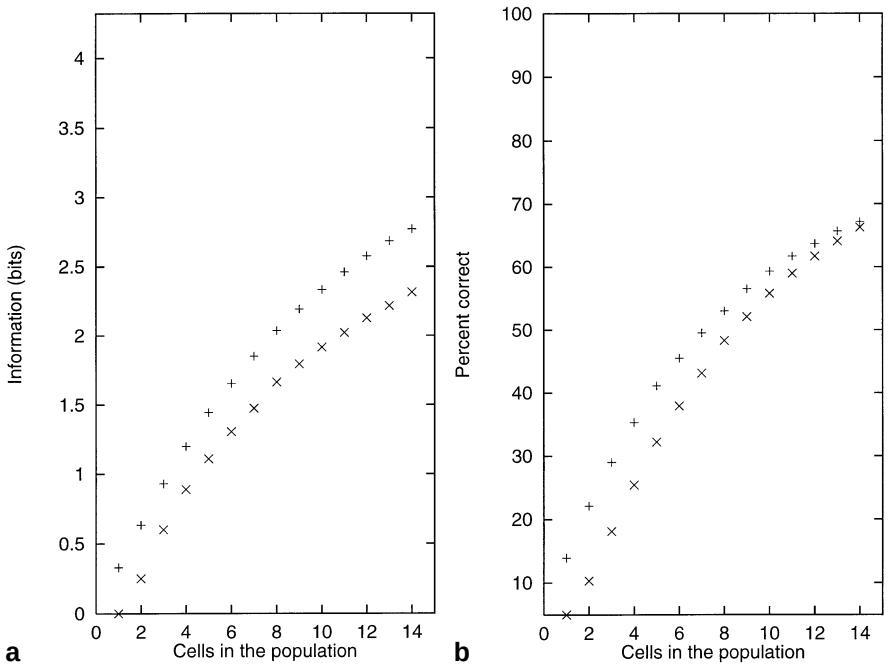

Fig. 4a The values for the mean information available in the responses of different numbers of these neurons in each trial, about which of a set of 20 face stimuli have been shown. The decoding method was dot product (×) or probability estimation (+). **b** The percentage correct for the corresponding data to those shown in **a**

on $I(s,s')$ and on the more efficient cross-validation procedure, although all of the following points can easily be seen to be valid in general.

Ceiling effects and redundancy

As shown in Figs. 1 and 2, both the information and the percentage correct show what amounts to a ceiling effect as the number of cells is increased towards 14. For percentage correct the ceiling is of course at 100%, and for information it is at $\log_2(S)$=20 (i.e. ≈4.3 bits) for our stimulus set. These ceilings correspond to the top of the respective figures. The deviation from a linear rise, in both measures, as more cells are included in the population, appears to be entirely due to the measures approaching their ceilings. This is shown in Fig. 3 by repeating the analysis with fewer stimuli than included in the full set. The multiplication signs correspond to information (Fig. 3a) and percentage correct (Fig. 3b) calculated for a subset (randomly chosen) of nine stimuli, and the stars for one of just four stimuli. The ceiling for information must be correspondingly lowered to $\log_2(9)$=3.17 and $\log_2(4)$=2 bits, and the baseline for percentage correct raised to 1/S=11.11% and 25%. Note that the initial slope in the rise of both measures depends on the particular subset of stimuli, not on its size alone. For example, the mean information per cell with four face stimuli was found to range from 0.04 to 0.57 bits with these cells and choosing an assortment of different random subsets of the 20 stimuli available. Sooner or later, depending on the initial slope, the increase with population size must slow down and saturate to stay below the ceiling value.

As for the rate of slowing down, it is intriguing that the way information depends on the number of cells in the population is fairly close to what would be predicted by a simple model (Gawne and Richmond 1993) of the relation between the information provided by different cells. In this model, each cell provides a fraction of the information required to discriminate the 20 stimuli perfectly [$1–\Phi=I_1/\log_2(20)$], and the information provided by any one cell has a random overlap with that provided by any other cell. Then, if a fraction Φ of the information is missing when looking at just one cell, a fraction Φ^2 is missing, on average, when looking at two cells, and so on. The mean fraction of information provided by C cells is $1–\Phi^C$, and this is the curve plotted with the data in Fig. 3a, where Φ has been chosen as a fit parameter. The curve is seen to be a reasonable match to the trend in the data, except perhaps at the higher numbers of cells.

