

OUT OF SIGHT BUT NOT OUT OF MIND: THE NEUROPHYSIOLOGY OF ICONIC MEMORY IN THE SUPERIOR TEMPORAL SULCUS

C. Keysers*

University Groningen, The Netherlands

D.-K. Xiao*, P. Földiák, and D. I. Perrett

University of St Andrews, UK

Iconic memory, the short-lasting visual memory of a briefly flashed stimulus, is an important component of most models of visual perception. Here we investigate what physiological mechanisms underlie this capacity by showing rapid serial visual presentation (RSVP) sequences with and without interstimulus gaps to human observers and macaque monkeys. For gaps of up to 93 ms between consecutive images, human observers and neurones in the temporal cortex of macaque monkeys were found to continue processing a stimulus as if it was still present on the screen. The continued firing of neurones in temporal cortex may therefore underlie iconic memory. Based on these findings, a neurophysiological vision of iconic memory is presented.

GENERAL INTRODUCTION

Iconic memory and information persistence

In the world around us, many important events can be fleetingly brief. A gazelle might get a brief glimpse at a lion while the wind is moving the bush hiding the lion. How do we process such brief visual stimuli? Partial report paradigms (Loftus, Duncan, & Gehrig, 1992; Sperling, 1960) show that if a complex visual stimulus is flashed very briefly on a screen, subjects can continue processing this stimulus for 200–300 ms after it has gone. This phenomenon is called “information persistence” or

“iconic memory” (Coltheart, 1980; Loftus et al., 1992). A suitable masking stimulus presented after the target is thought to overwrite this “icon” and therefore to terminate the information persistence of the previous stimulus. Despite the ubiquity of iconic memory in models of visual processing, little is yet known about the physiological basis of this phenomenon. In the present study, we examine the neural correlates of information persistence by comparing the performance of human subjects (Experiment 1) with that of single cells in the temporal cortex of macaque monkeys (Experiment 2) while rapid serial visual presentation (RSVP) was used to present naturalistic images to both groups.

Correspondence should be addressed to C. Keysers, BCN Neuro-Imaging Centre, University Groningen, A. Deusinglaan 2, 9713 AW Groningen, The Netherlands (Email: c.keysers@med.rug.nl).

*The first two authors contributed equally.

This work was supported by the BBSRC, the Wellcome Trust, the Boehringer Ingelheim Fond, and the Studienstiftung des deutschen Volkes. We thank E. Kohler and A. Perrett for critical comments, Mary Potter for suggesting that persistence might have an important role for higher brain functions, and the Sony Corporation for providing technical information about the GDM-20D11 computer screen.

RSVP with and without gaps

In RSVP, images are presented sequentially and continuously, with each image replacing the previous at the same location on the screen, one after another—much like a very rapid slide show. We have previously shown that human observers and pattern-sensitive neurones in the anterior superior temporal sulcus respond selectively to individual images in such RSVP sequences at presentation rates of up to one image every 14 ms (Keyser, Xiao, Földiák, & Perrett, 2001). Here we use the same technique to test how neurones and human subjects react to the introduction of gaps between successive frames in RSVP sequences. The experimental question is: If there is a gap between two images, will the iconic memory internally replace this gap with the “icon” of the preceding image? If so: How accurate is the icon, i.e., how does performance based on the icon compare to that based on the image itself?

To answer these two questions, two groups of RSVP sequences were tested (Figure 1a, left). Each group was composed of a set of three presentation conditions that could be compared. All stimuli were natural images (e.g., faces, real-world object), presented in the middle of a CRT computer screen with a refresh rate of 72 Hz. The images occupied only one quarter of the height of the video screen, and the phosphor decay time of the computer screen (Sony GDM-20D11) was short (≤ 1.2 ms to reach 10% of the original luminance, < 7.5 ms to reach 1% luminance). The presentation duration of a stimulus was varied by exposing it for an integer number of frames (1, 2, 4, or 8 consecutive frames) on the screen, and leaving 0, 3, or 6 frames of blank screen between consecutive stimuli (see Figure 1, left). When images are presented in the centre of a CRT screen, stimuli presented on consecutive frames are not actually continuously present on the screen (Bridgeman, 1998), but illuminated on the screen for ~ 4 ms on each frame, creating inevitable gaps of 9 ms between frames (see Figure 5). Since human subjects do not perceive these inevitable 9 ms gaps as gaps, in this paper conditions with 9 ms gaps will be called “nogap” conditions, while those with either

51 or 93 ms of blank screen will be called “gap” conditions.

The choice of conditions was dictated by a simple consideration. If human information processing based on the icon is as good as that based on the image itself, during the gaps human subjects should behave as if the stimuli had stayed on the screen. In addition to testing the performance in conditions with gaps, we therefore also tested performance in a condition referred to as “long-nogap,” in which we left the stimulus on the screen during the time that had been a “gap” in the gap condition. If, on the other hand, human information processing based on the icon is very poor, information processing should be limited to the time during which the actual stimulus is present. A third condition was therefore added to each group in which the stimuli were shown for the same time as in the gap condition, but were immediately followed by the next stimulus (“short-nogap” condition). All three conditions were tested for two different durations, creating a slow and a fast group of comparison.

