ISSN 0106-2646

NORDITA preprint

NORDITA - 95/24 S

INFORMATION FLOW AND TEMPORAL CODING IN PRIMATE PATTERN VISION

J. Heller[1].
Division of Applied Sciences, Harvard University, Cambridge, MA.

J.A. Hertz.
Nordita, Blegdamsvej 17, DK-2100 Copenhagen Ø, Denmark.

T. Kjær[2].
Laboratory of Neuropsychology, National Institute of Mental Health, Bethesda MD 20892, USA.

B. Richmond[3].
Laboratory of Neuropsychology, National Institute of Mental Health, Bethesda MD 20892, USA.

In press: J. Computational Neurosci.

[1]Current address: Jewish Theological Seminary, New York, NY.
[2]Current address: Nordita, Blegdamsvej 17, DK-2100 Copenhagen Ø, Denmark.
[3]Reprint requests: Laboratory of Neuropsychology, 49 Convent Drive, Bethesda, MD 20892-4415.

NORDITA · Nordisk Institut for Teoretisk Fysik

Blegdamsvej 17 DK-2100 København Ø Danmark

??, ??, 1–17 (??)

In press: J. Computational Neurosci.

Information Flow and Temporal Coding in Primate Pattern Vision

JOSHUA HELLER *
jheller@ln.nimh.nih.gov
Division of Applied Sciences
Harvard University
Cambridge, MA

JOHN A. HERTZ
hertz@nordita.dk
Nordita
2100 Copenhagen, Denmark

TROELS W. KJÆR **
kjaer@nordita.dk
Laboratory of Neuropsychology
National Institute of Mental Health
Bethesda, MD 20892

BARRY J. RICHMOND †
bjr@ln.nimh.nih.gov
Laboratory of Neuropsychology
National Institute of Mental Health
Bethesda, MD 20892

Received ??. Revised ??.

Abstract.

We perform time-resolved calculations of the information transmitted about visual patterns by neurons in primary visual and inferior temporal cortices. All measurable information is carried in an effective time-varying firing rate, obtained by averaging the neuronal response with a resolution no finer than about 25 ms in primary visual cortex and around twice that in inferior temporal cortex. We found no better way for a neuron receiving these messages to decode them than simply to count spikes for this long. Most of the information tends to be concentrated in one or, more often, two brief packets, one at the very beginning of the response and the other typically 100 ms later. The first packet is the most informative part of the message, but the second one generally contains new information. A small but significant part of the total information in the message accumulates gradually over the entire course of the response. These findings impose strong constraints on the codes used by these neurons.

Keywords:

1. INTRODUCTION

Recent studies have established [16], [14], [18], [22], [11] that neurons in the primate visual system convey information using the timing of their action potentials. However, very little is known about just how this information is represented in their responses. In this investigation we try to determine what features of the neuronal response carry information about a stimulus. We ask two questions in particular:

1. To what degree does the information transmission depend on the precise timing of the spikes?
2. What is the time course of the signal – for example, is the information transmitted in a single short burst, or distributed more evenly over the course of the response?

We address these questions in two brain areas, primary visual cortex (V1) and inferior temporal cortex (IT).

A visual system neuron may be seen as a communication system. The visual stimulus (a message) is translated by the neuron (a transmitter) into a set of neural firing patterns (a code). The neuron (acting as a channel) passes that encoded message on to other neurons (receivers). Ultimately, the signals are used by other brain centers (the destinations) to determine the nature of the original stimulus. In fact, information about the stimulus is carried by many neurons in the visual system, and, working as an ensemble, they can transmit much more information than any one neuron. However, our goal in this study was quantify the behavior of the building blocks of the system, individual neurons.

Probing the structure of this neuronal code requires a reliable tool for estimating transmitted information. We use an artificial neural network designed for this task[10], [11]. The inputs to the network are a representation of the neuronal responses. There is one output unit for each stimulus. The strength of each output is an estimate of the probability that its associated stimulus evoked the response used as the input. The transmitted information may be calculated in terms of these probabilities.

We can answer the first question by representing the neuronal response in different ways at the input layer of the network and comparing the resulting estimates of transmitted information. A given representation might capture some features and omit others (such as high frequency components). Finding an adequate "code" to represent the data has a twofold significance. Knowing which code represents the signal best provides insight into the coding scheme used by the neurons themselves. Features which are necessary to conserve the information in the signal are probably important in the cell's transmission of information. Conversely, features which are not needed to conserve the information of the analyzed signal are probably not relevant to the cell's internal representation of the stimulus. Secondly, an important part of analyzing data in statistically sound ways is choosing an appropriate representation for the data – one which makes it easier to identify and quantify sources of error. It is useful to have a representation which captures the important features of the data and can be used to eliminate error. We find strong evidence that the low-frequency components of the signal represent all of the stimulus-related information.

To answer our second question, we measure how the information content of the signal changes over time. Sliding narrow time windows along the response period and calculating the information carried in them permits us to estimate the instantaneous transmission rate as a function of time. We find that there is always a large burst of information early in the response, but subsequent neuronal firing also carries information. The net transmitted information generally continues to grow for several hundred milliseconds after its initial sharp rise, indicating that part of the message carried in the later parts of the response is new, and not a mere repetition of the message in the initial burst.

Many of the computations reported here were carried out as part of J. Heller's undergraduate thesis [8]. Some of these results have appeared in a short abstract [9].

2. METHODS

2.1. Data collection

The experiments that yielded the data analyzed here have been reported previously [17], [7]. They were collected from two different visual areas, V1 and IT, in awake rhesus monkeys, using standard extra-cellular recording techniques. The times of each action potential and each stimulus were recorded with a resolution of 1 ms.

19 cells were recorded from V1 in three monkeys. 13 produced enough data for our analyses. All 19 were complex cells, located in the supragranular layers. The receptive fields were located in the lower contralateral visual field, 1-3 degrees from the fovea [17] Data were also recorded from area TE of inferior temporal cortex. Of the IT neurons, we analyzed the responses of the 11 that had the largest amount of stimulus-related information [7].

* current address: Jewish Theological Seminary New York, NY
** current address: Nordita, 2100 Copenhagen, Denmark
† reprint requests: Laboratory of Neuropsychology, 49 Convent Drive, Bethesda, MD 20892-4415

The stimuli used for the studies were Walsh patterns: two-dimensional black and white 4×4 patterns based on Walsh functions, plus their contrast-reversed counterparts, making a total of 32 stimuli (Fig. 1).
[3]
, [15].

2.2. *Statistics*

Each stimulus is a possible message for the neuron to transmit. The mutual (or transmitted) information is

$$I(S;R) = \left\langle \sum_s P(s|\mathbf{r}) \log_2 \left[\frac{P(s|\mathbf{r})}{P(s)} \right] \right\rangle_{\mathbf{r}}, \tag{1}$$

where S is the set of stimuli s, R is the set of signals (here the neuronal responses) $\mathbf{r}$, $P(s|\mathbf{r})$ is the conditional probability of stimulus class s given an observed response $\mathbf{r}$, and $P(s)$ is the a *priori* probability of stimulus s. The brackets indicate an average over the signal distribution $P(\mathbf{r})$.