The degree to which the rise of information with the number of cells considered is below linearity has been linked in the literature to the notion of redundancy. For example, Gawne and Richmond (1993) took as a measure of redundancy the overlap between the information provided by pairs of nearby cells, an overlap that, when using the model above to fit results for single cells and for pairs, turns out to be just $1–\Phi$. For our three sets of stimuli this mean overlap (as extracted by fitting the model to the three data sets) would be roughly 8% for 20 stimuli, 9% for nine and 19% for four, but widely different values are obtained by selecting different subsets of stimuli. In fact, to the extent that the model provides a good fit, the value of the mean overlap is the same as the fraction of information provided by single cells, out of the maximum at the ceiling, and this fraction depends very strongly, as stated above, on both the size and the composition of the stimulus set. The implications of these results for the notion of redundancy is considered in the Discussion.

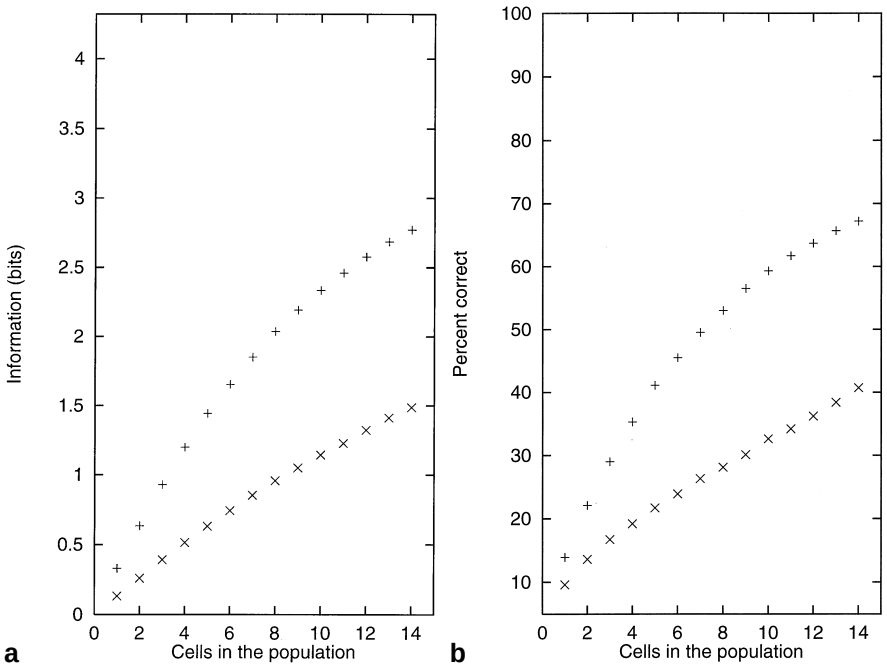

Fig. 5a The values for the mean information available in the responses measured in the usual 500-ms period (+), and in a shorter 50-ms period (×), both starting 100 ms after stimulus onset. The decoding method was probability estimation. **b** The percentage correct for the corresponding data to those shown in **a**

Biologically plausible decoding

The results of analyses to compare DP decoding and PE decoding are shown in Fig. 4. The percentage correct achieved with the DP algorithm comes progressively closer to that obtained with the PE algorithm (used in all previous analyses), as the number of cells increases from 2 to 14 (Fig. 4b). Thus, the DP decoding algorithm functions about as well, in indicating the actual stimulus, as the PE algorithm, and possibly would have functioned even better if larger numbers of cells had been included in the sample. The information extracted by the DP algorithm, instead, only comes to approximately 80-85% of that extracted by the PE algorithm. The information measure reflects the full range of estimated probabilities for all stimuli, and one should bear in mind that the DP algorithm does not really attempt to perform this estimation correctly. The reason that the measures are, respectively, zero and chance with one cell for DP decoding is, obviously, that then the DP of the test trial vector of cell responses with any of the mean response vectors to the stimuli is essentially meaningless.

Information encoded by an ensemble of neurons in a short time window

It is interesting to analyse whether considerable information is available from the population of neurons in short post-stimulus time periods, as has been found for single neurons (Tovee et al. 1993; Tovee and Rolls 1995). The values for the mean information [$I(s,s')$, extracted with the PE algorithm] available in a 50-ms period starting 100 ms post-stimulus, are displayed in Fig. 5a, together with those for the longer 500-ms windows. The same data produced the percentage correct predictions reported in Fig. 5b. It can be seen that both the information and the percentage correct rise more linearly with population size from their baseline level than those measured over 500 ms. The more linear increase with the short analysis period of 50 ms is because the information provided by the 14 cells, although quite high at 1.49 bits, has not approached the ceiling of 4.32 bits required to encode the set of 20 stimuli. The population of 14 cells provides, in the 50-ms period, 54% of the information it provides in the 500-ms period (1.49 vs 2.77 bits). Part of the reason this proportion is so high is that the information ceiling given the stimulus set size is being approached with the 500-ms period. For this reason, and because in principle apart from such a ceiling effect the information increases in proportion to the number of neurons in the ensemble, we report also that single cells on average provide 0.13 bits, or 40%, in 50-ms time periods, of what they provide over the 500-ms window, which is 10 times longer.