METHODS

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, and found always to correspond to the intended duration. After a given pixel is presented on the screen, the P22 phosphor requires ≤ 1 ms to reach 10% and < 7.5 ms to reach 1% of the original luminance of the pixel (data provided by the Sony Corporation). Hence, given that the stimuli occupied $1/4$ of the height of the screen, the actual stimulus is shown for $1/4$ of the frame duration, i.e., $1 \text{ s} \times 72 \text{ Hz}^{-1} \times 4^{-1} = 3.4$ ms. Adding to that the duration for the last pixel to reach 10% of its original luminance, the stimulus duration is ~ 4 ms. If one places the criterion at 1% of the original luminance, frame duration then reaches ~ 11 ms (see Figure 5).

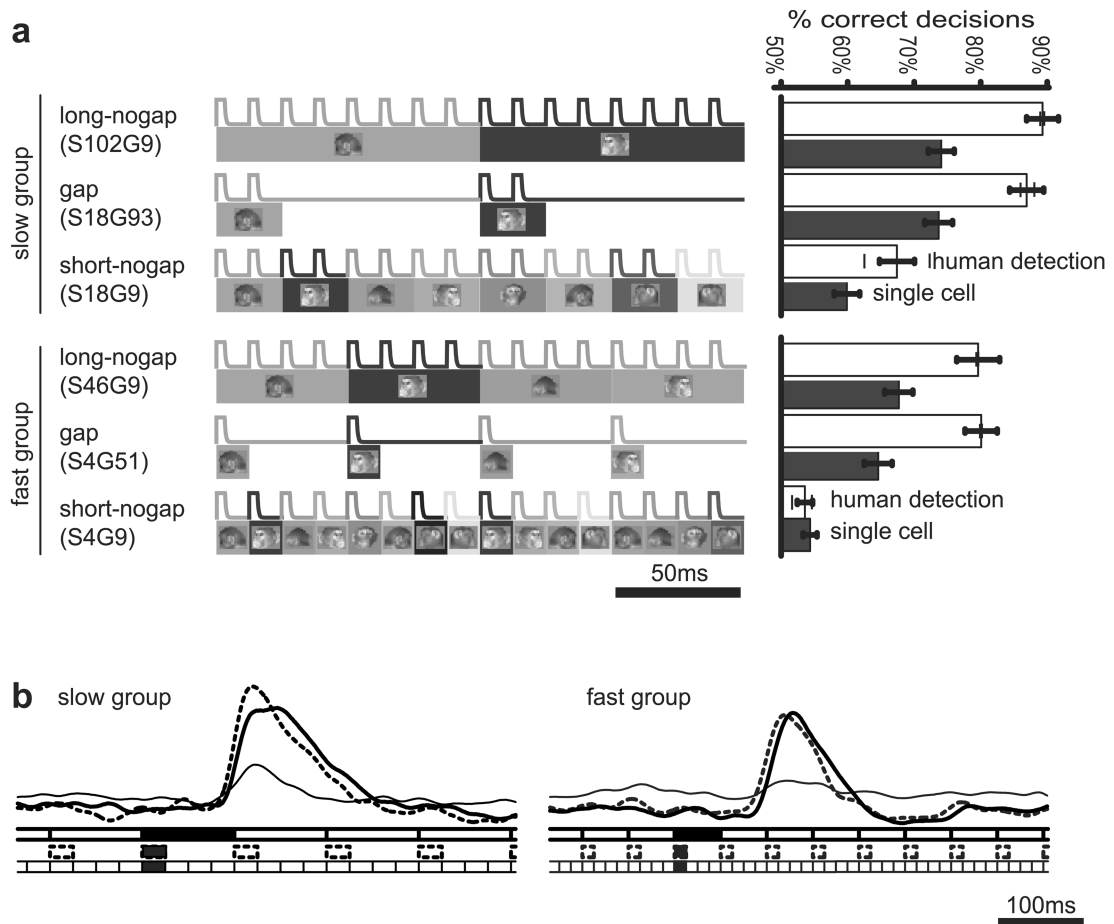


Figure 1. (a) Stimulus timing (left) and the corresponding accuracy (right) with which a single stimulus can be detected in an RSV sequence by human observers (open/green bars, the two black/red tick marks represent the individual mean accuracy for the two subjects) and by an ideal observer of STSa activity (grey/blue bars, as calculated using an ROC analysis). The x -axes of the left panel illustrate the stimulus timing in milliseconds, while the y -axes represent the energy on the screen of the stimulus shown below the curve. The timing of conditions is given as $SxGy$, where x is the duration in ms of the stimulus, and y that of the inter-stimulus gap. (b) Average normalised population response of the single neurones to the best stimulus in RSV sequences of different timing. The x -axes represent time, the y -axes the mean averaged latency-aligned population response. The horizontal "ladders" under the response represent the stimulus timing, with the filled squares representing the timing of the best stimulus onto which responses were aligned, while the open squares represent adjacent stimuli in the sequence occupied at random by one of the eight stimuli. The space between two squares represents the gap in gap sequences. The dashed curves corresponds to the long gap condition, while the thin and thick continuous curves correspond to the short and long-nogap condition, respectively. See text for details. To view this figure in colour, please see the online issue of the Journal.

