The Speed of Sight

C. Keysers[1], D.-K. Xiao[2], P. Földiák[2], and D.I. Perrett[2]

Abstract

■ Macaque monkeys were presented with continuous rapid serial visual presentation (RSVP) sequences of unrelated naturalistic images at rates of 14–222 msec/image, while neurons that responded selectively to complex patterns (e.g., faces) were recorded in temporal cortex. Stimulus selectivity was preserved for 65% of these neurons even at surprisingly fast presentation rates (14 msec/image or 72 images/sec). Five human subjects were asked to detect or remember images under equivalent conditions. Their performance in both tasks was above chance at all rates (14–111 msec/image). The performance of single neurons was comparable to that of humans and responded in a similar way to changes in presentation rate. The implications for the role of temporal cortex cells in perception are discussed. ■

INTRODUCTION

The mechanisms underlying biological object recognition are still poorly understood. The visual system is thought of as a chain of visual areas containing neurons with increasingly complex response properties. Single cells in the cortex of the anterior superior temporal sulcus (STSa) show remarkably selective responses to certain objects such as faces (e.g., Oram & Perrett, 1992), which make these cells well suited for participating in the recognition of the objects they selectively respond to (Logothetis, 1998). The aim of the present paper is to investigate the mechanisms involved in object recognition by pushing the visual system to its temporal limits. To achieve this, we studied visual responses to very rapid image sequences using rapid serial visual presentation (RSVP), a presentation method previously only used in behavioral studies (e.g., Chun & Potter, 1995; Potter & Levy, 1969).

A number of investigators have gained valuable insights into the mechanisms of object recognition by placing the visual system under extreme time constraints. Thorpe et al. investigated how quickly a visual scene can be categorized as containing an animal. They showed that correct motor responses can be generated as quickly as 235 msec (humans) or 190 msec (monkeys) after the scene was presented (Fabre-Thorpe, Richard, & Thorpe, 1998). The authors conclude that such short reaction times can only be the product of a feedforward brain architecture in which each neural stage generates its output based on the very first spikes arriving from the preceding stage (Thorpe, 1990).

Another line of investigation placing temporal constraints on vision uses the backward masking paradigm in which the time available to process one stimulus is limited by presenting a second stimulus (the mask) shortly after the first. Rolls, Tovee, and Panzeri (1999) and Kovács, Vogels, and Orban (1995) showed that visual masking (i.e., the reduced psychophysical detectability of a stimulus) occurs under conditions in which the responses of shape selective neurons in the temporal cortex are interrupted by the mask. This parallel between reduced neuronal responses and reduced perceptual performance in masking strengthens the putative link between individual neurons in temporal cortex and visual pattern perception (Logothetis, 1998). Unfortunately, backward masking paradigms have relatively long gaps between trials and follow a ''nothing–stimulus–mask'' sequence. This may give the stimulus a high salience in contrast to the ''nothing'' that precedes it, which might facilitate its processing. Natural viewing conditions, in which saccades often occur approximately three times per second, follow a continuous ''stimulus–stimulus– ... –stimulus'' sequence, and are, therefore, better investigated using the RSVP paradigm.

In RSVP, images are presented sequentially and continuously, with each image replacing the previous at the same location on the screen, one after another. RSVP is interesting not only because it simulates some aspects of natural saccadic scene acquisition but also because of its implication for the perception of TV transmission. TV images are presented at 25 frames/sec (PAL) or 30 frames/sec (NTSC). Videos or movies are typically composed of a sequence of related images: If a car moves through the screen, the individual frames are highly related. It is only at the transition between two scenes that the images are not related. Some TV programs (e.g.,

[1] Università di Parma, [2] University of St. Andrews

on MTV) make very frequent cuts between scenes, which raises the question of how many unrelated images can be perceived per second. Would a single frame (of a commercial for instance, i.e., 40 or 33 msec) have any impact on the visual system if embedded in an unrelated sequence?

While RSVP has been used extensively in behavioral investigations (e.g., Subramaniam, Biederman, & Madigan, 2000; Chun & Potter, 1995; Potter & Levy, 1969), it has so far not been used in conjunction with single cell recordings in the higher visual cortex, leaving open many questions regarding this powerful paradigm. What is the minimal image duration in RSVP necessary to have an impact on the higher visual cortex? How does the response of single cells in the higher visual cortex compare with the perception that occurs under similar circumstances?

Our study addresses such questions from both a neuronal and a behavioral approach. RSVP sequences composed of color photographs of faces, everyday objects familiar and unfamiliar to the subjects, and naturalistic images taken from image archives were used as stimuli and presented in the center of a computer screen. Within each sequence, all images were presented for the same duration. From one sequence to another, the presentation rate was changed to test the effect of presentation rate.

In the physiological experiments, neurons that responded selectively to complex patterns (e.g., faces; Oram & Perrett, 1992) were recorded in the STSa while the monkey fixated RSVP sequences for fruit juice reward. Each neuron was initially tested with up to 60 stimuli presented in random order at a moderate presentation rate (111 msec/image) to determine effective stimuli for that particular cell. Where reliable responses were found, 8 stimuli out of the 60 were selected (two best, two worst, and four intermediate at driving the neuron). Image sequences were then presented as permutations of these eight stimuli shown successively without interstimulus gaps at durations ranging from 14 to 222 msec/image (see Figure 1b and Table 1).

Figure 1 illustrates how the very same stimuli were used in separate psychophysical experiments to mea-