Estimating conditional probabilities for categorical data is a standard regression problem. We employ a conventional measure of goodness of fit, maximum likelihood. We select the parameters of the model that come closest to predicting our data, using as our cost function the negative log-likelihood

$$E = -\sum_{\mu} \log_2 P(s^{\mu}|\mathbf{r}^{\mu}), \tag{2}$$

where $\{s^{\mu}, \mathbf{r}^{\mu}\}$ are the data used to make the fit: s^{μ} is the stimulus and $\mathbf{r}^{\mu}$ the response in trial number μ. As E measures the degree to which the data and the fit form of $P(s|\mathbf{r})$ differ, we will refer to it hereafter as the "fit error", or simply as the error. It can be calculated both for the data used to make the fit (training error) and for independent data (test error).

2.3. *Neural network models*

A neural network can be trained using backpropagation so that given the input $\mathbf{r}$ its outputs provide an estimate of the conditional probabilities $P(s|\mathbf{r})$. Our backpropagation model was very much like the standard backpropagation model [20]. The model is pictured in Figure 2, and described in detail in Kjaer et al (1994).

The error (3) is used as the cost function for the backpropagation algorithm, leading the network to search for the parameters in the fit to $P(s|\mathbf{r})$ which give the largest log-likelihood. The learning rate η and the inertia α are used to control the speed/accuracy tradeoff of the learning. α was always .4 and η was set between .0001 and .001, depending on the representation and the cell. All of our networks had 6 hidden units.

Once we have trained the network so that its outputs $O_s(\mathbf{r})$ provide a good estimate of $P(s|\mathbf{r})$, we can substitute $O_s(\mathbf{r})$ for $P(s|\mathbf{r})$ in the expression (1) and average over a data set $\{\mathbf{r}^{\mu}\}$ to estimate the transmitted information:

$$I_{\text{est}}(S;R) = \frac{1}{n} \sum_{\mu,s} O_s(\mathbf{r}^{\mu}) \log_2 \left[\frac{O_s(\mathbf{r}^{\mu})}{P(s)} \right], \tag{3}$$

with $P(s)$ estimated as $n^{-1} \sum_{\mu} O_s(\mathbf{r}^{\mu})$, where n is the total number of samples.

A neural network can theoretically be trained to an arbitrary degree of accuracy as long as the number of samples is finite. However, at some point, the network's learning is based on features of the specific sample, rather than on features of the data in general. We used the "early stopping" procedure described in Kjaer et al (1994) to control such overfitting. We divided the data into training and test segments. The network was tested on the test set while the training set was used to drive the backpropagation algorithm. Training was stopped when the test set error reached a minimum. The data were divided into training and test sets at least 4 different ways for each analysis. In order to obtain the final estimates of information (1) and mean fitting error (2), the individual estimates for the four test sets were averaged.

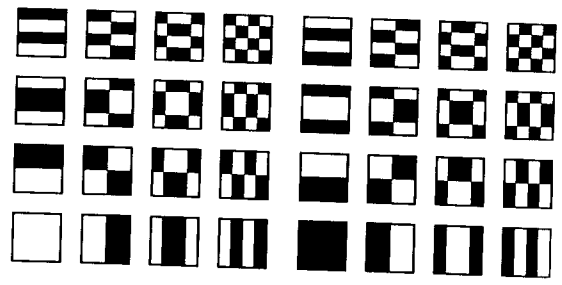

Fig. 1. The Walsh patterns and their contrast-reversed counterparts.

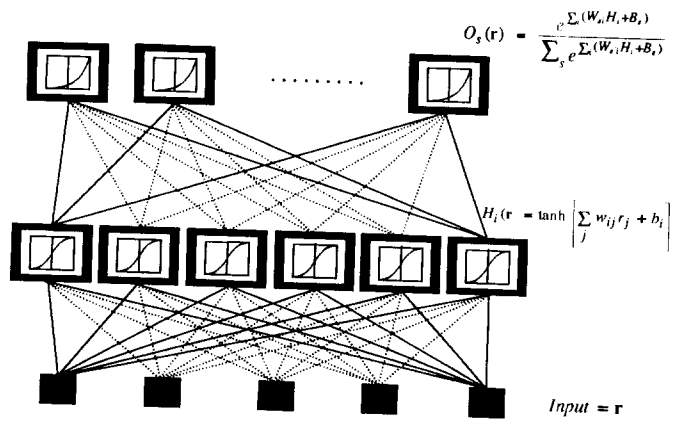

Fig. 2. The network. The components of the representations of the stimuli comprise the inputs, the hidden units have tanh activation functions, and the output units are normalized exponential functions of weighted sums of the hidden unit activations. There is one output unit per stimulus class, and the training targets are 1 for the stimulus class which evoked the response used as input, 0 for all other classes.

Since our test sets were used to determine the network training times, the correct procedure for estimating the transmitted information would be to perform the sample (3) over an entirely different data set, used neither in training nor in determining when to stop the training of the network. However, we have found in previous investigations on these data that, within the confidence limits estimated in the above manner, the values of $I_{\text{est}}(S;R)$ do not depend on whether the sampling is over our training data, our test data, or such new data [11].

2.4. *Data Windowing*

In order to examine the time course of the information flow, we evaluated the response using two different types of time windows. We derived an estimated latency ℓ, which was later updated to reflect more accurate calculations, for each cell. The first window began at ℓ and was 8 ms long.

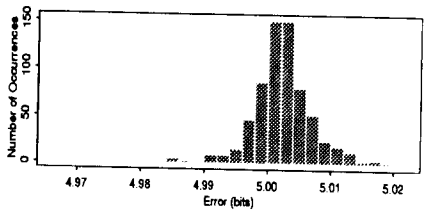

Fig. 3. Network performance on spontaneous activity. Fit errors based on the count representation (c) for a 16-ms wide window were calculated over 116 ms of spontaneous activity (29 window positions, at 4 ms intervals). The histogram shows the distribution of the fit error values. The distribution is almost normal and has a mean of 5.0021 bits and a standard deviation of .0057 bits. (5 bits is the fit error (2) for a network with all outputs equal to the a *priori* stimulus probability 1/32.) Since 4.985 bits is three standard deviations below the mean of this distribution, we assumed that values lower than that could be attributed to the presence of stimulus-related information.

Succeeding windows began at time ℓ but included more and more of the response: up to 320 ms, (the length of the stimuli) and even 368 ms. Other studies used what we called sliding windows. This means that for a given iteration, a set number of milliseconds (16, 24, 32 or 64) of the response, starting at some time t, were encoded using some representation, and the information calculations were performed on the data within this "window". For succeeding iterations, t was incremented by some small value, i.e., the window was effectively swept across the full response period.

All of the windows used in our final analysis were square windows, because our ultimate goal in using sliding windows was to localize features of the data within the time domain. We wanted to find the longest period over which the data could be represented by a single variable. We also tried windows with Gaussian tails, but concluded that while those, or other types of windows, might be relevant to how cells actually integrate responses, square windows provided us with the most accuracy in the time domain, and decided that we would use other techniques to assess the frequency domain.