For the monkeys, a fixation point appeared in the middle of the screen. Sequence presentation commenced when the subject's gaze remained within a fixation window of $\pm 5^\circ$ of the monitor centre for >500 ms and terminated after 10 s or

earlier if the monkey's gaze moved outside the fixation window. The fixation point was not visible during sequence presentation. The number of stimulus repetitions at a given presentation rate was adjusted to equate the total presentation time

at each rate. On average, testing involved a total recording duration of approximately 1 hr per neurone. For humans, a fixation dot was shown together with a warning tone, the fixation dot then disappeared, and the target image was shown for 300 ms followed by a 500 ms gap. A sequence of seven images was then presented at rates shown in Figure 1. In 50% of the cases this sequence contained the target image in position 3, 4, or 5. Subjects pressed a button if the target image was in the sequence.

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). Neurones were localised to the upper and lower banks of the STSa (12–18 mm anterior to the inter-aural plane) on the basis of x-ray visualisation of micro-electrodes. Recording sites were confirmed through MRI and histology with markers placed at the site of neurone recording [MRI: Magnavist, Schering Health Care Ltd, Burgess Hill, UK; Histology: micro-lesions 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 and spike arrival times and measured stimulus onset times.

Human subjects

The subjects included one naïve subject (TJ) and one author, aged 35 and 26 respectively. Both were male and had normal vision. In a very similar task we tested five subjects and found no difference between naïve and author subjects (Keysers et al., 2001).

Neural 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 obtained after the convolution with a $SD = 10$ ms Gaussian kernel then reflect the systematic response to the stimulus of alignment surrounded by activity evoked by all stimuli.

The responses of each neurone to each of the eight stimuli during the S102G9 condition were measured in a time window starting at 100 ms post-stimulus onset and lasting 111 ms. The stimulus eliciting the largest responses was defined as the neurone's "best" stimulus; the remaining seven stimuli were defined as "rest" stimuli. Response onset latency of a given neurone was computed off-line from trials for the "best" stimulus pooled across all presentation rates with 9 ms gaps. The latency of response onset was defined as the first 1 ms time bin of the spike density function (SDF) at which the firing rate exceeded the mean + 2.58 SD (i.e., $p < .005$) of activity measured in a control period 250 ms before stimulus onset, for at least 25 consecutive bins. Latency-aligned responses refer to responses time-shifted by the difference between an individual neurone's response onset latency and the population average (117 ms).

Population responses

Neurones differed in their response onset latency (56–171 ms). Hence, to investigate the duration of responses at population level, responses of all neurones were shifted onto the average latency of the population (117 ms). Neurones also differed with respect to their peak firing amplitude. In the S102G9 condition a post-stimulus time histogram (PSTH) of the response to the best stimulus was calculated with 1 ms bin-size. The bin with the largest spike count was then used to divide activity in all conditions.

Time window of interest

Based on the latency-aligned and normalised responses of the population of neurones, a window of analysis was defined separately for the fast and the slow comparison groups. The aim of this

window is to identify the period of time during which responses are due to the stimulus of alignment, and not to the random stimuli occurring before or after the stimulus of alignment in the sequences. Hence, for each group, the period of time during which responses were different for different stimuli of alignment was identified. For each millisecond relative to response onset, the average normalised SDF values for each cell and stimulus were compared in an eight-stimulus within-cell ANOVA (see Figure 4). The window of analysis is then the period in which, from the three conditions of one group, at least one differentiates between stimuli, i.e., the period of time for which the ANOVA yields a p value for stimulus $< .01$. In the fast group, at least one condition showed stimulus discrimination in the -11 to 95 ms period, and in the slow group, in the -24 to 133 ms period, relative to response onset. These windows were then used to count spikes for the receiver operator curve (ROC) analysis and integrate the surface under the SDF for the response comparison. The windows started before $t = 0$, the no-gap response onset latency, because at the population level, the additional statistical power reveals significant differences between responses earlier than the less powerful single cell analysis used to determine the response onset latency.

Uncontaminated responses

Response to a stimulus X is considered “uncontaminated” if it was flanked by a sufficient number of consecutive “rest” stimuli (R, i.e., all but the best stimulus) 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 (S102G9, S18G93), RRRXRRR (S46G9, S4G51), RRRXRRRR (S18G9, S4G9).