Relationship between information and percentage correct

One can see from Fig. 1 that the rise of information with population size parallels very closely that of percentage correct. This is made explicit in Fig. 6, where the information values are plotted against percentage correct. The information axis (ordinate) is normalized so that it reaches full scale when full information about the stimulus set, i.e. $\log_2(20) \approx 4.3$ bits is obtained from the population responses. The percentage-correct axis (abscissa) extends from chance level (which in our case is at 1/20=0.05) to 1. The fact that the data points are very close to the 45° line has some implication for the struc-

ture of the stimulus set as coded by these cells (Treves 1997; Treves et al. 1996). This is because, while percentage correct is a measure that is not affected by how good the estimated probabilities were, apart from the highest one, information is affected and thus reflects the structure of perceived similarities and differences between different stimuli. To understand this point, the two lines shown with the data points illustrate the dependence of $I(s,s')$ on $P(s^P=s)$ in two simple, idealized cases. In one, which yields the upper curve, stimuli are supposed to group naturally, as perceived by this population, into classes of equal size Z; different classes are perfectly discriminated by the cells, whereas in each class there is such similarity that individual stimuli are not at all discriminated one from the other. Obviously Z is taken to vary to give the required percentage correct, which in this case is just $1/Z$; the information is $\log_2(20/Z)$. In the second case, which yields the lower curve, no class structure exists, and stimuli are either discriminated individually (with probability q) or confused with all the others (with probability $1–q$). The parameter q is taken to give the required percentage correct, which is $q+(1–q)/20$; the information results in: $(q+(1–q)/20)\log_2(20q+1–q)+(19/20)(1–q)\log_2(1–q)$. The fact that the experimental relationship turns out to be intermediate between these two extremes indicates that the stimuli are coded by these cells as having a structure of similarity to each other that is slightly more complex than in the two trivial situations mentioned. Not much more can be inferred, as an infinite number of different structures would generate identical intermediate relationships; for example, one still very simple situation in which the fraction of information available equals the fraction of percentage correct above chance (close to what the data show) is when classes exist but also each stimulus can be discriminated, with a certain probability, within its class. It would be interesting to find out whether the extensive averaging that underlies our numerical results in itself might tend to produce data very close to the 45° line, because for example that is where the highest density of similarity structures may be packed. Preliminary results, however, (unpublished observations) indicate a very different relationship between percentage correct and information when analysing the responses of cells in different brain areas and to different external correlates.

Discussion

The analyses described here elucidate a quantitative approach to analysing the representation provided by a population of neurons. First, the mean information about a set of 20 equiprobable stimuli, available on any trial $[I(s,s')]$, was found to increase approximately linearly with the number of cells from which the best estimate was made: from 0.33 bits, with one neuron, up to 2.77 bits, available from all 14 neurons. To be able to analyse in detail this quasi-linear dependence with the limited number of trials available in this experiment (and in many similar experiments with mammals), we have shown that it is important to select with great care the exact type of information quantity and the procedure to measure it from the data. We showed that failure to cross-validate, failure to apply a correction for limited sampling (i.e. the limited number of trials available) and calculation of the information contained in the frequency table of P^F_N could all result in unreliable estimates of the information encoded by the ensemble of neurons. Overcoming these difficulties has been responsible for the slower development of information-theoretic analyses of mammalian neuronal recordings, relative to their use with invertebrate recordings, in which very many trials of data can be obtained (Bialek et al. 1991). In the next four paragraphs we discuss the central aspects of this quasi-linear relationship before returning to the more general issues.