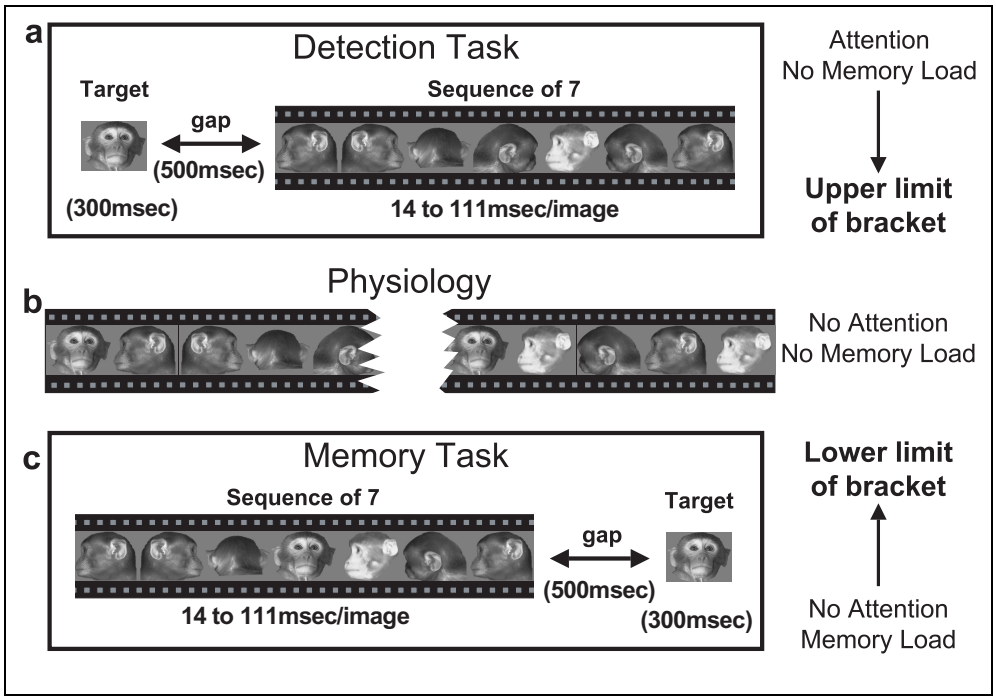


Figure 1. Experimental design. (b) In the physiological experiments, a continuous sequence of naturalistic images was presented to the monkey. For each neuron, eight stimuli were selected to range in effectiveness from that causing the strongest (best) to that causing the weakest (worst) responses for the cell. The ''movie strip'' represents schematically the RSVP sequence; the break in the strip represents the fact that in the physiology, the RSVP sequences were very long, containing up to 720 consecutive images. No report of perception was required, and, therefore, no systematic attention towards one of the stimuli was generated. The goal of the separate psychophysical experiments (a, c) was to define a bracket of performance indicating how much perception of individual images occurred in the physiological experiments. The performance in a detection (a) and a memory (c) task was used to estimate the upper and lower limit of this perception respectively. We used the 23 stimulus sets containing eight stimuli used previously in (b), and for each set, the stimulus that had been ''best'' for the cell was chosen as the ''target,'' and the other seven stimuli as ''distracters.'' In the detection task (a), each target was presented for 300 msec, followed after a 500-msec gap by a rapid test sequence of seven images either containing only the seven distracters, or containing the target in positions 3–5 surrounded by six distracters. The human subjects signaled perception of the target in the sequence by pressing a key. Trials with each of the 23 stimulus sets were intermixed and presentation rate (111, 56, 42, 28, 14 msec/image) for the test sequences varied randomly between trials. Due to the selective attention induced by presenting the target before the sequence (see text), this performance gives an upper limit to the amount of perception occurring in (b). In the memory task (c), the sequence was shown first, followed by a 500-msec gap and a 300-msec target. The subjects signaled their memory of having perceived that target in the preceding sequence by pressing a key. This memory performance was taken as the lower limit of the amount of perception occurring in (b), since the subjects could not selectively attend to the target during presentation and the stimuli had to be both perceived and memorized for a brief interval.

Table 1. Statistics of Neurone Performance as a Function of Presentation Rate

msec/ Image	*Images/sec*	*Response onset*[a]	*Response duration*	*Response–stimulus duration*	*Number of neurons*	*Percent neurons discriminating*[b]
222	4.5	108	284	62	34	97
111	9	108	168	57	34	97
56	18	108	112	56	34	94
42	24	108	86	44	23	91
28	36	108	93	65	34	79
14	72	108	71	57	23	65

The images were presented as a continuous sequence, so that the number of images per second is the inverse of the duration of each image. The duration of the population response was defined as shown in Figure 4.

[a]Time of latency alignment and average detected discrimination onset.
[b]Percent neurons with significant ANOVA testing the effect of stimulus (1–8) on neuron response ($p < .05$).

sure the perception possible within RSVP sequences. Two tasks were used (Figure 1a,c). In the detection task (Figure 1a), subjects selectively attended to a target thereby privileging its processing at the expense of other stimuli (Chelazzi, Duncan, Miller, & Desimone, 1998; Chun & Potter, 1995). This selective attention, present in the detection task but absent from the physiological test situation, makes the behavioral performance in the detection task an estimate of the upper limit of the perception possible in the RSVP sequences used in the physiological testing situation (Figure 1b). In the memory task (Figure 1c), subjects not only had to perceive the target, but also to remember it briefly. This additional requirement makes the behavioral performance in the memory task an estimate of the lower limit of the perception possible in the physiological testing situation.

Separate psychophysical assessments were made because measuring perception during the physiological recordings would have interrupted the data acquisition and underestimated perception (if a memory task had been used) or would have biased the brain in favor of a target (if a detection task had been used). The latter would overestimate the brain's unbiased processing performance and would not establish the extent to which very brief stimuli can be processed without selective attention.

Finally, the performance of single cells and human subjects was compared using corresponding signal detection analyses to investigate the extent to which the psychophysical decision of humans could be based upon the responses of neurons such as those recorded in the macaque STSa.

RESULTS AND DISCUSSION

Single Neurons Show Selective Responses at 14 msec/Image

One hundred thirty-seven neurons were tested. Out of 137, 103 were not fully tested, because none of the 60 stimuli tested evoked robust responses (~70% of cases) or because an effective stimulus was found, but the cell could not be recorded for long enough to complete testing (~30%). Thirty-four neurons were recorded long enough to complete testing at presentation rates of 222, 111, 56, and 28 msec/image, and 23 of these neurons were tested additionally at 42 and 14 msec/image. In RSVP sequences, cells respond not to a single stimulus in isolation but to a continuous sequence composed of many stimuli. To measure the response to a particular stimulus in the sequence, we created peristimulus rastergrams (e.g., Figure 2) by realigning the continuous recording at the time of each occurrence of the stimulus in the sequence. The average spike density functions (e.g., Figures 2 and 3) then reflect the systematic response to the stimulus of alignment surrounded by activity evoked by all stimuli. Where the phrase ''response to a stimulus'' is used below, it refers to the systematic response extracted according to this method.

Figure 2 depicts the aligned responses of one neuron to the occurrences of face and half-profile views of a monkey in the sequence. These stimuli caused the largest and second largest response, respectively, from those tested for the neuron in the 111-msec/image reference condition (and, hence, are defined as the ''best'' and ''second best'' stimulus for the neuron). This neuron was sharply tuned within the stimulus set to the ''best'' stimulus: Its response to the second-best stimulus was barely detectable. Other neurons had broader tuning. The responses of this neuron occurred at a relatively constant time interval (~90 msec) after onset of the ''best'' face stimulus and lasted longer than the stimulus itself. Responses to the ''best'' stimulus are evident on the majority of trials, especially for rates of 222–42 msec/image. A stimulus-differentiating response is apparent, even at the faster rates (28–14 msec), as a small peak in the spike density function for the ''best'' stimulus compared to that for the second-best stimulus. The neuron illustrated in Figure 2 and others responded