Neurometrics

For each of the 21 neurones tested, spike counts were obtained for uncontaminated stimulus

sequences that were comparable to the target present and target absent sequences in the psychophysics. “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 human 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. A receiver operator curve was then calculated using all possible thresholds, and the surface under this curve taken as the neurometric performance of an ideal observer (see Celebrini & Newsome, 1994, for details). 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 neurone to determine whether the target was present or not. The mean trial number used for the ROC analysis in the slow group was 47 trials per stimulus and cell, and 95 trials per stimulus for the fast group. These large trial numbers lead to reliable ROC estimates.

EXPERIMENT 1: PSYCHOPHYSICS

Introduction

Information persistence and visual persistence

Coltheart (1980) introduced an important distinction between two seemingly identical phenomena: information persistence and visual persistence. Both phenomena refer to the fact that when a brief visual stimulus is flashed on a screen, the processing of this stimulus can outlast the stimulus itself. Information persistence refers to the fact that if a subject needs to process the brief stimulus, performance will be better if a masking stimulus is not presented immediately after the stimulus. This indicates that some processing is still going on after the end of the stimulus presentation, and that the mask interrupts this continued processing. Measurement of information persistence does not

involve asking the subject to report if she still mentally “sees” the stimulus after it has ended. Visual persistence, on the other hand, refers to the fact that subjects subjectively have the impression that the stimulus doesn’t suddenly vanish from the screen, but rather slowly fades. If one then measures the time from beginning of the subjective sensation of seeing the image to the end of this perception, the total subjective duration is substantially longer than the stimulus duration. It is intuitively appealing to believe that *information* persistence arises from the fact that the subjects process the *visual* persistence of the stimulus. Hence visual and information persistence should have similar time courses. Unfortunately, quantitative analyses of the two phenomena show that this is not the case (see Coltheart, 1980, and Loftus & Irwin, 1998, for reviews). In the present paper, information persistence will be examined, not visual persistence.

Human perception and gaps

The effect of gaps between stimuli in RSVP was measured in two male human observers using a simple target detection task. Only two subjects were used, because a previous psychophysical study (Keysers et al., 2001) using the same task had revealed that the performances of five subjects were so similar to one another, that two subjects were sufficient to yield very representative data. Figure 2a illustrates the time course of a single trial. In each trial, the subject viewed a “target” image for 306 ms (22 frames), followed after 500 ms by an RSVP sequence of seven stimuli containing either seven distracter images (i.e., images different from the target) or six distracter images and the original target image itself in the 3rd, 4th, or 5th position of the sequence. Images were photographs of humans and monkeys (e.g., Figures 1a and 2a), laboratory objects, and images drawn from image libraries on

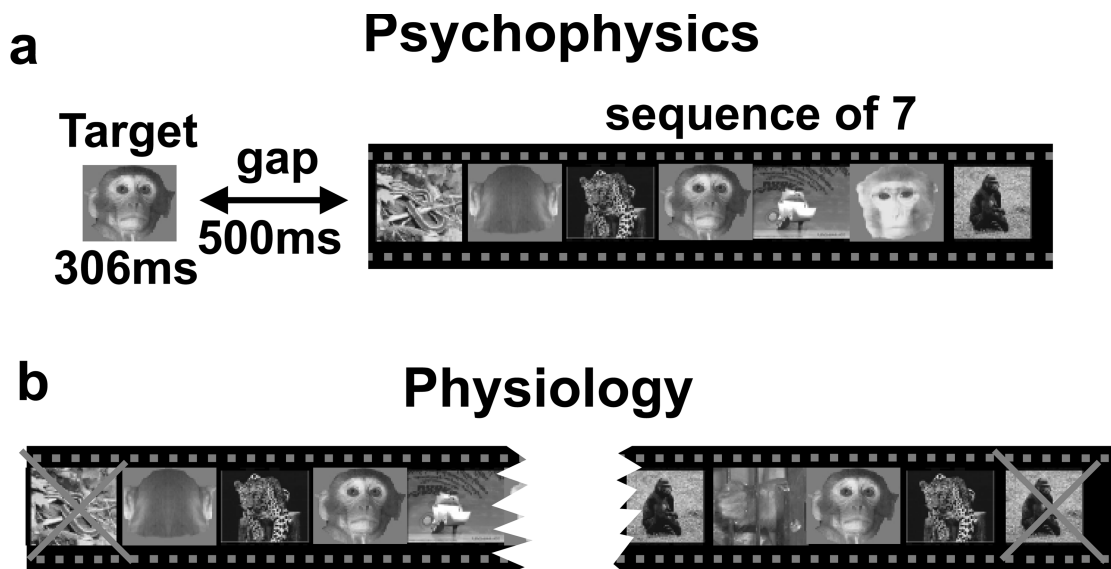


Figure 2. *Experimental design. (a) In the psychophysical detection task (Exp. 1), a single stimulus (the target) is shown for 306 ms, followed after a 500 ms gap by a sequence of seven stimuli presented using the different timing conditions. If this sequence contained the target stimulus, the subject had to press a button. (b) In the physiological recordings (Exp. 2), a continuous sequence of naturalistic images was presented to the monkey. For each neurone, 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. The first and last stimuli of each sequence were ignored in the analysis. To view this figure in colour, please see the online issue of the Journal.*