2.5. Representations

Once a time-segment of the response had been selected for study, it had to be represented in a form which could be used to train the network. We used all of the formats in Table 1 for growing windows of widths 16, 32, 64, 128 and 320. The representations were compared on the basis of their average test set errors. Once preliminary results had been obtained, a few representations were used on a wider selection of widths. A few of these representations were then used on sliding windows of widths 16, 24, 32 and 64. All of these studies were performed on each cell individually.

2.6. Latency

A visual system neuron can change its firing rate earlier in response to some stimuli than to others, or the firing rate might not change at all. Latency is often defined as the delay until the cell's first response to stimulus. However, we were specifically interested in response features that could be used to differentiate among stimuli. Therefore we defined an information latency as the time when enough change had taken place in the response that stimuli could be discriminated at a threshold level. To determine this threshold, we first computed the test error and transmitted information obtained when the network was trained on spontaneous (non-stimulus-related) activity, using a 16 ms-wide window which was swept from 100 ms before the stimulus to 12 ms after it. The windows were spaced at 4 ms intervals. Within each window, the data were assigned to classes and encoded using only the count code c. For each cell, we were able to obtain both fit error and information values, with their means and standard deviations, for 29 window positions (Fig. 3). The information never dropped below 0 bits, and its distribution shows a tail which extends up to about .02 bits. The distribution of the fit error has a mean of 5.0021 and a standard deviation of .0057. Only 1.4% of the values fell below 4.985 (three standard deviations below the mean), and we used this value as our criterion for establishing a relationship between response and stimulus. We assumed that error values higher than this indicated only spontaneous activity, while values lower than this

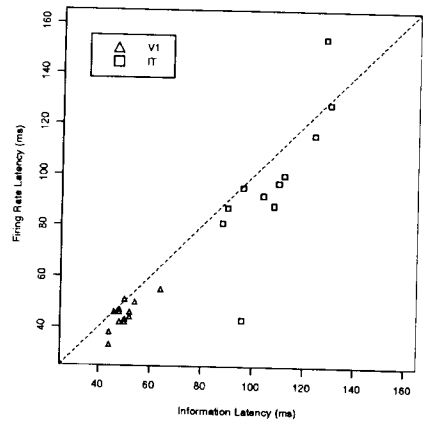

Fig. 4. Latency based on firing rate and information rate. We calculated spike density over all stimuli as a function of time for each cell and set the firing rate latency λ as the first time the spike density rose 20% of the way from its baseline level to its peak. The information latency ℓ was placed at the time when the network error was first significantly below 5 bits (see Fig. 3). The scatterplot compares λ to ℓ for all V1 and IT cells. Two dotted line shows the times of equal latencies for the two measures. The two values fall within 4ms of each other for only 5 V1 cells and 3 IT cells. In 8 V1 and 7 IT cells, the information latency is at least 4 ms longer than the firing rate latency. In 2 IT cells, the firing rate latencies we calculated are not meaningful, but the cells are included for the sake of completeness.

indicated significant differentiability among stimuli. To determine the information latency, we performed this calculation of the mean fit error as the 16-ms wide time window was moved along into the response period. When we reached the point where the error fell below our criterion, we set the latency 2 ms earlier than the end of the window. This procedure defined the latency values ℓ used for growing windows in all of our subsequent analyses.

We also established a criterion for a firing-rate latency, λ, as follows. We summed the spikes in each millisecond over all trials with all stimuli and smoothed the resulting curve with a Gaussian (σ = 5 ms) to produce a spike density estimate. We defined the firing rate-based latency as the first time at which this smoothed spike density

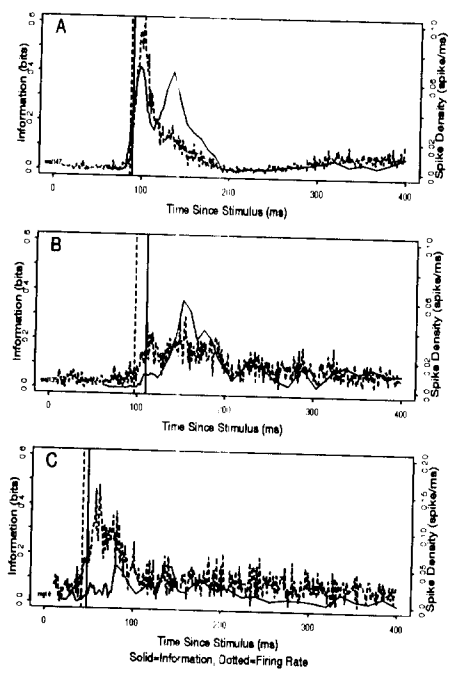

Fig. 5. Firing rate and information rate. We calculated the spike density over all stimuli as a function of time for each cell (dotted curve) and used that to determine the firing rate latency λ (dashed line). A 16-ms sliding window, represented using a count code, was used to produce information transmission as a function of time. (solid curve), and a related statistic the error, was used to compute the information latency ℓ. A. In 8 cells the initial response carried information about which stimulus was being presented, and the two latencies effectively coincided. B. In others the first few ms of the response did not convey any information about which stimulus was being presented: $\ell > \lambda$. C. Sometimes, increased firing rate did not lead to increased information transmission. Also, there could be peaks later in the response in the information transmission without a corresponding rise in firing rate (A).

reached a value 20% of the way from its spontaneous rate to its peak rate.

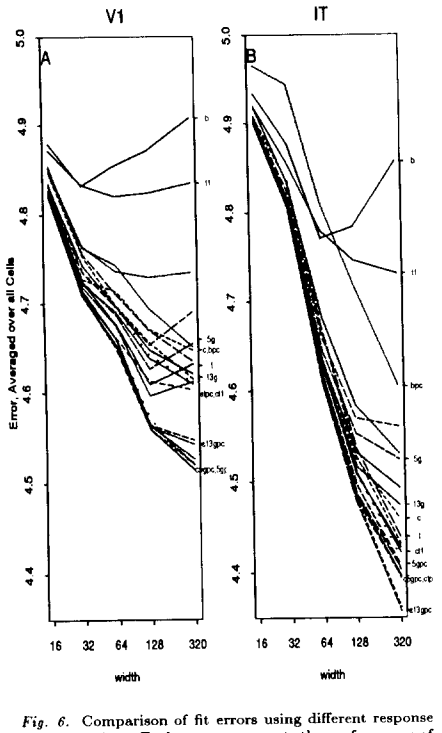

Fig. 6. Comparison of fit errors using different response representations. Each curve represents the performance of one particular representation across a number of window sizes, averaged over all cells of a particular type. The x axis indicates the width, in ms, of the window used (all windows began at the estimated information latency ℓ). The y axis shows the fit error for that representation and that window size, averaged over all cells of that type. A. Fit errors for V1 cells. The lowest error for V1 is with `c5gpc` (see Table 1), which is the spike count and the principal components of data smoothed with a $\sigma = 5$ ms Gaussian kernel. The order of the representations for the entire 320-ms response is (from best to worst): `c5gpc c3gpc 3gpc 5gpc c8gpc 8gpc c13gpc 13gpc ctpc ct1 ct3 13g tpc ct 8g t c bpc 5g 3g t3 t1 b`. B. Fit errors for IT cells. The lowest error for IT is with `c13gpc`, the count and principal components of data smoothed with a $\sigma = 13$ ms Gaussian. The order here is: `c13gpc c8gpc 13gpc 8gpc c5gpc ctpc 5gpc c3gpc ct3 3gpc ct1 ct tpc t c 13g 8g 5g t3 3g bpc t1 b`.