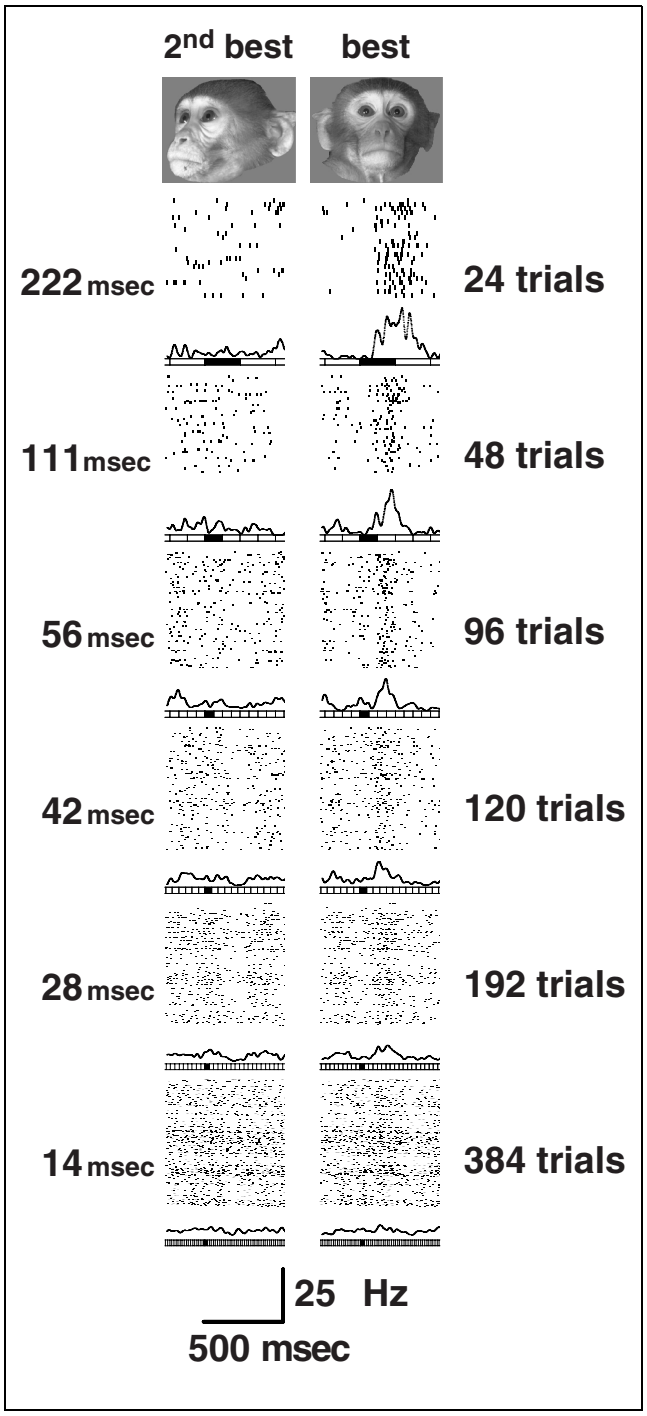


Figure 2. Responses of one neuron to the ''best'' (face) and ''second best'' (half profile) images at the different presentation rates. For each rate, rastergrams record activity during image sequences aligned for each occurrence of the face or profile images. Spike density function (SDF, smoothed using a Gaussian with σ = 10 msec) combine the activity across all occurrences of the image (trial numbers are indicated on the right of each rastergram). The black horizontal rectangles represent the time of presentation of the stimulus of alignment, the unfilled rectangles the timing of the other randomized stimuli in the sequence.

in a consistent manner, time-locked to the onset of the ''best'' stimulus.

The average firing rate was typically higher in the more rapid presentations, because effective stimuli (e.g., ''best'') are shown more often and closer to one another in time. Despite this elevation in average firing rate, at 14 msec/image the neuron illustrated in Figure 2 did not show the flat spike density function that would have occurred if presentation at this extremely high rate had caused the eight overlapping images to fuse completely into a single compound image. Levick and Sacks (1970) demonstrated that if two brief (2 msec) light flashes are presented one after another, at interstimulus intervals (ISIs) longer than 60 msec, the two flashes produce completely separate responses in the ganglion cells of the cat's retina, lasting for ~60 msec each. If ISI is reduced below 60 msec, the two responses start to overlap; and for ISI smaller than 24 msec, the joined response to the two individual flashes becomes indistinguishable from the response to a single flash of equal energy lasting throughout the ISI. If consecutive flashes integrate in the retina, it is hard to conceive how they would not integrate in later stages of processing, such as the STSa. It is, therefore, likely that some form of integration between neighboring stimuli in the RSVP sequence occurred in our experiment within a time window of ~60 msec. This integration may account for some of the response reduction, particularly at high presentation rates.

The classical concept of a prestimulus baseline is problematic in RSVP. In other paradigms such as masking, the activity occurring before the stimulus presentation reflects the spontaneous activity level of the neuron. In our RSVP testing, because each stimulus is preceded by other randomly selected stimuli, the activity occurring before a stimulus of alignment is not spontaneous activity, but the average response to all eight stimuli used in the sequence (including, in 1/8 of the cases, the ''best'' stimulus for the cell). Testing the response to a stimulus against this prestimulus activity underestimates the magnitude of the response. Selectivity of cells was, therefore, analyzed (see below) by contrasting the responses to different stimuli.

Indeed, there are two positions in the sequences in which the baseline is less elevated: directly before and directly after the ''best'' stimulus. This is because a stimulus was never followed by itself in the sequence, to avoid contaminating responses at a given presentation rate by cases of ''double presentation,'' which would in effect create responses to the stimulus at half the intended presentation rate. This requirement resulted in a short period of low neuronal activity directly preceding and following the response to the ''best'' stimulus. This effect may be supplemented by a rebound period of low activity following the intense firing to the ''best'' stimulus.

If more than eight stimuli are used in RSVP, the responses to the most effective stimulus will contrast