the Internet. All target images used in this study were the most effective stimuli for one neurone of the physiological study (Exp. 2) presented below. For 16 of the neurones, the best stimulus was a face. For 10/16 the face was that of a monkey, and in 6/16 it was the face of a human. Two of these cells selective for faces were tested with the target face mixed with 7 other faces, while the remaining 14 were tested against other stimuli including inanimate objects (e.g., apples, coca-cola cans, a newspaper). Five cells responded best to particular views of an entire body including the face. The remaining cells responded best to stimuli not associated with the face such as a syringe, an orange, an apple, a trash-bag, a bear, a spider, and a duck. If the observer recognised the target image in the sequence, he was required to press a button.

Figure 1 illustrates the stimulus timings used in the study. Six timing conditions were selected to fall into two comparison groups: a fast and a slow comparison group. The slow comparison group was composed of a "gap" condition, having a stimulus duration of 18 ms followed by 93 ms of blank screen; a "short-nogap" sequence, with the same stimulus duration (18 ms) but no gap (except the inevitable 9 ms); and a "long-nogap" sequence, having the same stimulus onset asynchrony (SOA = 111 ms) as the gap sequence but with the stimulus filling the entire SOA. The fast group was identical except that half as many frames were used for stimuli and gaps.

For precision, these groups and conditions are also referred to by the duration of stimulus (S) and blank screen (G) (i.e., slow group: gap S18G93, long-nogap S102G9, short-nogap S18G9; fast group: gap S4G51, long-nogap S46G9, short-nogap S4G9).

Twenty-one different target images were used with 16 trials each (8 target present and 8 target absent trials). These were tested for all six timing conditions. The distracter stimuli accompanying target stimuli were the same as those used in the testing of particular neurones described below.

Results

The results in terms of proportion correct decisions were calculated for each target image separately, but

pooling trials from the two subjects (the performance of the two subjects did not differ, Binomial test $p > .05$). The results were then averaged over images and shown as the open bars (\pm sem) in Figure 1a (right panel). To compare the performance in the three conditions of each group, due to the binary nature of the response (yes/no), a χ^2 statistic was adopted, comparing the total correct responses (i.e., pooled over stimuli and subjects) in one condition with those in another condition. For both groups, performance in the gap sequence was not different from that in the long-nogap sequence but was better than that in the short-nogap sequence. Slow group: null hypothesis H_0 : S18G93 = S102G9, $\chi^2(1) = 1.82$, $p > .17$; H_0 : S18G93 = S18G9, $\chi^2(1) = 72$, $p < .001$; Fast group, H_0 : S4G51 = S46G9, $\chi^2(1) = 0.04$, $p > .83$; H_0 : S4G51 = S4G9, $\chi^2(1) = 106$, $p < .001$.

Unsurprisingly, performance in gap and long-nogap conditions with long SOAs (111 ms) in the slow group was superior to comparable conditions in the fast group with shorter SOAs (55 ms), all $\chi^2(1) > 11$, $p < .001$. The lack of difference between the gap and long-nogap condition of the fast group cannot therefore be due to a ceiling effect.

Discussion

The performance in gap sequences was equal to that in long-nogap conditions and performance in gap sequences was much better than that in short-nogap conditions. Hence, when it comes to detecting the presence of a particular photograph in a sequence, humans perform as well based on the icon of the image as based on the photograph itself. The icon of the image is thus "as good" as having the image itself in the context of this task. This was true for gaps of up to 93 ms.

The retina and V1 cannot account for this persistence

Despite the central role played by the concept of information persistence and iconic memory in models of visual perception, little is yet known about the physiology of this phenomenon. In the

17th century, Newton suspected that the continuing activity in the eye was responsible for perceptual persistency. From such a perspective, the continuing activation in the retina would simply be sent on to all cortical areas processing this input.

Levick and Sacks (1970) measured the duration of responses in ganglion cells of the retina, and found those responses to be small and to last 50–70 ms for an 8 ms stimulus and to be strong and last about 180 ms for a 128 ms stimulus of the same intensity. Hence, for the stimuli used in the present experiment, responses should be much longer and stronger for the long-nogap compared to the gap condition. Stronger and longer responses would lead to better performance in the long-nogap condition if performance was based on the iconic properties of the retina alone. This was clearly not the case here. While it may be that the retina is responsible for part of the information persistence, retinal processing cannot by itself explain the strong persistence observed in Exp. 1.