Most of the deviation from a simple linear increase was shown, utilising subsets of the complete stimulus set, to be due to a simple ceiling effect. The ceiling effect occurs because with the total information in the stimulus set being limited (to 4.32 bits, that required to code the 20 stimuli in the set), larger numbers of cells in the population considered are forced to be more redundant, that is, to have greater overlaps. When the limit on the amount of information is lowered by including fewer stimuli, obviously the information tends to saturate at lower values. Conversely, if single cells carry, on average, a smaller proportion of the maximal information, then values close to saturation are approached only for larger populations, and the increase in information is much more linear as cells are added to the sample. This is exactly what is shown in Fig. 5, in which the analysis period for the neuronal response was 50 ms, the information available from one cell was 0.13 bits (and 9.4% correct for the 20 stimuli) and the information increased more linearly to 1.49 bits (and 41% correct) as the number of cells in the population was increased to 14. This supports the explanation for the ceiling effect and suggests that, were the stimulus set much larger (so that much more information could be extracted from it), and if the absolute amount provided by individual cells were to stay roughly the same (this being dependent on the choice of stimuli), then the information would continue to increase almost linearly over a wider range of population sizes. This issue is also addressed in a separate paper based on computer simulations (Abbott et al. 1996), and the results of those simulations support the hypothesis based on the neurophysiological results described here: that the number of stimuli that can be encoded by a population of neurons in this part of the visual system increases approximately exponentially as the number of cells in the sample increases. That is, the log of the number of stimuli increases approximately linearly as the number of cells in the sample is increased. This is in contrast to a local encoding scheme (of grandmother cells), in which it is the number of stimuli encoded that increases linearly with the number of cells in the sample.

The conclusion based on the neurophysiological data described here is subject to the limits of the actual stimulus set (20 stimuli) and cell population (14 neurons) available from real experiments; and the study by Abbott et al. (1996) extends the conclusions to much larger stimulus sets by simulation.

The hypothesis that the deviation from linearity is due *entirely* to the ceiling effect is supported by the reasonable fit to the data provided by the simple model introduced by Gawne and Richmond (1993), which requires as a fitting parameter only the mean information from single cells (the initial slope in the rise). The model is that in which each cell provides a mean fraction $1–\Phi=I_1/\log_2(S)$ of the total information required to discriminate S stimuli perfectly, any element of information has therefore a "chance" (Φ) to be missed by a cell, and, being what is coded by one cell in a random relation to what is coded by other cells, it has a chance Φ^C to be missed by C cells. $I_C=(1–\Phi^C)\log_2(S)$ is therefore the information provided by C cells. The crucial aspect of the model (which appears to have been overlooked by Gawne and Richmond 1993, in their discussion of the correlations between cell pairs) is that it predicts zero redundancy in the limit of very large S, provided the information provided by single cells does not grow proportionally to the ceiling. This is because the overlap, or degree of redundancy, is, again, just $1–\Phi$ ($1–\Phi=y$ in the notation of Gawne and Richmond 1993), which would tend to zero for very large sets of stimuli. The model may well not be exact, but it remains true that the observed redundancy should in no way be taken to characterize the absolute information processing capabilities of single temporal lobe visual cortex cells or of groups of cells, but only their performance given the stimulus set. In quasi-real-life conditions, it may be that the absolute amount of information provided by one cell could be approximately constant, as the stimulus set is greatly enlarged to include a naturalistic set of stimuli. This would happen if, when stimuli are added, the overall statistics of the responses remained relatively constant, as in the simulations of Abbott et al. (1996). However, in that case the fraction of the maximum (and with it, presumably, the redundancy overlap, which in the model is numerically identical) will not be a constant, but rather will decrease as the inverse log of the size of the stimulus set.

Another point to note in relation to previous work with similar goals (Gochin et al. 1994) is that failure to take ceiling effects into account, to select the appropriate information quantities and to measure them by correcting for limited sampling easily results in artifactual findings, such as the postulated tendency of novel information to decrease with the inverse square root of the population size (Gochin et al. 1994; the novel information is the information provided by a population of C cells divided by C times the mean information provided by single cells). In the simple model used above, the novel information is $(1–\Phi^C)/[C(1–\Phi)]=(1/C)\Sigma_{k=0}{}^{C-1}\Phi^k$; that is, it scales in a different way from an inverse square root. We note that, with only five stimuli in their set, Gochin et al. (1994) had a rather low ceiling on the amount of information required to encode the set, and the failure to take this into account, together with the very few population sizes considered, makes their conclusions subject to artefact.