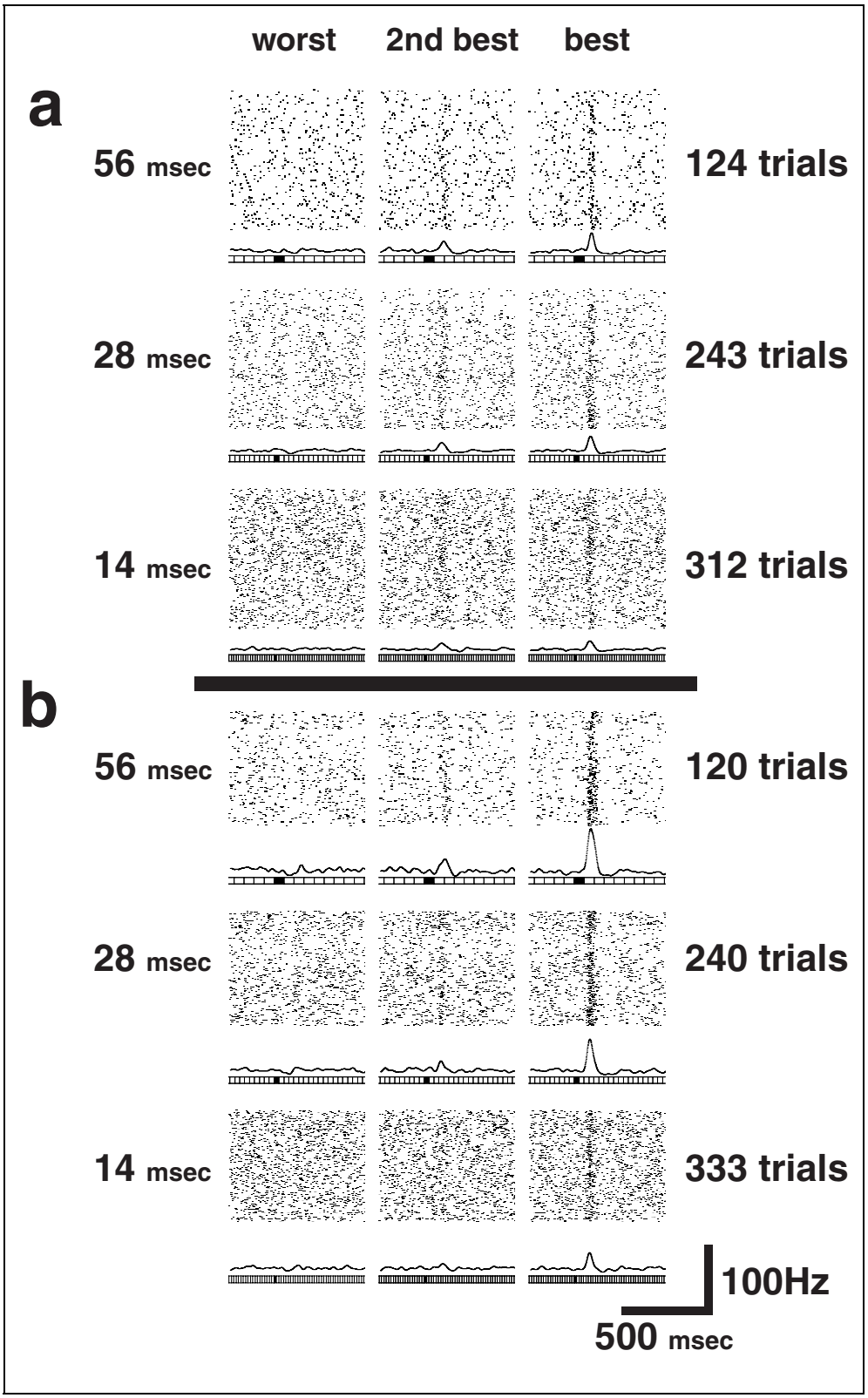
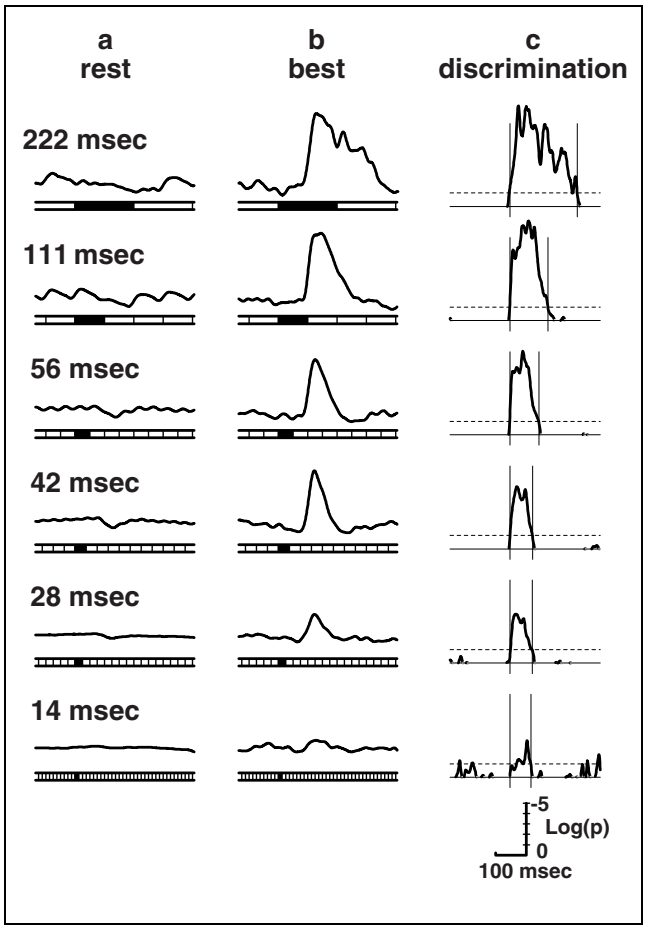


Figure 3. Responses of two cells at different presentation rates of 30 stimuli. Using the same conventions as Figure 2, the responses at presentation rates of 14, 28, and 56 msec/image are displayed for the "best," "second best," and "worst" stimulus. These two cells were tested with 30 rather than 8 stimuli, resulting in a more pronounced response and a less elevated average firing rate in the rapid presentation conditions due to the less frequent appearance of the most effective stimulus. All SDFs have the same scale indicated in the bottom right corner of the figure.

more against the activity in the period prior to this stimulus. In a separate set of experiments, 21 neurons were recorded with RSVP sequences involving 30 rather than 8 stimuli. While these cells will not be included elsewhere in this paper, they show that using more stimuli results in clearer responses after alignment because the effective stimuli occur less often in the period surrounding the stimulus of alignment (1/30 vs. 1/8). The responses of two such cells are shown in Figure 3. They illustrate that a clear response can occur to 14-msec stimuli embedded in a continuous sequence of 30 unrelated images.

The Population of Neurons Discriminates at 14 msec/Image

A pattern similar to Figure 2 was evident when the neurons that had been tested individually were considered as a population (Figure 4). Response latency varied between neurons with a mean of 108 msec (range 56–171 msec). Figure 4a,b depicts the average responses of all neurons to their best and to the other stimuli (''rest'' = average of all stimuli except the ''best'') after the responses have been aligned to the same latency (108 msec, see Methods). Increasing the presentation rate resulted in a progressive decrease of both the peak firing rate and response duration. Nevertheless, a selective population response remains apparent certainly at 28 msec/image, and to some extent even at the fastest presentation rate. Responses were visible as a peak to the ''best'' stimulus, and a dip to the ''rest'' stimuli. The dip to the ''rest'' stimuli represents the average response to ''rest'' stimuli alone (excluding the ''best'' stimulus), which is necessarily less than elsewhere in the sequence where the ''best'' stimulus occurred 1/8 of the time. Similarly, the small dips observable on either side of the average response to the ''best'' stimulus are due to the fact that directly next to the ''best'' stimulus, the ''best'' stimulus never occurs.



While Figure 4a and b displays the time-course of the average neuron's responses to the ''best'' and to the ''rest'' stimuli, Figure 4c shows how significantly these responses discriminate between ''best'' and ''rest'' as a function of time and presentation rate. Discrimination arises on average at 108 msec. This means that neurons discriminate between stimuli as soon as they start responding (Rolls et al., 1999; Ringach, Hawken, & Shapley, 1997; Celebrini, Thorpe, Trotter, & Imbert, 1993; Oram & Perrett, 1992; Thorpe, 1990; Thorpe, Fize, & Marlot, 1996). The duration of the discrimination exceeded stimulus duration by ~60 msec for all presentation rates (Table 1, columns 4 and 5) displaying a neuronal form of ultrashort memory. As mentioned earlier, Levick and Sacks (1970) demonstrated that in ganglion cells of the retina of the cat, responses to brief flashes of light also outlast stimulus duration by ~60 msec. This raises the possibility that, throughout the visual system, responses to a stimulus outlast the stimulus' duration by ~60 msec, if the stimulus is followed by a mask, and by even longer durations if the stimulus is not followed by a mask (Rolls et al., 1999). By transforming a 14-msec stimulus into a 71-msec response, the