Duysens, Orban, Cremieux, and Maes (1985) recorded the responses of single V1 neurones to optimally oriented bars. They found that the on-responses of the cells were either relatively transient, lasting for only ~25 ms independent of how long the optimally oriented stimulus was presented for. If responses were less transient, the on-responses to 100 ms stimuli were longer and stronger than those to 12.5 ms stimuli. These findings, too, would predict that an ideal observer of V1 activity would perform differently to 100 ms stimuli compared to 12.5 ms stimuli in psychophysical tasks. Duysens et al. and Levick and Sacks (1970) recorded from the cat, leading to results that may differ from those in the monkey. But Gur and Snodderly (1997) recorded from the macaque monkey, presented optimally oriented bars continuously on a 60 Hz monitor, and showed that the duration of the response of V1 cells actually reflects the duration of the stimulus on the screen, responding for only a few milliseconds on each refresh cycle of the screen. Taken together, these findings suggest that the time characteristics of responses in early visual cortex cannot account for the strong information persistence observed in Exp. 1. It should be noted, however, that this latter

result was obtained from monkeys that were trained to fixate a light-emitting diode. The stimulus itself thus had no task relevance for the monkey. It might be that in the context of a different task where the stimuli were relevant, the timing of the V1 neuronal responses would have differed.

To be able to explain the psychophysical data of Exp. 1, a neural correlate of iconic memory would need to fulfil three requirements. First, it would need to provide a representation of a stimulus, which continues beyond the stimulus's physical duration. Second, presenting a new stimulus should terminate this persistence. Third, responses in gap conditions should be as informative as the responses to long-nogap conditions but substantially more informative than those to short-nogap conditions.

EXPERIMENT 2: PHYSIOLOGY

Introduction

STSa neurones and perception

Here we propose that iconic memory is implemented in the brain by the persistent firing of visual neurones in the higher stages of visual shape processing (e.g., in the temporal cortex). When a stimulus is flashed on a screen, it causes selective responses in visual cortex. The firing of neurones in macaque inferior temporal cortex (IT) and the cortex lining the anterior superior temporal sulcus (STSa) correlates with perceptual reports (Keyser et al., 2001; Logothetis, 1998). These same neurones are known to continue responding for some hundreds of milliseconds after the end of a briefly presented stimulus (Kovács, Vogels, & Orban, 1995; Rolls & Tovee, 1994). Finally, presenting a new stimulus is known to terminate the persisting responses to a previous stimulus for these neurones (Keyser et al., 2001; Kovács et al., 1995; Rolls & Tovee, 1994). Taken together, the response patterns of the temporal cortex neurones may thus provide adequate neural correlates of information persistence. It is surprising, therefore, that temporal cortex responses have not been explicitly tested in relation to iconic memory.

Iconic memory vs. working memory

Temporal cortex neurones have been implicated in working memory (Fuster & Jervey, 1981; Naya, Sakai, & Miyashita, 1996). A typical paradigm for the study of working memory is the delayed matching to sample task. A stimulus is shown first, and then, after a retention interval of several seconds, this “old” stimulus is shown together with another “new” stimulus. The monkey has to remember the old stimulus during the retention interval to be able to indicate which of the two stimuli shown afterwards is the old stimulus. In such paradigms, visual neurones in temporal cortex responding to the presentation of the old stimulus continue to fire, albeit to a lesser extent, for some seconds during the retention period. If the monkey has a different task, this delay activity disappears (Naya et al., 1996). Hence the delay activity observed in these tasks appears to reflect a deliberate retention of the stimulus, akin to working memory in humans. Iconic memory differs from working memory in that the former lasts for hundreds of milliseconds and is automatic, while the latter lasts for several seconds and is deliberate. If STSa neurones provide a neural correlate of iconic memory then they should show persistent activity even in the absence of a memory task.

Experimental paradigm

The responses of single neurones in the STSa of two macaque monkeys were tested using images and methods that have been described previously (Keysers et al., 2001). The data reported here were acquired at the same time as those presented in that paper. Briefly, 137 neurones were initially tested by presenting 60 different stimuli in RSVP sequences using a relatively slow presentation rate (8 frames per stimulus, S102G9). RSVP sequences were created by presenting all 60 stimuli once, then permuting their order and presenting them again and again, without pauses and in random order, while the monkey fixated the centre of the computer screen (± 5 degrees). The monkeys were trained to fixate but were not trained in any memory-related tasks. Single cell responses to a particular stimulus

were assessed by aligning responses to the onset of each occurrence of this stimulus and creating an average spike density function (SDF) of the response around the onset of this stimulus. If at least one of the 60 stimuli caused a visible response in the so-created SDF, the stimulus producing the largest response was selected as the “best” stimulus for that cell, and tested using the six timing conditions outlined in Figure 1. The best stimulus for the cell was presented with seven “rest” stimuli selected to span the range of responses from second best to worst stimulus out of the original 60 (see Figure 2b).

Results

For 72/137 of these neurones, none of the 60 stimuli were effective. Another 31/137 cells were lost before completing testing. Finally, 34/137 cells were tested in all conditions of the slow group, and 21 of these cells were also tested for the conditions of the fast group. The first 13 cells were not tested in the fast group because the ability to present sequences at one frame per image (S4G9) was not available before the 14th tested cell.