3. Results

3.1. Latency

The information-based latencies ℓ that we computed using the count code within a 16-ms sliding window coincided within 4 ms of the firing-rate latencies λ in only 5 V1 cells and 3 IT cells. In 8 V1 and 7 IT cells, the rise in information is delayed with respect to the earliest change in firing rate. In 2 IT cells, although there was stimulus-dependent information, there was no comparable rise in the average firing rate across all trials to all stimuli. Figure 4 shows the relationship between the firing rate and information rate latencies.

Fig. 5 shows plots of the information carried in the 16-ms-wide windows as they are slid along in time from through the onset of the response. The three panels illustrate a case where the two latencies are nearly the same, another where the information latency is substantially greater than the firing-rate latency, and a third where very little information is transmitted, despite considerable firing. These cells are typical in that changes in information rate do not necessarily correspond to changes in the firing rate averaged over all stimuli.

3.2. Comparison of representations

All of the codes (representations) were tested on a set of windows of durations 16, 32, 64, 128 and 320 ms, all of which had their early edges aligned at ℓ. Each representation was tested on each window size for each cell. For each window size in each brain region, the test error values (Fig. 6) and the information for a given representation were averaged across all cells and ranked (Fig. 7). For small window widths, the count code (`c`) did as well as any other representation. For wider windows, it usually carried a large proportion of the information conveyed in the best representation. On average, this fraction was 76% ±25% (sd) in V1 and 85%±13% in IT.

The binary code, (`b`) performed worse than all the other representations. The principal components of this code (`bpc`) performed better, but still not as well as other representations. When the binaries were convolved with truncated Gaussian kernels and resampled at a lower rate (every 4 ms, 8 ms in some studies with wider Gaussians), larger values of σ provided increasingly better codes (in both panels of Fig. 6A, the representations `13g`, `8g`, `5g`, and `3g` are ranked in that order). When the principal components of the Gaussian-filtered binaries (`3gpc`, etc.), were used differences in error

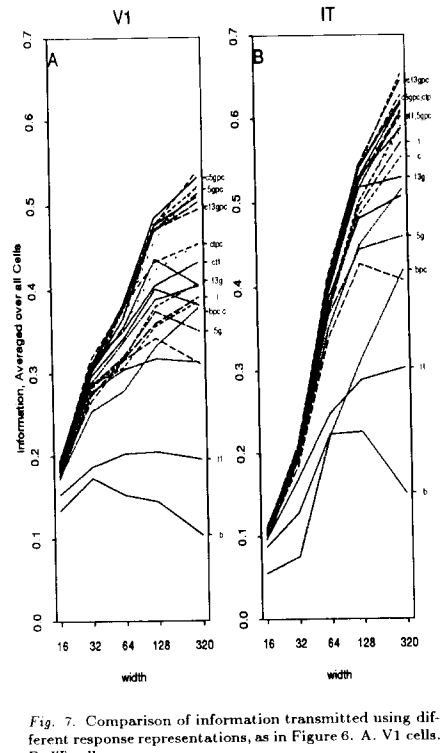

Fig. 7. Comparison of information transmitted using different response representations, as in Figure 6. A. V1 cells. B. IT cells.

due to changes in σ became much smaller (a few hundredths of a bit).

For the larger windows, the best representation was the count and principal components of the Gaussian-smoothed data, taken together. Up to 5 principal components were needed in the optimal representation for both V1 and IT cells, with 3 the most common number. A representation using only one principal component was optimal only for one V1 cell. The mean gain in information from adding these principal components to the count alone was 0.09 bits (15% of the total) in IT and 0.12 bits (24% of the total) in V1 ($p < 0.01$, Wilcoxon signed-rank test). When the principal components were employed, the width of the Gaussian smoothing kernel made less of a difference than it did when the smoothed, resampled time series were used directly, but was still significant. A wide kernel ($\sigma = 13$ ms) performed better on the full response (320 ms) IT data than a narrow one ($\sigma = 5$ ms) (p , .05, paired t-test). The narrow kernel ($\sigma = 5$ ms) performed somewhat, but not significantly (p =.09), better on the full response in V1, with an average improvement of 0.02 bit. With narrower time windows, the gap between representations narrowed: for the window of width 16, the count alone was within 0.01 bits of the fit error achieved by the best representation.

The times representation (`t`) produced lower error values than the raw binaries (`b`) did, but not as low as those obtained with the Gaussian-smoothed and resampled binaries. In an initial study, the intervals (`i`) and inverse intervals ($\mathtt{i}^{-1}$) always did worse than the times and were dropped from further analysis. In the shorter windows, the times did well. Adding the count and taking the principal components of the times helped significantly, but none of the times-based representations did as well as the optimal representation in longer windows

For a 320 ms window, the first three times (`t3`) accounted for about 80%±28% of the information recoverable from the times in general (`t`) in V1, about 90%±14% in IT. Sometimes, the first three times of occurence were more useful than all the times. In those cases, any information found in the remaining spike times was masked by noise so that the network could not recover it.

We also measured the information carried by the first spike time alone (`t1`), i.e., the single-response latency. It accounted for 35%±20% of the total information in V1 and 48%±22% in IT. The combination `ct1` of the first spike time and the spike count accounted for 94%±17% of the total information in IT, but only 84%±22% in V1.

To set confidence limits on our estimates of the fit error and information transmitted by the different representations, we used the standard errors of the means of these statistics over the four different test sets for which the fitting procedure was carried out. These were generally around 0.04 bits for V1 cells and 0.02 or smaller (because there were more trials in the experiment) for IT cells. Fig. 8

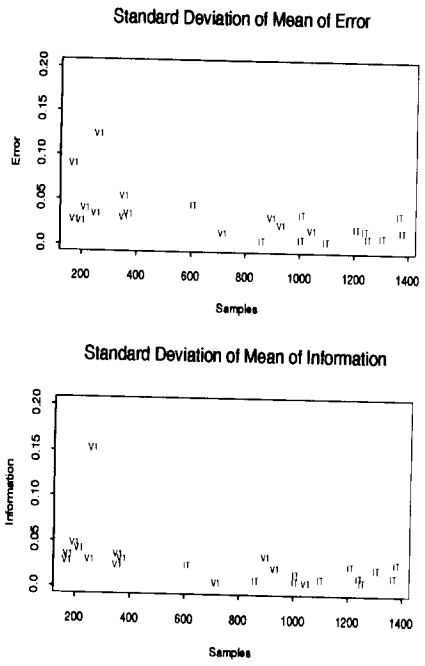

Fig. 8. Standard deviation of the mean. In running the network, each test set produces a different fit error and information estimate. The standard error of the mean increased with the amount of information detected and decreased with the number of samples. In our study, we only used one division (4 runs of the network), and in some cases, the estimated standard error of the mean was as high as 0.1 bit, though usually it was much lower.

shows these standard errors for both statistics for the optimal representations.