To conclude with the implications of the present data for exponential encoding and redundancy in the brain, we stress again some of the important limitations in our findings. First, the responses of only 14 cells to only 20 face stimuli formed the basis of the analysis. More face cells were recorded, in part also with larger sets of face stimuli, and qualitatively the data looked similar, but these additional cells either were in a second monkey or were not recorded for enough trials per stimulus to be part of the present quantitative analysis. Coding for a few thousand faces (of the order of the number that humans may feel confident at discriminating) presumably involves an order of magnitude more face cells. The applicability of our conclusions to face encoding remains thus an extrapolation, although it is an extrapolation supported by very plausible simulation results, as shown by Abbott et al. (1996). Second, the cells recorded were all located in a restricted portion of cortex and were all face cells responding to face stimuli. It is thus in principle possible that face encoding may display rather different features in other parts of cortex, and it is quite likely and entirely natural that the encoding of very different categories of visual stimuli may proceed along different principles. Seemingly different types of population encodings have, for example, been discussed for arm movements in three-dimensional (3D) space (by motor cortex neurons; Georgopoulos et al. 1988) and for the prevailing (1D) direction of motion of random dots (by visual cells of the middle temporal area; Zohary et al. 1994). We regard our face cell data as more indicative of population encoding of classes of stimuli spanning a high-dimensional space, as, e.g. discussed by Tanaka (1993). Third, the cells were not recorded simultaneously, and thus trial-to-trial correlations in their responses were not included in the analysis. Including such correlations might result in either higher or lower information values, depending on the type of correlations found. It is easy to construct response structures that would lead to either result. Significant correlations in the firing of small groups of cells have been found, e.g. in the early visual cortex stimulated with moving bars (Gray et al. 1989) or in the frontal cortex in relation to behavioural state (Abeles et al. 1995). We are aware of no positive evidence for such correlations, however, directly relevant to stimulus encoding in higher visual cortices. The findings of Gawne and Richmond (1993) and Gochin et al. (1994) indicate weak correlation effects, and similarly shuffling the simultaneously recorded responses of rat hippocampal place cells to control for such effects made little difference (our unpublished observations on data kindly provided by the McNaughton laboratory). It remains very important, as noted below, to supplement the present analyses with analyses performed on simultaneously recorded responses from a similar population of face cells

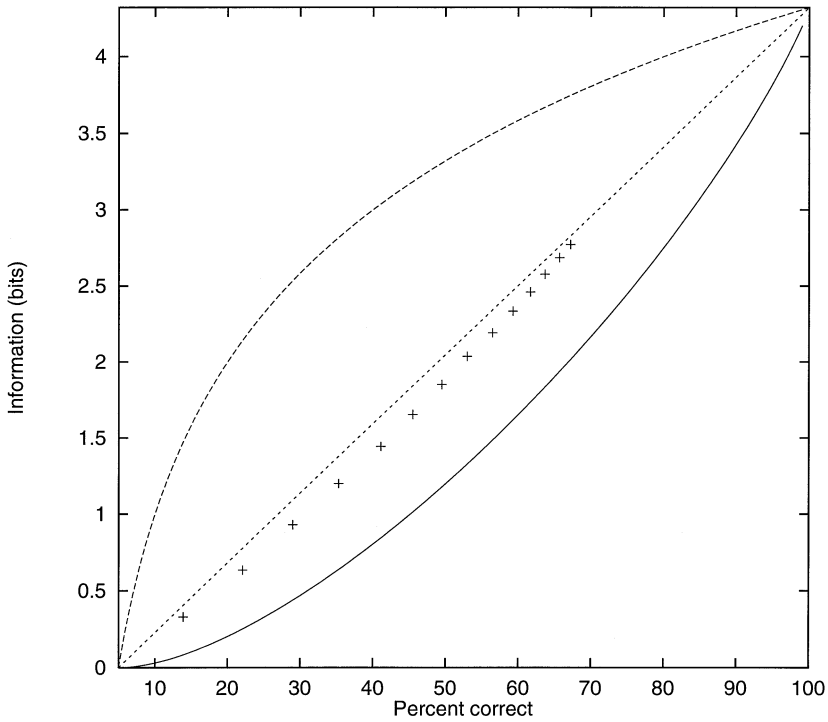

Fig. 6 The corrected values for the information $I(s,s')$ (with PE decoding and the efficient cross-validation procedure) shown as a function of percentage correct when both are estimated from the responses of subsets of the 14 neurons to the 20 face stimuli

responding to a similar set of face stimuli. Finally, we note the possibility that the details of the experimental paradigm used here (the level of attention required, the degree of familiarity with the stimuli used, etc.) might have had an influence on the results obtained.

The second general issue arising from the present investigation is that it was found that the percentage correct behaved fairly similarly to the information in the way it increased as the number of neurons from which the best estimate was made increased from 1 to 14 (see Fig. 6). Performance with one neuron was approximately 14% correct (chance was 5% correct) and with 14 neurons was 67% correct. The nearly identical behaviour of information and percentage correct indicates that the stimulus set is coded by these neurons as having a structure of mutual similarities (an intrinsic metric) in which at least two cluster sizes are present. More generally it suggests the use of specific choices of stimulus sets that might generate results at variance with the present ones. The point here is that the percentage correct measure takes into account only the best estimate of which stimulus was presented; it is unaffected by how good the second best, third best, etc. guesses might have been. If the best guess was wrong and the second best guess would have been right, then the percentage correct measure does not reflect this. On the other hand, the information measure does reflect structure of this type. The fact that the percentage correct and information measures are above the lower line in Fig. 6, and below the upper line, thus has interesting implications for the way in which these neurons categorise stimuli: for example it could indicate that second-best guesses are better than chance. More detailed analysis (Treves 1996) is required, though, to make these implications explicit, because averaging by itself could produce a behaviour intermediate between extremes.