Figure 4. Normalized average neuron responses for effective and ineffective stimuli as a function of presentation rate. Normalization was performed by dividing all SDF values of a given cell by a single normalization factor. The normalization factor reflected the peak firing of each neuron in the 111-msec/image condition, and was the 1-msec bin with the highest spike count from the PSTH to the ''best'' stimulus at 111 msec/image. For each test rate, the responses to the ''best'' (b) and the ''rest'' (a) of the stimuli (defined at the reference test rate of 111 msec/image) were latency aligned (see Methods), normalized, averaged over neurons, and smoothed (Gaussian with σ = 10 msec). See Table 1 for number of neurons per condition. The black horizontal rectangles represent the time of presentation of the stimulus of alignment, the unfilled rectangles the timing of the other randomized stimuli in the sequence. The vertical scale is identical for all presentation rates. (c) Probability of discrimination between stimuli as a function of time. For each neuron, and presentation rate, a latency aligned SDF was calculated using a Gaussian (with σ = 5 msec) for the ''best'' stimulus, and, separately, an equivalent function was calculated for the ''rest'' (remaining seven stimuli together). For each presentation rate, at the population level, the SDFs for ''best'' and ''rest'' were compared using one entry for each neuron in a sliding matched pair t test performed separately for each millisecond. No probability was computed when the response to the ''best'' stimulus was less than the response to the ''rest.'' Discrimination onset and offset were defined as the first millisecond where 30 consecutive 1-msec bins have t test, $p <$ .05 and $p >$.05, respectively. The time of discrimination onset is relatively independent of presentation rate: 104 (222 msec/image), 106 (111 msec/image), 106 (56 msec/image), 107 (42 msec/image), 109 (28 msec/image), and 118 msec (14 msec/image). The small increase in detected discrimination onset latency as presentation rate is increased is probably due to the decrease in response magnitude. The average discrimination onset detected across presentation rate is 108.3 msec (also the time of latency alignment) and is taken as the onset for the ''time window for response analysis'' for all test rates. The vertical lines represent the beginning and end of this window, the dashed horizontal line the p = .05 criterion. Note that the p values are not Bonferroni corrected and are used only to determine the ''time window for response analysis'' while the significance of the population response is assessed separately.

visual system is released from some of the time constraints of the outside world. This provides extra time for signal processing and might be a key to the brain's ability to handle very brief stimuli.

The fact that single cells respond ~60 msec longer than stimulus duration is likely to create a temporal overlap between the neuronal representations of successive stimuli in the sequence: One cell will still be responding to an old stimulus, while another cell will already respond to a new stimulus. In binocular rivalry, two stimuli are also represented at the same time in early visual cortex (see Logothetis, 1998 for an excellent review). In the latter case, the neural representations of stimuli have been shown to compete strongly against each other. Indeed, strong rivalry between stimuli is not restricted to cases in which the competing stimuli enter through different eyes (e.g., Andrews & Purves, 1997) and, even in binocular rivalry, competition occurs between stimuli and not between ocular channels (Logothetis, Leopold, & Sheinberg, 1996). The neural mechanisms responsible for binocular rivalry may, therefore, not be restricted to the case of dichoptic stimulation. Through the same neuronal competition mechanisms, the temporally overlapping neural stimulus representations in RSVP sequences may also lead the stimulus representations to compete against each other. Since response duration outlasted stimulus duration by a constant ~60 msec, the neuronal representations of stimuli would overlap most at the most rapid presentation rates. Competition would, thus, predict response magnitude to decrease with increasing presentation rate due to increasing overlap between the stimulus representations, which is what was observed in our experiments. RSVP, binocular rivalry, masking, and possibly even attention (Chelazzi et al., 1998), while all having their own specific properties, may, thus, be part of a general class of phenomena in which the neuronal representations of stimuli compete against each other.

Following methods of Sáry, Vogels, and Orban (1993), the population's capacity to signal the presence of specific stimuli at fast rates was assessed. An average tuning curve (Figure 5) was computed based on the activity of the population in the entire response duration at each rate (Table 1, columns 3 and 4). The stimulus that evoked the largest response at 111 msec/image also produced the largest response at the higher presentation rates. Indeed, the rank order of stimulus effectiveness remains relatively stable across test speed indicating a preserved stimulus coding ability.

Sixty-five Percent of the Individual Cells Discriminate Significantly Between Visual Stimuli at 14 msec/Image

To examine the extent to which single neurons exhibit a preserved stimulus coding ability, neuron by neuron

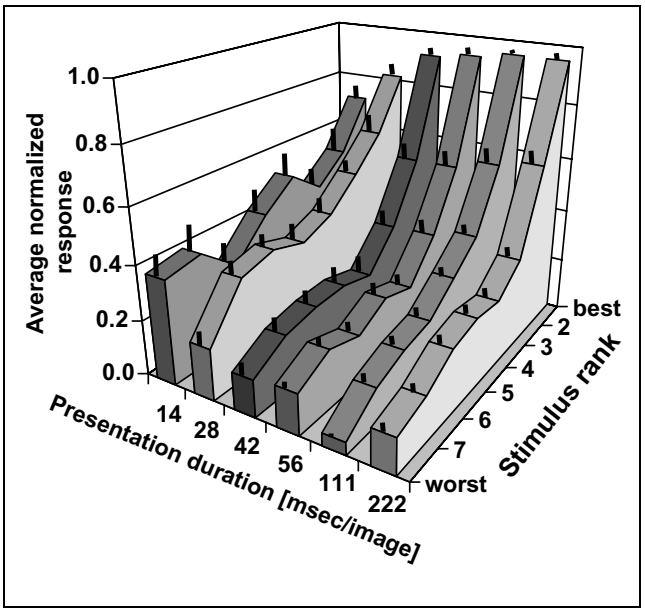


Figure 5. Response as a function of presentation speed and stimulus rank order. The stimulus ranking for each neuron was defined in the 111-msec/image condition. The neurons' responses to these same stimuli were assessed from the number of spikes occurring in the response windows starting at the neurons' latencies and lasting for the durations defined in Table 1, column 4. These responses were normalized per neuron to range from 0 to 1 at each presentation rate, and then averaged across available neurons. The black bars represent the standard error of the mean (*SEM*). (Sáry et al., 1993).

statistical analyses were performed for each occurrence of the stimuli, comparing the response to the different stimuli with one spike-count entry for each stimulus occurrence. ANOVA performed on the spike counts of each neuron during the response duration window (see Table 1, column 4) indicated a significant ($p < .05$) stimulus discrimination for the majority of the neurons at each testing rate (Table 1, rightmost column): 65% showed significant stimulus discrimination at the highest rate of 14 msec/image. Comparing spike counts during the response window to the "best" stimulus against spike counts taken during the albeit inflated prestimulus baseline yields similar results.

The stimulus' discriminatory ability emerges suddenly and early in the response. The rapid emergence is apparent in the population response (Figure 4c) and also at the level of single neurons: For each neuron, stimulus discrimination was tested immediately before and after the neuron's response onset latency. Half the neurons (17/34) discriminated between stimuli in the 20-msec interval after but not before the onset latency (*t* test, uncontaminated (see Methods) spike counts across all presentation rates, "best" > "rest"; $p < .05$). For 10-msec intervals, 38% (13/34) had this property. Thirty percent of neurons showed rapidly emerging discrimination even in the 14-msec/image condition alone (20 msec analysis). Stimulus discrimination can, thus, arise within 10–20 msec of response onset (Rolls et al., 1999; Ringach et al., 1997; Celebrini et al., 1993;

Oram & Perrett, 1992; Thorpe, 1990; Thorpe et al., 1996) for a substantial proportion of STSa neurons during RSVP.

A modest, ~30% decrease in firing rate to the "best" stimuli was observed overall between the first and last third of testing for each cell—a period of typically 1 h. Changes in response rate did not vary with presentation rates. This diminution of measured responses may reflect habituation or other factors such as the deterioration of the cell or the recording.

Implications of Very Rapid Neuronal Processing

These findings demonstrate that neuron selectivity can be measured using a RSVP method and that selectivity is preserved at high presentation rates for many STSa neurons. This opens the prospect of confronting neurons with thousands of stimuli rather than the more restricted and potentially biased stimulus sets used currently in visual neurophysiology.