9.3. Instantaneous information rate

The instantaneous rate of information transmission was estimated as a function of time by sliding a window along the response, as described above under latency. However, while each window used in the latency analysis was identified by its starting point, each window used in measuring instantaneous information rate was identified by its center point, so features found using sliding windows of different widths could be compared and aligned. For each window position, we estimated the information using two different codes: the count (`c`), and the combined code (`cg5pc`) consisting of the count together with up to three principal components of the Gaussian-smeared ($\sigma = 5$ ms) data.

We first used a 16-ms wide sliding window. Combining the principal components with the count did not yield higher information rates than the count for either IT or V1 data. The same held true for a 24-ms wide sliding window. This means that one number, the spike count (`c`), is a contains all the stimulus-related information in the data on the 16-ms or 24-ms time scale. If there were additional features with smaller resolution, the PC's would have reflected that fact and carried additional information. Thus, 16 ms of a response can be used to estimate an instantaneous baud rate as the information conveyed in a time window centered at the time in question divided by the window duration. There are rises and falls in this rate, as can be seen in V1 cells in Fig. 9 and IT cells in Fig. 10. The peak baud rates observed in different cells range from about 2 to about 30 bits/s. We experimented with even shorter sliding windows (12 ms), but the fit error values were higher and more variable, reflecting the relative lack of spikes in such short time intervals.

For 32-ms sliding windows, all of the cells showed one or two periods early in the response, about the time of the peak in the instantaneous information rate, when two principal components were necessary for the optimal representation (`cg5pc`). However, the differences between the results with the principal components and those based on the spike count alone were small for most IT cells and some V1 cells. The 64-ms sliding window shows greater discrepancies between the information conveyed by the count alone and that carried by the count with the principal components added, particularly at the points of highest information (though not in all cells). In those places, 2 or even 3 principal components are needed (Fig. 9 and Fig. 10, panels C and D).

The count code in the 64-ms sliding window usually carries more information than in the 32-ms window. However, for most V1 cells, at the time of peak information rate, it actually carries less information than the count in a 32 ms sliding windows centered on the same point (Fig. 9, panels A and B). This phenomenon was not found in any of the IT cells we studied (Fig. 10, panels A and B).

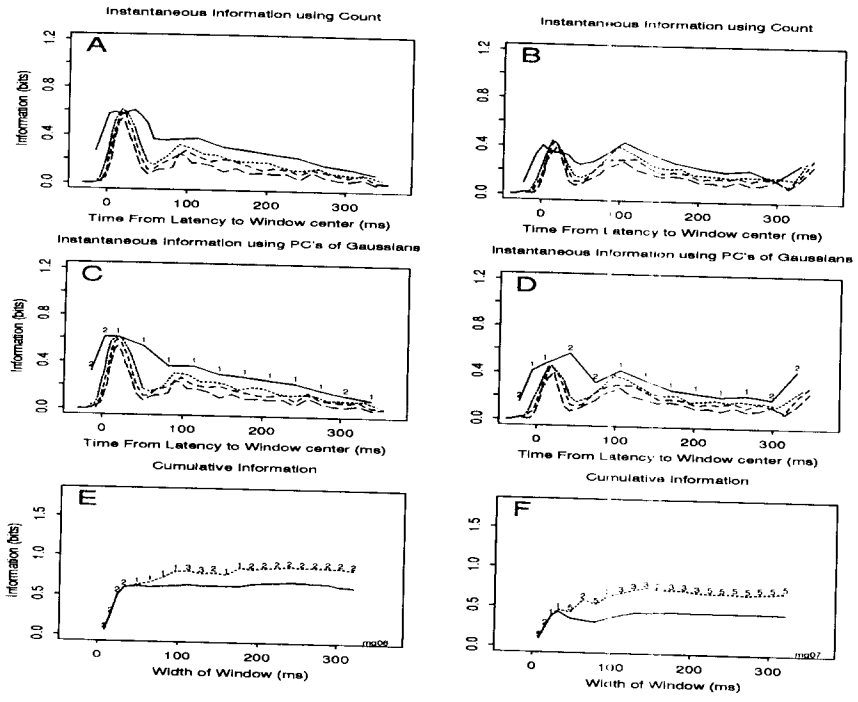

Fig. 9. Temporally-resolved information transmission for typical V1 cells. The y axes are in bits. In panels A-D, The x axis is the time at the center of the window in ms, measured from the information latency ℓ. A. and B: Information transmitted using the count representation for 4 different window sizes: 16 (long-dashed line), 24 (short-dashed), 32 (dotted) and 64 ms (solid). The x axis is the time at the center of the window in ms, measured from the information latency ℓ. Note how the information rate rises and falls with time. In most cells, there are at least two bursts of information. Widening the windows around the first burst does not increase the information carried. C and D: Information transmitted using representations including several principal components, for 4 window sizes (same sizes as in A and B). For each window, the optimal number of principal components was chosen. The numbers above the 64 ms curve indicate how many principal components were necessary for this optimal representation at each window position. Starting with the 32- and 64-ms windows, additional principal components can capture more information than the spike count (A and B) can, particularly at points where the information content is highest. E and F: Each point in the curve indicates the total information available using a particular representation if all of the data from latency until that time is used. In each plot, the solid line is the result of using the count code c. The dotted line is the result of using the count and the optimal number of principal components of Gaussian-smeared data for that length window. The optimal number of principal components is listed above some data points. The Gaussian had $\sigma = 5$ ms. Information values rise for both representations rise together in the first 30-40 ms. Then, the representations using principal components begin to outperform that using the count alone (which may actually perform worse after more than 100 ms). This happens in almost every cell.

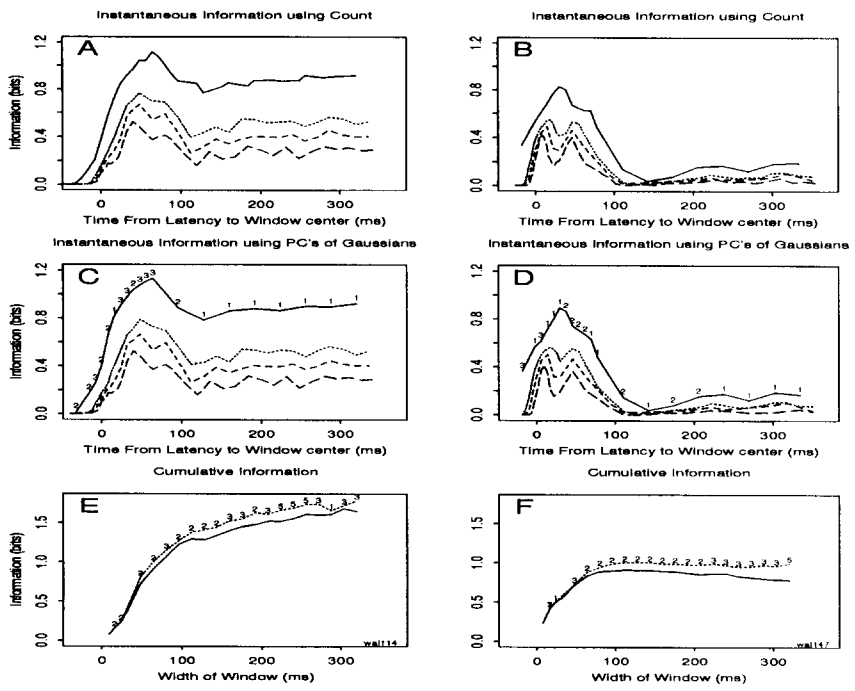

Fig. 10. Temporally-resolved information transmission for typical IT cells. See the caption for Fig. 9 for details. Differences from the V1 cases are: (1) The width of the Gaussian kernel used in the optimal representations is 13 ms. (2) Widening the windows in A - D increases the information carried. (3) The information carried by the spike count continues to grow throughout the entire response period.