Results indicate that, in STSa, the response of neurones during the presentation of gap sequences is equal to that during the presentation of long-nogap sequences but better than that during short-nogap sequences (Fig. 1a grey/blue bars and Fig. 1b). The response of a single neurone to its best stimulus (the image of an apple) is shown in Figure 3. Note how the apparent end of the response (black arrow) depends not on the duration of the stimulus, but on the time of onset of the next stimulus. The response to a longer stimulus presentation (S213G9) also demonstrates that the similarity in the responses in the S102G9 and S18G93 conditions is not due to a ceiling effect: Longer stimuli are able to produce longer responses. The same pattern is evident in the population, and was assessed quantitatively using two analyses.

ROC analysis

First, a receiver operator analysis (ROC; see, for instance, Celebrini & Newsome, 1994) was

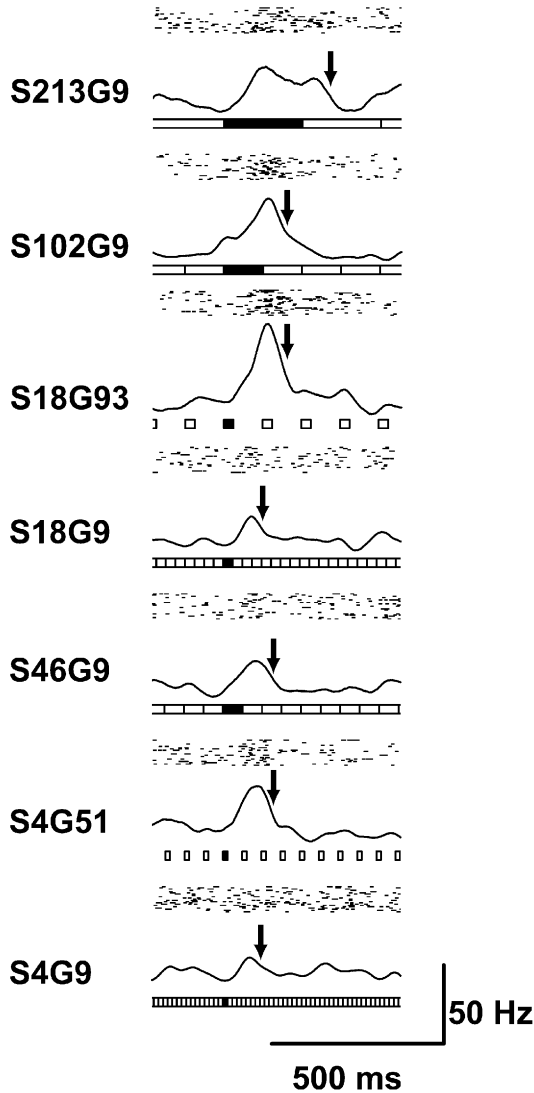


Figure 3. Rastergram and sdfs of a single neurone response in STS_a to the best stimulus for that cell (the image of an apple). Conventions as in Figure 1. For comparability only the first 21 trials are shown in each rastergram. The total trial numbers for this cell from top to bottom are 21, 42, 42, 168, 84, 84, 336. All trials are considered in the sdf. The duration of the responses decreases with decreasing SOA. The apparent ends of the responses are marked by black arrows. Comparing S102G9 with S18G93, two conditions with equal SOA (111 ms) but very different stimulus duration, the offset of the response (arrow) occurred at approximately the same time, despite the large difference in stimulus duration (102 vs. 118 ms). The same is true for

performed for each neurone in each condition based on the spike count occurring in the “time window of interest.” The time window of interest (see Methods for details) is the time during which the neurones discriminated between the best stimulus and the rest of the stimuli. Figure 4 shows the population response to the best and the rest of the stimuli, together with the probability that the cell discriminated between the best and the rest of the stimuli at that moment in time. Given that this window was slightly different for the different conditions, we considered the time window of interest to be the entire interval within each comparison group during which at least one of the conditions displayed significant stimulus discrimination. This time window reflects the interval during which the neurones respond to a particular stimulus, and was -24 to 133 ms (slow group) and -11 to 95 ms (fast group) relative to the response onset latency of each neurone. It might appear paradoxical that this time window contains negative values relative to the single cell latencies, but it should be noted that the latency is calculated cell by cell, while the time window of interest is calculated at the population level, with more statistical power. The ROC analysis yields a measure between 0 and 1, approximating how well an ideal observer of the activity of a single neurone could determine if the best or another stimulus had been presented in the sequence. Chance decision corresponds to .5 and a perfect decision corresponds to 1. To compare these results with those obtained in the psychophysical testing, this measure was calculated separately for each of the 21 cells tested in all conditions and whose best stimuli had been used as targets in the psychophysical experiment described above. Results are shown in Figure 1 as grey/blue bars (\pm sem). The results show that humans are better in their

S46G9 vs. S4G51. Surprisingly, in both cases the condition with the shorter stimulus had the larger peak firing rate but, as can be seen from Figure 1, this difference was not robust enough to be apparent at the population level. The response in the S132G9 condition is also shown in this figure to illustrate that the similarity between the S18G93 and S102G9 condition is not due to some ceiling effect: responses can last longer.