The abrupt emergence of stimulus discrimination at response onset (Figure 4c) indicates that some information about the stimuli is present within the initial volley of inputs to the STSa (Rolls et al., 1999; Ringach et al., 1997; Celebrini et al., 1993; Oram & Perrett, 1992; Thorpe, 1990; Thorpe et al., 1996). This conclusion is also suggested by the capacity of individual STSa neurons to discriminate stimuli at the highest presentation rate (14 msec/image) as soon as responses commence. At this presentation rate, early stages of visual processing will have very little time to pass on information (e.g., in 14 msec, individual neurons can typically produce only 0 or 1 spike; even if cells fire for ~60 msec longer than the stimulus duration, only a few spikes can be transmitted). Under these conditions, a postsynaptic neuron cannot base its initial output firing on the firing rate of single presynaptic neurons because the output firing must be generated before a firing rate can be assessed. Together, these new findings suggest that the neural code utilized by the visual system can rely on population rate, i.e., how many neurons of a given type fire rather than how much one neuron fires (Perrett, Oram, & Wachsmuth, 1998; Földiák & Young, 1995), in short time epochs (Rolls et al., 1999; Fabre-Thorpe et al., 1998; Thorpe, 1990; Thorpe et al., 1996). Our results are consistent with feedforward processing (Thorpe, 1990) and place constraints on the type of feedback loops (between STSa to lower processing stages) involved in the stimulus discrimination within such rapid sequences.

Although the neurons at the fastest rate convey information about stimuli, the quality of the neuronal representation decreases as presentation rate increases. This effect is similar to the neural findings in masking experiments (Rolls et al., 1999; Thompson & Schall, 1999; Morris, Ohman, & Dolan, 1998; Kovács et al., 1995) and is evident in the decrease in response magnitude and discrimination probability (Figures 2, 3, and 4), the flattening of neuron tuning (Figure 5) and the reduction in number of neurons discriminating (Table 1, rightmost column) with higher test rates. The reduction in response quality at fast rates could reflect interference or competition (Chelazzi et al., 1998; Logothetis, 1998; Logothetis et al., 1996) between stimuli or the decrease in stimulus duration alone.

RSVP is a greater challenge to the visual system than the masking paradigms used in studies of higher visual cortex. In backward masking, the target stimulus is presented after a period without complex visual patterns. This may place the subject in an attentive state that may facilitate the processing of the first image. In RSVP, where stimuli are presented continuously, no such primacy exists: The processing of each stimulus suffers from both forward and backward masking from multiple stimuli. Indeed, combining forward and backward masking is known to have unpredictable effects on psychophysical performance that are in excess of the sum of the effects of forward and backward masking alone (Uttal, 1969). The present study, therefore, provides the first neuronal evidence for the higher visual system's capacity to deal with brief stimuli embedded in continuous streams of unrelated images.

Indeed, at the fastest presentation rate (14 msec/image), the visual system has to deal with the fact that during the time separating the presentation of a stimulus and the beginning of the response in STSa (~108 msec) more than seven stimuli have been presented to the eye:

$$\frac{\text{latency}}{\text{image duration}} = \frac{108 \text{ msec}}{14 \text{ msec/image}} = 7.8 \text{ images}$$

Although RSVP is undoubtedly more challenging for the brain than natural viewing conditions in which a succession of images follows rules (e.g., a rotating head), RSVP does give us insight into the upper limit to the brain's processing capacity and guides our understanding of the visual system by excluding models unable to account for such rapid serial identification.

The significant neuronal stimulus discrimination at 14 msec/image raises the question of what—if any—perceptual experience occurs in the RSVP?

Humans Can Detect and Remember Stimuli in Very Rapid Sequences

As in other studies (Rolls et al., 1999; Macknik & Livingstone, 1998), human subjects were used to examine how well single images are perceived under similar RSVP conditions. Both the ability to detect a target image in a subsequently presented RSVP sequence (detection, Figure 1a) and the ability to remember if a target had been present in a previously presented RSVP sequence (memory, Figure 1c) were tested. Subramaniam et al. (2000) did not obtain above

chance explicit or implicit memory for images presented at 126 msec/image in RSVP sequences, while Potter and Levy (1969) reported very modest recognition memory performance at 125 msec/image (their fastest rate). By contrast, Figure 6a shows that in our study, both the memory (Figure 6a, dashed black line) and the detection performance (Figure 6a, solid black line) were above chance at all presentation rates ($p < .05$, binomial test of proportion of correct trials across all five subjects and 23 stimulus sets compared to chance, i.e., 50%), indicating that human perception and memory can occur at surprisingly fast rates (14 msec/image) for stimuli in RSVP.

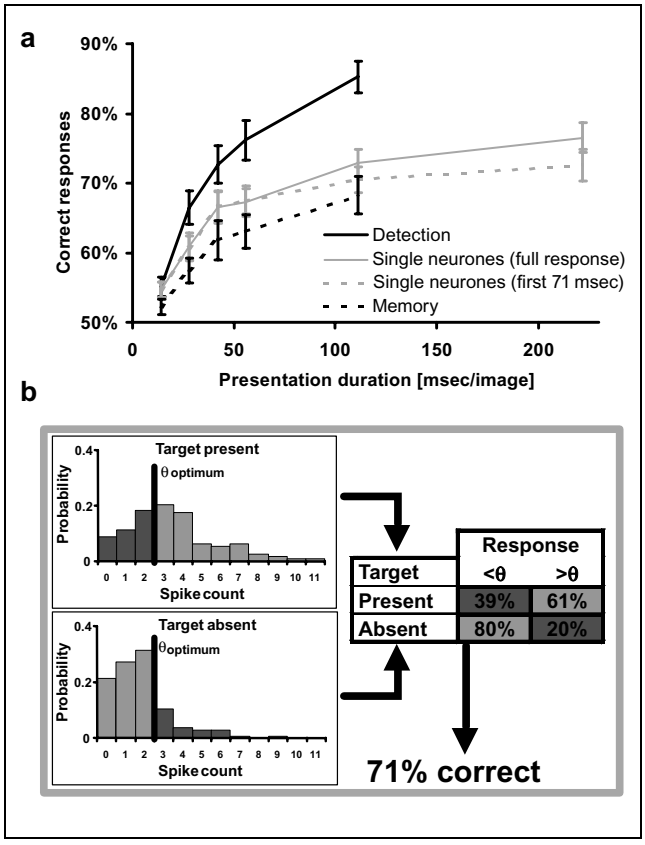


Figure 6. Psychometric and neurometric performance as a function of presentation rate. (a) The average (±*SEM*) human psychophysical performance in the memory (dashed black line) and the detection task (solid black line) as a function of presentation rate together with the neurometric performance (gray lines) calculated as shown in (b). (b) Spike count histograms for uncontaminated (see Methods) target present and target absent subsections of the single neuron record. The spikes were counted starting at the cell response latency either for the response duration defined in Table 1 (solid gray line) or for the same 71-msec duration in all conditions (dashed gray line). The similarity of the two gray lines indicates that most of the information is contained in the initial part of the response. Trial by trial spike counts are displayed for one neuron at 14 msec/image in histogram form. A signal detection table was computed by choosing a threshold θ and responding "target present" if spike count > θ and "target absent" if spike count < θ, and combining the data from both histograms (arrows). In both the table and the histogram, dark gray indicates errors (miss and false alarm) and light gray represents correct decisions (hit and correct rejection). The table expresses the percent correct and incorrect decisions for target present and target absent trials respectively (i.e., each row adds to 100%). This table was then used to compute percent correct decisions if target absent and target present were equiprobable. The procedure was repeated for all possible values of θ, and that yielding the highest performance ($\theta_{optimum}$) was used to yield the final result.