3.4. Cumulative transmitted information

In determining the best representation (as described above), we measured the information carried in different time-periods of the response. These results (Fig. 8B) can be interpreted as crude cumulative information curves that answer the question: how much could be determined about the stimulus by observing the response for a specific amount of time? To answer this question more completely, we used windows that grew in relatively small increments (8-16 ms). For both V1 and IT cells we used the count alone, and the "best" representation as obtained above: the count and the principal components of the Gaussian-filtered binaries, with $\sigma = 5$ ms (`c5gpc`) for V1 and $\sigma = 13$ ms (`c13gpc`) for IT. The IT data were resampled at 8 ms, rather than 4, because a Gaussian of $\sigma = 13$ ms eliminates features of the response on the scale of 4-8 ms.

For 10 of the 13 V1 cells, the information values for the two representations rise together steeply for the first 30-40 ms. Then, the count code reaches a practical maximum, and possibly decreases in effectiveness thereafter. The combined code also rises steeply in the early part of the response, but it then either levels off or continues to rise slowly. In 9 of 11 IT cells the information transmission rates for both representations rise over the first 100 ms. At that point the PC's gain an advantage over the count, but both then continue rising more gently through the remainder of the response. Examples of these curves for V1 and IT cells are shown in Figs. 9 and 10 respectively, panels E and F of each figure. An examination of how many principal components are necessary for a maximal representation at each stage is also informative. More principal components are necessary later in the response. After about 32 ms, the average number of principal components needed for optimal representation rises to over two, and continues rising, eventually typically reaching a value of 3 for the whole response.

3.5. Independent messages

All of the cells showed a peak in transmission rate early in their responses. This peak was between 2 and 40 ms (mean 21 ms) after the latency ℓ in V1, between 6 and 56 ms (mean 29 ms) in IT. On average, the 16-ms window centered on the peak carried 58% of the total information for V1 cells and 32% for IT cells. All of the V1 cells and 9 of the IT cells showed an additional subsequent peak in transmission rate. In V1, this peak came between 52 and 232 ms after latency (mean time 101 ms). In IT, it came between 48 and 216 ms after latency (mean time 94 ms). We calculated the combined information of the two bursts using a combined code – the count code from 16 ms of the first burst and the count code from 16 ms of the second burst, to compute the combined information, c. If b_1 is the information carried in the first 16-ms period and b_2 is that carried in the second one, the fraction of b_2 that is *new* information is $(c - b_1)/b_2$. On average, in V1 this fraction is 56%±56%, and in IT it is 66%±27%. Thus, in 22/24 cells the second information bursts carry new information not present in the first ones. (We emphasize that the "bursts" we describe here have nothing to do with what is often called "bursting" firing patterns of neurons. Our bursts are simply local maxima in the instantaneous information transmission curves.) In some cases, the information available from the combined code was greater than that available from the two bursts individually.

4. Discussion

These results extend considerably our understanding of temporal coding of information about visual patterns in primates. Previously, it had been established that neurons in V1 and IT code some of this information in some way in the pattern of timing of their spikes [16], [14], [18]. Reliable calculations of the magnitude of this information are now possible[10], [11], but very little was known about the way the messages were coded in time. Our results permit us to determine the temporal resolution with which the messages transmitted by single neurons need to be read and how the information they convey varies in time in the course of the response. These findings, in turn, place strong constraints on possible codes these neurons use.

4.1. The temporal resolution limit

Characteristic temporal resolution limits of the order of 25 ms (V1) and 50 ms (IT) emerge from our results at several points. The first pieces of evidence come from the comparison of the many different representations of the response. In both V1 and IT, the code that allowed our neural network to achieve the lowest test error had the spikes smeared by a Gaussian kernel, followed by a Karhunen-Loeve transform truncated to 3 or 4 principal components. The best fit was achieved when the Gaussian had $\sigma = 5$ ms for V1 data and $\sigma = 13$ ms for IT. A Gaussian of $\sigma = 5$ ms reduces all frequencies over 27 hz by 3 db. This means that Fourier components of the responses which are smaller than 1000/27=37 ms in period are cut to 1/3 in magnitude. Thus the smoothing effectively suppresses variations on the scale of about half that - 18 ms. For IT, where the optimal Gaussian kernel has a width of 13 ms, the corresponding argument leads to a suppression of variations in the response shorter than 48 ms.

The truncation at a small number of principal components introduces additional low-pass filtering, so rapid variations in the response are suppressed even further in both these optimal representations, and we can be sure that very little information is coded near the cutoff frequencies of 27 hz (V1) or 10 Hz (IT). Nevertheless, the fact

that the narrow Gaussian kernel produced a significantly (albeit marginally) better fit than the wide one for V1 (and vice versa for IT) implies a difference in the temporal resolution scales of the messages in the two areas.

Nearly the same timescales emerge from a comparison of the performance of different representations in different time windows. The codes which are optimal for the maximal window, encompassing the entire 320 ms of the response, lose much of their advantage over other representations in the shorter time windows. In particular, it is evident in Fig. 6 that for windows of 16 ms in V1 and 32 ms or less in IT, the spike count yields a fit virtually as good as the optimal code and conveys essentially all the information it does. Thus, no improvement would occur by analyzing changes in firing rate on timescales shorter than these. On the other hand, for windows wider than these, information is lost if the temporal variation of the firing rate within the window is not taken into account. This places the temporal resolution limit between 16 and 32 ms for V1 and between 32 and 64 ms for IT.

One can see this same transition in the more detailed cumulative information curves (Figs. 9 and 10, panels E and F). At short times, the spike count and the optimal representations transmit essentially the same information, but at longer times the curves for the two codes diverge, revealing that temporal variation of the response within the window carries information. The characteristic times are once again consistent with the 25 and 50 ms identified in the two areas above.

These calculations show that at some times in the windows longer than these values, it is necessary to take into account the temporal variation of the firing rate on these timescales to extract all the information in the signal. The sliding-window calculations allow us to see at what points in the response this resolution is necessary.

We find that codes including principal components give better fits and more information during the portions of the response in which the information rate is high. For V1, in such periods, more information is transmitted by codes including several principal components (Fig. 9, panels C and D) than by the spike count alone (panels A and B), whenever the window is wider than 24 ms. However, at other times, the spike count appears to be an adequate measure of the response even over 64-ms periods. Thus, our 25-ms temporal resolution figure is relevant only to these relatively brief periods periods of high transmission rates.

A corresponding result is found in IT. Employing more than one principal component of the response generally leads to better fits and higher information for 64-ms windows only around the times of local maxima in the information rate. At other times, the characteristic temporal resolution is apparently larger.