Third, alternative algorithms were used to estimate which of the mean response vectors (one for each stimulus) most closely matched the vector of cell responses being produced by a test stimulus. The PE algorithm, which approximates a theoretically correct estimate of the relevant probabilities (that a given response had been elicited by any one stimulus), may not be the way the (next population of neurons in the) brain would decode the neuronal responses. It was found that, with a neurally plausible algorithm (the DP algorithm) that calculates which mean response vector the neuronal response vector was closest to by performing a normalized DP (equivalent to measuring the angle between the test and the mean response vector), the same generic results were obtained, with similar percentage correct and only a 15–20% reduction in information compared with the more efficient (PE) algorithm. This is an indication that the brain could utilise 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 DP 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; Treves and Rolls 1991). One of the important points made here is that, because the representational capacity of a set of neurons increases exponentially, neurons in the next brain region would each need to sample the activity of only a reasonable number

(e.g. a few hundred) of what might be a much larger cell population and yet still obtain information about which of many thousands of stimuli had been shown. In particular, the characteristics of the actual cells described here indicate that the activity of: 15 neurons would be able to encode 192 face stimuli (at 50% accuracy); 20 neurons, 768 stimuli; 25 neurons, 3072 stimuli; 30 neurons, 12288 stimuli; and 35 neurons, 49152 stimuli (Abbott et al. 1996. The values are for the optimal decoding case and are derived using somewhat different methods from the present ones; they are given here for the purpose of illustration.)

Fourth, the results shown in Fig. 5 with 50-ms decoding highlight the value of having large numbers of neurons of the type described here, for they make it clear that part of the value is that information can be made available very rapidly about which stimulus is present if the responses of a population of neurons, rather than just a single neuron, are considered. Moreover, the fact that the representation provided by each neuron is apparently in a random relation to that provided by other neurons means that the information is available very rapidly from whichever subset of neurons is taken. This rapid availability of information from a *population* of neurons is one factor that contributes to the very rapid processing of information from stage to stage in the visual cortical areas (see Rolls 1994; Rolls and Tovee 1994), for it means that the information from one cortical area can be extracted very rapidly (in, e.g. 20 ms) by the next.

Fifth, it is of great interest that the information about which stimulus is present can now be read off from the end of the visual system to identify *what* is being seen, and that this read-out of information can be performed excellently if only the firing rates of the neurons are taken into account. The results in this paper thus provide good evidence that the temporal relationships between the spike times of different neurons are not at all a necessary part of the neural code used at the end of the visual system (cf. Engel et al. 1992). It is of course possible that if the temporal relationships between the spike firings of the neurons in the ensemble were taken into account more information would be available about which stimulus was shown. However, the results described in this paper show that a very great deal of information can be read out from the responses of an ensemble of neurons about which stimulus was shown without taking into account the relative timing of the spikes in the different neurons. The results described here show that using just the firing rates would be sufficient for the good operation of at least this part of the temporal visual cortex.

However, a point that certainly merits further investigation is the effect of generating pseudosimultaneous trials, rather than recording simultaneously from large populations of cells (Wilson and McNaughton 1993). Particularly in exploring fine points such as the presence of trial-to-trial correlations in the responses, it is important that the present studies be integrated with new ones once simultaneously recorded data become available for these cells. However, we believe it likely that, if the present data had been recorded simultaneously, then, if anything, more information would have been available in the neuronal responses of the population, because any general shift in excitability of the cell population from trial to trial – a particularly simple and common type of correlation – would be compensated for by a neurophysiologically plausible decoding procedure such as computing the DP, whereas the same shift would increase the noise level when pseudosimultaneous trials are constructed. It would only be in more unlikely situations of small sets of cells having rather idiosyncratic variability in their activity that the current procedure might overestimate the information available from a simultaneously recorded population of neurons. Further, we have been able to apply the same kind of analyses to data recorded simultaneously in the rat and kindly provided by the laboratories of Bruce McNaughton and Gabriele Biella (unpublished observations). With simultaneously recorded data, it is possible to control for the effects of generating pseudosimultaneous trials by simply shuffling trials independently for each cell. Such a control procedure resulted in very small differences, if any, thus indirectly supporting the assumption that also with our data the effect of generating pseudosimultaneous trials is minimal. A further point is that simultaneous recordings from the thalamus provide a counterexample of a case in which there appears to be no linear increase of information with population size (Panzeri et al. 1995; this is true both before and after shuffling), therefore removing the doubt that the near-linear increase found for the visual cortex cells described here might have been an almost automatic consequence of either the decoding procedure or the generation of pseudosimultaneous trials.