The experiments of Subramaniam et al. (2000) differed in a number of important aspects from the present investigation. Subramaniam et al. used line drawings, while we used photographs. In addition, in their Experiment III, Subramaniam et al. used 92 different stimuli per sequence, while we used only seven different stimuli in our psychophysical tasks. Indeed, when Subramaniam et al. used 32-item sequences, their recognition performance was at 58% correct, close to significance. In the work of Subramaniam et al., subjects did not expect a recognition test to be administered after the presentation of the sequence and, thus, made no efforts to actively memorize the stimuli. On the other hand, in our experiment, subjects were aware of the nature of the task and had only seven items to remember, enabling subjects to use active working memory. All of these differences may explain why Subramaniam et al. did not find above chance recognition at RSVP presentation rates at which we did.

Human Performance is Similar to the Performance of Single Neurons

The visual systems of macaque monkeys and humans are similar. The two species show very similar performance in rapid categorization (Fabre-Thorpe et al., 1998) and masking tasks (Bridgeman, 1980). Indeed, Kovács et al. (1995) showed that when a stimulus is followed by a mask, and the stimulus is to be identified, the performance of humans and rhesus monkeys is within 1–4% of each other. Assuming that the human perceptual performance is, therefore, a fair indicator of monkey perceptual performance, we can compare the single cell's performance in rhesus monkeys with the psychophysical performance in humans to investigate the relation between single cell responses and behavioral report of perceptual experience. Sequences from the neurophysiology that were comparable to those used in the psychophysics were identified (see Methods), and spike counts for the occurrence of target present and target absent sequences were computed (Figure 6b). Using optimal thresholds, the neurometric performance (Newsome, Britten, Movshon, & Shalden, 1989) that would be achieved using the output from each neuron separately was computed, and the average across neurons is displayed as a function of presentation rate (Figure 6a, gray lines).

As can be seen, the performance of single neurons is affected by presentation rate in a way similar to human psychophysical performance. This is true when the

entire response period is considered at each presentation rate (solid line) and also when the shortest response assessment period (71 msec) is used for calculations at all rates (dashed line). The reduction in neurometric performance with increasing presentation rate is, thus, not due to measuring spike counts in increasingly shorter response windows. In addition, the absolute performance of a single neuron falls on average between the two psychophysical tasks (see also Figure 1), and is closer to the memory task. STSa neurons do, therefore, carry signals appropriate for supporting psychophysical performance. This conclusion remains unchanged if we take into account that the psychophysical performance of monkeys (Kovács et al., 1995) can be somewhat (up to 4%) less than human performance when a stimulus is followed by a mask.

We found that responses last for ~71 msec for image presentation rates of 14 msec/image. This finding constrains the type of neural coding used by the visual system to encode stimulus identity. At 14 msec/image, we observed above chance human perceptual performance. If this performance is based on a temporal code in the neural response, then that code cannot rely on responses that last more than 71 msec. As seen in Figure 4c, the initial components of the responses are very similar despite changes in presentation rate. These early components could, thus, still convey some form of temporally coded information.

Together, these findings demonstrate that both single cells in STSa and human observers can detect complex patterns displayed very briefly and embedded in remarkably fast sequences of unrelated images. While at 28 msec/image, the present study reveals very clear responses, at 14 msec/image, the responses are more modest, although they remain significant for a majority of cells and clearly visible for some (e.g., Figures 2 and 3). If one was to doubt that the modest responses at 14 msec/image were of significance for the system, the robust responses at 28 msec/image still indicate that single frames can be recognized at presentation rates higher than both NTSC (33 msec/image) and PAL (40 msec/image) TV even when unrelated images are used. In addition, we have shown that some single cells in the macaque STSa are as ''smart'' as humans: They provide as much information about the presence of a brief stimulus as is evident from the behavior of observers in the memory task. It is, therefore, possible that the firing of very few cells in STSa might be sufficient to allow us to recognize an object that is presented as briefly as a single TV frame.

METHODS

Physiological Subjects and Recording Techniques

Awake subjects (two male *Macaca mulatta*, age 4–6 years) were seated in a primate chair and head restrained. Neural signals were recorded with standard methods (Oram & Perrett, 1992). Neurons were localized to the upper and lower banks of the STSa (12–18 mm anterior to the interaural plane) on the basis of x-ray visualization of microelectrodes. In one subject, recording sites were confirmed through MRI and histology with markers placed at the site of neuron recording (MRI: Magnavist, Schering Health Care, Burgess Hill, UK; Histology: microlesions and DiI [as used in Snodderly & Gur, 1995], Molecular Probes, Europe). The subjects' eye positions were monitored (accuracy $\pm 1^\circ$; IView, SMI, Germany). A 486 PC and Cambridge Electronics CED 1401 interface recorded eye position, spike arrival times and measured stimulus onset times.

Stimulus Presentation

Stimuli (256 × 320 pixels) were presented centrally on a Sony GDM-20D11 monitor (72 Hz refresh rate, image size: $10^\circ \times 12.5^\circ$, attached to an Indigo2 Silicon Graphics workstation). Onset and duration of the stimuli were measured using light sensitive diodes on the monitor screen. If the measured stimulus duration differed from the intended duration by more than 1 msec, the data for that stimulus sequence were discarded. Sequence presentation commenced when the subject's gaze remained within a fixation window $\pm 5^\circ$ of the monitor center for >500 msec and terminated after 10 sec or earlier if the subject's gaze moved outside the fixation window. Fixation was rewarded with fruit juice delivery. Activity relating to the first and last image of each sequence was discarded. The number of stimulus repetitions at a given presentation rate was adjusted to equate the total presentation time at each rate. On the average, testing involved 196 repetitions for each stimulus at 28 msec/image (range 108–336) and a total recording duration of approximately 1 hr per neuron.

Response Analysis

To measure the response to a particular stimulus in the sequence, we created peristimulus rastergrams by realigning the continuous recording on the time of each occurrence of the stimulus in the sequence. The average spike density functions then reflect the systematic response to the stimulus of alignment surrounded by activity evoked by all stimuli.

The responses of each neuron to each of the eight stimuli during the 111-msec/image testing rate were measured in a time window starting at 100 msec post-stimulus onset and lasting 111 msec. The stimuli eliciting the largest and smallest responses were defined as the neuron's ''best'' and ''worst'' stimuli. Response onset latency of a given neuron was computed off-line from trials for the ''best'' stimulus pooled across all presentation rates. The latency of response onset was defined as the first 1 msec time bin at which the firing rate exceeded the mean + 2.58 *SD* (i.e., $p < .005$) of activity measured

in a control period 250 msec before stimulus onset, for at least 25 consecutive bins. Latency aligned responses refer to responses time-shifted by the difference between an individual neuron's response onset latency and the population average (108 msec). For each neuron, a ''time window for response analysis'' was defined as commencing at the neuron's response onset latency and lasting for the duration of population discrimination at that presentation rate (Table 1, column 4).