4.2. The time course of information flow

Deeper insight into the nature of the neural code is obtained through our systematic measurements of information flow using sliding and growing time windows. The former reveal a persistent pattern in the instantaneous transmission rate, for both V1 and IT neurons. There is always an initial burst, peaking very quickly after latency (as we have defined the latter). A 16-ms window placed over the center of this peak contains, on the average, half the total information in the neuronal message in V1. In IT, the figure is lower (30%), presumably because 16 ms is too short a sampling time for these relatively slower neurons. Then, in both areas, there is usually another burst, typically about 100 ms later, carrying considerable new information. In addition, some information is carried between and after these bursts. Its magnitude varies rather irregularly in time.

The cumulative information curves tell the same story. They are almost invariably characterized by a sharp rise at the very beginning of the response, coinciding with the first peak in the sliding-window curves. Most of the eventual total information is transmitted in this first period of 30-40 ms (V1) or 100 ms (IT). The shape of the curves after this point is somewhat more variable, but the net transmitted information generally rises to a somewhat higher value (25% higher on average) during the entire 320-ms course of the response.

If all the information calculated in non-overlapping sliding window segments were independent, the integral of the sliding-window curves would match the cumulative curves. This is not

the case. A good deal of the information carried in the sliding windows after the first burst is redundant. On the other hand, some of it is new; otherwise the cumulative curves would not continue to rise.

The peak transmission rates we find in the initial bursts of some of our cells are of the same order as those measured by Eckhorn and Pöpel in the LGN of the cat [5], [6] and somewhat smaller than those recorded by Bialek *et al.* in the H1 motion-sensitive neuron in the blowfly [4]. It is an open question whether our cells would maintain such rates in response to rapid changes in stimulus. It is important to note that the H1 cells each represent a large fraction of all visual information carried by the system, while each of the V1 and IT cells we studied is one of many sub-channels that carry information in primate visual cortex, and carries only a small faction of that information.

4.3. Implications for the neural code

These results provide insight into the way these neurons code information. They tell us just what aspects of their spike timing carry information to cells that receive their signals. Of course, we have not determined whether those receiving cells make use of all the information we measure or to what extent signals from a population of cells may be combined synergistically. Nonetheless, we can place limits on the codes the brain actually uses at the single cell level. In the following, we consider different kinds of codes that have been suggested and examine the implications of our findings for them.

Abeles [1] has argued that the primate cortex may compute using very precisely-timed sequences of spikes. His group has found evidence for systematically-repeated patterns of activity, with spike timing precision of 1-2 ms, in recordings from frontal cortex[2]. These patterns appear to be related to the tasks the monkeys are performing.

Our results offer no evidence of a role for such precise timing in V1 or IT cortices. It is evident in Fig. 6 that codes such as the spike times (`t`), or combinations of them with the count (`ct, ct1, ct3`) are essentially just as effective representations of the signal as the optimal ones. However, even for the shortest windows, they never do better than the optimal ones, despite the fact that they preserve (different amounts of) exact spike timing, information that is destroyed by the low-pass filtering in the optimal ones. Thus spike timings more precise than the resolution limits we have identified do not carry any information about the patterns in these experiments.

It should be emphasized, however, that we have only analyzed information about stimuli. Recently, Lestienne [12] analyzed the same V1 data and reported evidence of repeated, precisely-timed patterns in the spike trains. Our findings imply that these patterns do not convey any information about stimuli beyond what can be extracted from a strongly low-pass filtered version of the responses. It is conceivable that the brain uses different codes for different kinds of messages, and it is even an appealing idea to use different frequency bands for information about sensory stimuli and internal processing. However, we have no evidence that bears on this conjecture.

De Ruyter van Steveninck and Bialek carried out an extensive analysis of short portions of spike trains from the H1 motion-sensitive neuron in the blowfly [19]. They found that temporal resolution of spikes on the order of 5-10 ms did convey information. Our primate neurons apparently operate with considerably less temporal precision than theirs.

Thorpe and Imbert [21] have suggested that for the visual system be able to carry out recognition tasks as quickly as it does, a lot of information must be carried in the time of the first spike. (This argument presupposes that some other neurons fire indiscriminately in response to any stimulus, to give a reference time with respect to measure these first spike times.) We find that the `t1` representation accounts for only about half the total information. This may be enough for the system to carry out the tasks in their experiments, but it seems implausible to us to conclude that the brain wastes half the information in these signals.

It has often been assumed that the neural code is simply the total spike count, taken over some interval. If we take this interval to be the entire response period (320 ms), we find this not to be a bad approximation. It misses about a quarter of

the total information in V1 and 15% in IT. However, the rest of the information we have measured is statistically significant, and a full description of the neural code for these cells must take the coding of this additional information into account.

Tovee *et al.*[22] have also carried out information measurements on IT cells. They found that when the first 120 ms of the response was removed from the interval analyzed for information about the stimuli, the net information dropped slightly. This reduced amount could be accounted for almost entirely in terms of firing rate alone. Attributing the drop when the initial part of the response was excluded to loss of information about when the spike train started, they suggested that a combination of "onset characteristics of the spike train" and the subsequent mean firing rate carried essentially all information in these cells. If the only onset characteristic we use is the first-spike latency, this combination is exactly our code `ct1`, and in our IT cells it accounted for 94% of the total information. In V1, it was not as successful, yielding 81% of the information carried by the optimal code. However, even in IT, the optimal codes consisting of the spike count plus principal components did even better than `ct1`, and we have learned, moreover, that in IT no spike timing more precise than 32 ms carries information.

The first 50 ms of the response of our IT neurons typically contain only 2 spikes, so a single spike can be said to carry a good deal of information. However, apparently it is the presence or absence of these spikes in this period, not their exact timing, that is most informative.

If "onset characteristics" is taken to mean the spike rate in the first, say, 64 or 100 ms, the code proposed by Tovee *et al.* amounts to a simple combination of two numbers: the spike count in that first interval and that in the rest of the response. In fact, this bivariate code is not far from the optimal representations we find. All our results are consistent with the hypothesis that all information about the stimulus is carried in an effective time-varying firing rate defined by averaging the spike train over a suitable time window. This window should be about 25 ms wide in V1 at times of high information transmission rate, about 50 ms wide in IT. Elsewhere in the response, longer averaging times, perhaps 100 ms or more, are adequate. The extra degrees of freedom of the response after the initial burst are necessary to give the slow rise in information seen over the entire response period for most cells. The bivariate code approximates these degrees of freedom by a single number, the post-burst spike count. The trivariate code ("early", "middle", and "late" firing rates) proposed by Miller *et al.*[13] goes one step beyond this. These are not qualitatively bad approximations, and for the 16 out of our 24 cells for which the optimal number of principal components was 3 or less, the latter is essentially equivalent to our optimal model. However, for the remaining 1/3 of our cells, our optimal representations have higher dimensionality, and they capture more information.

In V1, the extra information rise after the initial burst is apparently due to new features of the response (characteristic changes in firing frequency), which provide new, independent information. This is evident from the fact that the information carried by the spike count generally remains constant or even drops a bit after the fast initial rise period (Fig. 9, panels D and E). The extra principal components in the optimal representation are necessary to capture the change that occurs in the nature of the message.