The results described here thus provide evidence that when real neuronal responses in the brain are considered, so that the trial to trial variability of individual neuronal responses is taken into account, and also the degree to which each stimulus produces a somewhat different set of responses in a population of neurons from other stimuli, then the essentially exponential increase in coding capacity that is potentially a property of distributed representations can be realised.

It has been argued elsewhere (Rolls and Tovee 1995) that the rather widely distributed encoding found in this population of neurons allows a relatively large amount of information about a set of stimuli to be provided by such a population, provided of course that they do not have the same profile of responsiveness to the set of stimuli. Such a representation would be ideal for *discrimination*, for the maximum information suitable for comparing fine differences between different stimuli would be made available across the population. However, a representation as distributed as this would not be appropriate for a memory system, in which the aim is to store a large number of memories. In an associative memory containing neurons with continuously variable firing rates, such as the autoassociative memory believed to be implemented in the hippocampus (Rolls 1989; Treves and Rolls 1994), the maximum number of firing patterns that can

be retrieved increases approximately with the inverse of the sparseness (a) of the neuronal representation (Treves 1990; Treves and Rolls 1991). It is therefore proposed that these fundamentally different constraints, representational capacity versus storage capacity, account for the different sparsenesses of representations found in the high-order sensory cortices such as the temporal cortical areas described here and by Rolls and Tovee (1995), and in memory systems such as the hippocampus (Treves and Rolls 1994), amygdala, and orbitofrontal cortex (Rolls 1989, 1992a, c). In the sensory cortex, a relatively distributed representation may be used in order to optimize discriminative ability. In memory systems, much more sparse representations may be used in order to maximize the number of memories that can be stored. We note that many of the cells described here have other properties that make them suitable for discrimination between faces, including invariance with respect to size, spatial frequency, translation and even view in some cases (see Rolls 1992a, 1994, 1995; Rolls and Tovee 1995).

Acknowledgements This research was supported by Medical Research Council Grant PG8513790 to Professor E. T. Rolls. We are very grateful for helpful comments and assistance provided by Professor L. A. Abbott and Drs. S. Panzeri and R. Baddeley.

References

Abbott LA, Rolls ET, Tovee MJ (1996) Representational capacity of face coding in monkeys. Cereb Cortex 6:498–505

Abeles M, Bergman H, Gat I, Meilijson I, Seidemann E, Tishby N, Vaadia E (1995) Cortical activity flips among quasi-stationary states. Proc Natl Acad Sci USA 92:8616–8620

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

Barlow HB (1972) Single units and sensation: a neuron doctrine for perceptual psychology? Perception 1:371–394

Baylis GC, Rolls ET, Leonard CM (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 GC, Rolls ET, Leonard CM (1987) Functional subdivisions of temporal lobe neocortex. J Neurosci 7:330–342

Bialek W, Rieke F, Ruyter van Steveninck RR de, Warland D (1991) Reading a neural code. Science 252:1854–1857

Bruce C, Desimone R, Gross CG (1981) Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. J Neurophysiol 46:369–384

Churchland PS, Sejnowski TJ (1992) The computational brain. MIT Press, Cambridge, MA

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

Desimone R, Gross CG (1979) Visual areas in the temporal lobe of the macaque. Brain Res 178:363–380

Desimone R, Albright TD, Gross CG, Bruce C (1984) Stimulus-selective properties of inferior temporal neurons in the macaque. J Neurosci 4:2051–2062

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

Foldiak P (1993) The ideal homunculus – statistical inference from neural population responses. In: Eeckman FH, Bower JM (eds) Computation and neural systems. Kluwer Academic, Boston, MA, pp 55–60

Gawne TJ, Richmond BJ (1993) How independent are the messages carried by adjacent inferior temporal cortical neurons? J Neurosci 13:2758–2771

Georgopolous AP, Kettner RE, Schwartz AB (1988) Primate motor cortex and free arm movements to visual targets in three-dimensional space. II. Coding of the direction of movement by a neuronal population. J Neurosci 8:2928–2937

Gochin PM, Colombo M, Dorfman GA, Gerstein GL, Gross CG (1994) Neural ensemble encoding in inferior temporal cortex. J Neurophysiol 71:2325–2337

Golomb D, Hertz JA, Panzeri S, Richmond BJ, Treves A (1996) How well can we estimate the information carried in neuronal responses from limited samples? Neural Computation 8 (in press)