Uncontaminated Responses

Response to a stimulus X is considered ''uncontaminated'' if it was flanked by a sufficient number of consecutive ''rest'' stimuli (R) to ensure that responses to nearby ''best'' stimuli did not contaminate the window of analysis for X. Contamination arises because response duration exceeds stimulus duration. The exact criterion therefore depended on presentation rate: RRXRR (for 111–222 msec), RRRXRRR (42–56 msec), RRRXRRRR (14–28 msec).

Neurometrics

For each of the 23 neurons tested at 14 msec/image, spike counts were obtained for stimulus sequences that were comparable to the target present and target absent sequences in the psychophysics (see Figure 1). ''Target present'' and ''target absent'' spike counts were obtained by selecting uncontaminated responses to targets (i.e., the ''best'' stimuli) and distracters (i.e., any but the ''best'' stimulus), respectively. This selection was required because in the psychophysics tasks there was at most one target stimulus per sequence. Spike count distributions for the target present and target absent sequences were calculated for each presentation rate separately. Figure 6b shows examples of response distributions for one neuron at 14 msec/image and illustrates the analysis performed on the spike counts (see Figure 6b legend). The goal of the analysis is to determine the percentage of correct responses that would occur if an ideal observer used just the spike counts of that single neuron to determine whether the target was present or not.

Psychophysics

In 50% of the trials, the RSVP sequence contained only the seven distracters in random order; in 50% of the trials, it contained six distracters and the target in positions 3, 4, or 5 of the sequence (see Figure 1a,c). The memory and the detection task were performed on different sessions; three subjects performed the detection task first and the memory task second, for two subjects, it was the other way round. The 23 stimulus sets were tested 16 times (eight target present, eight target absent trials) in each task at presentation rates of 14, 28, 42, 56, and 111 msec in pseudorandom order for each of the five subjects. The data from the five subjects were pooled. The experiment was conducted using the same display as that used in physiology. Subjects included three authors (CK, DX, and PF) and two naive subjects (TJ, RE). None of the subjects was trained at the task, but CK, DX, and PF had been exposed to the stimuli previously during electrophysiological recordings. No systematic differences between the authors and the naive subjects were observed.

Acknowledgments

This work was supported by the BBSRC, the Boehringer Ingelheim Fond, and the Studienstiftung des deutschen Volkes. We thank S. Celebrini for the advice on the neurometric analysis, and H. Barlow, I. Biederman, E. Bowman, V. Brown, J. Bullier, A. Cugliandolo, P. Dayan, V. Gallese, C. Koch, D. Milner, M. Mon-Williams, M. Oram, A. Perrett, W. Singer, S. Thorpe, and B. Wicker for the critical comments on the manuscript.

Reprint requests should be sent to: Christian Keysers, Instituto di Fisiologia Umana, Università di Parma, 43100 Parma, Italy, or via e-mail: keysers@nemo.unipr.it.

REFERENCES

Andrews, T. J., & Purves, D. (1997). Similarities in normal and binocularly rivalrous viewing. *Proceedings of the National Academy of Sciences, U.S.A., 94,* 9905–9908.

Bridgeman, B. (1980). Temporal response characteristics of cells in monkey striate cortex measured with metacontrast masking and brightness discrimination. *Brain Research, 196,* 347–364.

Celebrini, S., Thorpe, S., Trotter, Y., & Imbert, M. (1993). Dynamics of orientation coding in area V1 of the awake primate. *Visual Neuroscience, 10,* 811–825.

Chelazzi, L., Duncan, J., Miller, E. K., & Desimone R. (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search. *Journal of Neurophysiology, 80,* 2918–2940.

Chun, M. M., & Potter, M. C. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 109–127.

Fabre-Thorpe, M., Richard, G., & Thorpe, S. J. (1998). Rapid categorization of natural images by rhesus monkeys. *NeuroReport, 9,* 303–308.

Földiák, P., & Young, M. P. (1995). Sparse coding in the primate cortex. In M. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 895–898). Cambridge: MIT Press.

Kovács, G., Vogels, R., & Orban, G. A. (1995). Cortical correlate of pattern backward masking. *Proceedings of the National Academy of Science, U.S.A., 92,* 5587–5591.

Levick, W. R., & Sacks, J. L. (1970). Responses of the cat retinal ganglion cells to brief flashes of light. *Journal of Physiology, 206,* 677–700.

Logothetis, N. K. (1998). Single units and conscious vision. *Philosophical Transactions of the Royal Society, London, Series B, 353,* 1801–1818.

Logothetis, N. K., Leopold, D. A., & Sheinberg, D. L. (1996). What is rivalling during binocular rivalry? *Nature, 380,* 621–624.

Macknik, S. L., & Livingstone, M. S. (1998). Neuronal correlates of visibility and invisibility in the primate visual system. *Nature Neuroscience, 1,* 144–149.

Morris, J. S., Ohman, A., & Dolan, R. (1998). Conscious and unconscious emotional learning in the human amygdala. *Nature, 393,* 467–470.
Newsome, W. T., Britten, K. H., Movshon, J. A., & Shalden, M. (1989). Single neurones and the perception of visual motion. In D. M.-K. Lam & C. D. Gilbert (Eds.), *Neural mechanisms of visual perception, proceedings of the retina research foundation* (pp. 171–198). The Woodlands, TX.
Oram, M. W., & Perrett, D. I. (1992). Time course of neural responses discriminating different views of the face and head. *Journal of Neurophysiology, 68,* 70–84.
Perrett, D. I., Oram, M. W., & Wachsmuth, E. (1998). Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformations. *Cognition, 67,* 111–145.
Potter, M. R., & Levy, E. I. (1969). Recognition memory for rapid sequence of pictures. *Journal of Experimental Psychology, 81,* 10–15.
Ringach, D. L., Hawken, M. J., & Shapley, R. (1997). Dynamics of orientation tuning in macaque primary visual cortex. *Nature, 387,* 281–284.
Rolls, E. T., Tovee, M. J., & Panzeri, S. (1999). The neurophysiology of backward visual masking: Information analysis. *Journal of Cognitive Neuroscience, 11,* 300–311.
Sáry, G., Vogels, R., & Orban, G. A. (1993). Cue-invariant shape selectivity of macaque inferior temporal neurons. *Science, 260,* 995–997.
Snodderly, D. M., & Gur, M. (1995). Organization of striate cortex of alert, trained monkeys (*Macaca fascicularis*): Ongoing activity, stimulus selectivity, and widths of receptive field activating regions. *Journal of Neurophysiology, 74,* 2100–2125.
Subramaniam, S., Biederman, I., & Madigan, S. A. (2000). Accurate identification but no priming and chance recognition memory for pictures in RSVP sequences. *Visual Cognition, 7,* 511–535.
Thompson, K. G., & Schall, J. D. (1999). The detection of visual signals by macaque frontal eye field during masking. *Nature Neuroscience, 2,* 283–288.
Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381,* 520–522.
Thorpe, S. J. (1990). Spike arrival times: A highly efficient coding scheme for neural networks. In R. Eckmiller, G. Hartman, & G. Hauske (Eds.), *Parallel processing in neural systems* (pp. 91–94). North-Holland: Elsevier.
Uttal, W. R. (1969). The character in the hole experiment: Interaction of forward and backward masking of alphabetic character recognition by dynamic visual noise (DVN). *Perception and Psychophysics, 6,* 177–181.