In contrast, the information carried in the spike count alone in IT cortex generally continues to rise throughout the response period. This finds a natural explanation in a model where the information is carried in a time-independent firing probability. This probability can be estimated better if the spike train is observed for a longer time: simple arguments give an uncertainty in frequency, $\Delta f \sim 1/t$, where t is the observation time, leading to information $\propto \log t$, which is in at least qualitative agreement with the variation seen in many IT cells, including the ones in Fig. 10. However, this agreement does not mean that an underlying firing rate constitutes the code of IT neurons. For them, just as for our V1 ones, better fits are achieved and more information is transmitted using representations that include several principal components, indicating that temporal variations in the firing rate are part of the neural code.

The fact that, of the representations we tried, ones employing principal components of the response were optimal does not imply that the

cells which receive the response actually perform a principal component decomposition. However, the fact that we could not find any better ones does suggest that downstream neurons can gain more information if their processing can accommodate the features in the temporal structure of the signals that our analysis has identified as carrying information. The minimum temporal resolution we have identified for the initial burst period in V1 neurons matches well with typical cortical membrane time constants. We do not know how, or even whether, the system achieves the integration over longer periods necessary to extract maximum information from later portions of the response. However, this is a problem not just for our codes, but for any code based on firing rates averaged over more than 20 ms or so.

Another noteworthy feature of the best representation is that it includes the spike count. It might seem that the first principal component, which corresponds to the largest source of variance among the responses, would duplicate the count. Although the correlation coefficient between the count and the first principal component is high [15], [18], the first principal component represents the extent to which the response can be approximated by a specific waveform, while the count is just an average. This means that the difference between the two is the extent to which the actual waveform of the response differs from the first principal component. The count and principal components always do better than the principal components alone. This suggests that the count conveys information which is not conveyed by the first few principal components.

The importance of the spike count and firing rates averaged over times ranging from 25 ms upward suggests another simple hypothesis: that variations in information transmission are a direct consequence of temporal changes in firing rate, so high information transmission is achieved when and only when the firing rate, averaged over all stimuli, is high. As was the case for other simple hypotheses we tested above, this is not a bad approximation, and it holds approximately for a majority of our cells, but it fails for about a third of them (see Fig. 5).

We have also shown that for some cells there are two distinct latencies – one which measures the time until the firing rate rises as a result of stimulus presentation, and the other which measures the time until different stimuli can be distinguished on the basis of the cell's response. For some cells, these two latencies are simultaneous. For others, they are separated by over 10 ms. This means that some cells in the visual system start firing when a new stimulus is presented, but this change in firing rate provides no information about what the stimulus was.

Our study leaves a number of important questions unanswered. A particularly interesting one is just what patterns or features in those patterns can be discriminated on the basis of a neuron's response. In recent work, we have addressed this question with respect to the entire 320-ms response. However, our findings here about the detailed temporal course of the information transmission raises the question of whether information transmission about different kinds of spatial pattern features (for example, low and high spatial frequencies) follows different time courses. Exploration of this question would extend our knowledge of the neural code significantly.

Acknowledgements

We would like to thank Dr. Richard Kronauer of the Harvard University Division of Applied Sciences for his support and many insightful discussions. We would also like to thank Dr. Mortimer Mishkin for his support of our work and for careful and helpful criticism of the manuscript.

References

1. Abeles, M. (1991) Corticonics Cambridge University Press, Cambridge.
2. Abeles, M., Bergman, H., Margalit, E. and Vaadia, E. (1993) Spatiotemporal firing patterns in the frontal cortex of behaving monkeys. J Neurophysiol 70: 1629-38.
3. Ahmed, N. and Rao K.R. (1975) Orthogonal transforms for digital signal processing. Berlin: Springer-Verlag.
4. Bialek, W., Rieke, F. de Ruyter van Steveninck, R.R. and Warland, D. (1991) Reading a neural code. Science 252: 1854-1857.
5. Eckhorn, R. and Pöpel, B (1974) Rigorous and extended Application of information theory to the afferent visual system of the cat. I. Basic concepts. *Kybernetik* 16: 191-200

6. Eckhorn, R. and Pöpel, B. (1975) Rigorous and extended Application of information theory to the afferent visual system of the cat. I. Experimental results. *Biol Cybernetics* 17: 7-17.
7. Eskandar, E.N., Richmond, B.J. and Optican, L.M. (1992) Role of inferior temporal neurons in visual memory. I. Temporal encoding of information about visual images, recalled images, and behavioral context. *J Neurophysiol* 68: 1277-95.
8. Heller, J. (1994) Using neural networks to study information in the structure of neuronal responses in the primate visual system. Undergraduate thesis, Division of Applied Sciences, Harvard University.
9. Heller, J., Kjær, T.W., Hertz, J.A. and Richmond, B.J. Dynamics of information transmission by single neurons in the visual system. Soc. Neuroscience Abst. 20: 314, 1994.
10. Hertz, J.A., Kjær, T.W., Eskandar, E.N. and Richmond, B.J. (1992) Measuring natural neural processing with artificial neural networks. *Int J Neural Systems* 3, sup: 91-103.
11. Kjær, T.W., Hertz, J.A. and Richmond, B.J. (1994) Decoding Cortical Neuronal Signals: Network Models, Information Estimation and Spatial Tuning. J Computational Neuroscience 1: 109-139.
12. Lestienne, R. (1994) Frequency insensitive measures of temporal correlations in spike trains. Extended Abstracts Book, Dynamics of Neural Processing 68-72.
13. Miller, E.K., Li, L. and Desimone, R. (1993) Activity of neurons in anterior inferior temporal cortex during a short-term memory task. J. Neurosci. 13: 1460-1478
14. Optican, L.M. and Richmond, B.J. (1987) Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex III. Information Theoretic Analysis. J Neurophysiol . 57: 162-178.
15. Richmond, B.J., Optican, L.M., Podell, M. and Spitzer, H. (1987) Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex I. response characteristics. J Neurophysiol . 57: 132-146.
16. Richmond, B.J. and Optican, L.M. (1987) Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex II. Quantification of Response waveform. J Neurophysiol . 57: 147-161.
17. Richmond, B.J., Optican, L.M. and Spitzer, H. (1990) Temporal encoding of two-dimensional patterns by single units in primate visual cortex I. Stimulus-Response Relations. J Neurophysiol . 64: 351-369.
18. Richmond, B.J. and Optican, L.M. (1990) Temporal encoding of two-dimensional patterns by single units in primate visual cortex II. Information Transmission. J Neurophysiol . 64: 370-380.
19. de Ruyter van Steveninck, R., and Bialek, W. (1988) Real-time performance of a movement-sensitive neuron in the blowfly visual system: coding and information transfer in short spike sequences. Proc R Soc Lond B 234: 379-414.
20. Rumelhart, D.E., McClelland, J.L. and the PDP Research Group (1986) Parallel Distributed Processing. MIT Press, Cambridge, MA, US.
21. Thorpe, S. and Imbert, M. (1989) Connectionism in perpective Elsevier Science Publishers B.V.
22. Tovée, M.J., Rolls, E.T., Treves, A. and Baylis, R.P. (1993) Information encoding and the responses of single neurons in the primate temporal visual cortex. J Neurophysiol . 70: 640-654.