Gray CM, Konig P, Engel AK, Singer W (1989) Oscillatory responses in cat visual cortex exhibit intercolumnar synchronization which reflects global stimulus properties. Nature 338:334–337

Gross CG, Desimone R, Albright TD, Schwartz EL (1985) Inferior temporal cortex and pattern recognition. Exp Brain Res [Suppl] 11:179–201

Hinton GE, McClelland JL, Rumelhart DE (1986) Distributed representations. In: Rumelhart DE, McClelland JL, (eds) Parallel distributed processing. MIT Press, Cambridge, MA, pp 77–109

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

Kjaer TW, Hertz JA, Richmond BJ (1994) Decoding cortical neural signals: network models, information estimation and spatial tuning. J Comput Neurosci 1: 109–139

Maunsell JHR, Newsome WT (1987) Visual processing in monkey extrastriate cortex. Annu Rev Neurosci 10:363–401

Optican LM, Gawne TJ, Richmond BJ, Joseph PJ (1991) Unbiased measures of transmitted information and channel capacity from multivariate neuronal data. Biol Cybern 65:305–310

Panzeri S, Treves A (1996) Analytical estimates of limited sampling biases in different information measures. Network 7:87–107

Panzeri S, Biella G, Sotgiu ML, Treves A (1995) Information theoretical analysis of thalamocortical ensemble coding during noxious stimulation. Soc Neurosci Abstr 21:114

Perrett DI, Rolls ET, Caan W (1982) Visual neurons responsive to faces in the monkey temporal cortex. Exp Brain Res 47:329-342

Rolls ET (1981a) Processing beyond the inferior temporal visual cortex related to feeding, learning, and striatal function. In: Katsuki Y (eds) Brain mechanisms of sensation. Wiley, New York, pp 241–269

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

Rolls ET (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 ET (1989) Functions of neuronal networks in the hippocampus and neocortex in memory. In: Byrne JH, Berry WO (eds) Neural models of plasticity: experimental and theoretical approaches. Academic, San Diego, pp 240–265

Rolls ET (1991) Neural organisation of higher visual functions. Curr Opin Neurobiol 1:274–278

Rolls ET (1992a) Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. Phil Trans R Soc 335:11–21

Rolls ET (1992b) The processing of face information in the primate temporal lobe. In: Bruce V, Burton M (eds) Processing images of faces. Ablex, Norwood, New Jersey, pp 41–68

Rolls ET (1992c) Neurophysiology and functions of the primate amygdala. In: Aggleton JP (ed) The amygdala. Wiley-Liss, New York, pp 143–165

Rolls ET (1994) Brain mechanisms for invariant visual recognition and learning. Behav Proc 33:113–138

Rolls ET (1995) Learning mechanisms in the temporal lobe visual cortex. Behav Brain Res 66:177–185

Rolls ET, Tovee MJ (1994) Processing speed in the cerebral cortex and the neurophysiology of visual masking. Proc R Soc Lond B Biol Sci 257:9–15

Rolls ET, Tovee MJ (1995) The sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. J Neurophysiol 73:713–726

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

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

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

Suga N (1989) Principles of auditory information-processing derived from neuroethology. J Exp Biol 146:277–286

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

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

Tovee MJ, Rolls ET (1995) Information encoding in short firing rate epochs by single neurons in the primate temporal visual cortex. Vis Cogn 2:35–58

Tovee MJ, Rolls ET, Treves A, Bellis RP (1993) Information encoding and the responses of single neurons in the primate temporal visual cortex. J Neurophysiol 70:640–654

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

Treves A (1990) Graded-response neurons and information encodings in autoassociative memories. Phys Rev A 42:2418–2430

Treves A (1997) On the perceptual structure of face space. Biosystems 40:189–196

Treves A, Panzeri S (1995) The upward bias in measures of information derived from limited data samples. Neural Comp 7:399–407

Treves A, Rolls ET (1991) What determines the capacity of autoassociative memories in the brain? Network 2:371–397

Treves A, Rolls ET (1994) A computational analysis of the role of the hippocampus in memory. Hippocampus 4:374–391

Treves A, Panzeri S, Rolls ET (1994) Correcting for limited sampling in estimates of the information carried by neuronal responses. Soc Neurosci Abstr 20:577

Weiskrantz L, Saunders RC (1984) Impairments of visual object transforms in monkeys. Brain 107:1033–1072

Wilson M, McNaughton BJ (1993) Dynamics of the hippocampal ensemble code for space. Science 261:1055–1058

Zohary E, Shadlen MN, Newsome WT (1994) Correlated neuronal discharge rate and its implications for psychophysical performance. Nature 370:140–143