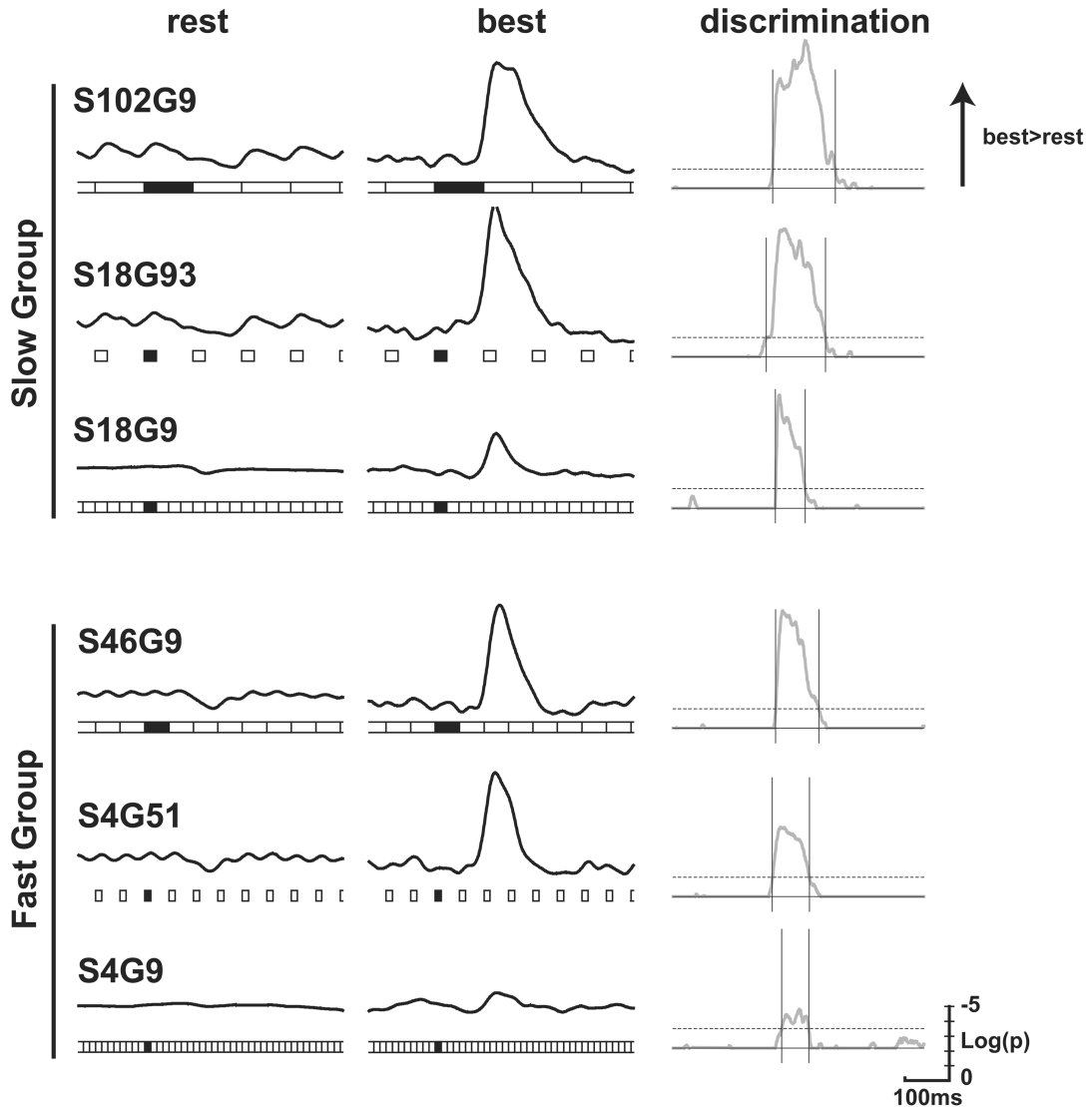


Figure 4. Time window of interest. The black curves represent normalised, latency-aligned population sdf to the rest (left) and best stimuli (middle column) in the three conditions of the slow and the fast group. The rightmost column illustrates how well, over time, the population of neurones discriminates between the best and the rest of the stimuli. It shows the $-\text{Log}(p)$ of the probability that the population of neurones does not discriminate between the stimuli, as assessed by a floating ANOVA (see Methods). High $-\text{Log}(p)$ values thus suggest high discriminative capacity between the stimuli. The dotted horizontal lines represent the $p = .05$ threshold, the black solid line the $p = 1$ level, and the vertical lines, the beginning and end of the period during which discrimination is above $p = .05$. Only deflections in the direction best > rest are shown. The scale on the bottom right applies to the x-axes of all plots, and to the y-axes of all floating ANOVA plots (grey). ANOVA p -values are two-tailed and not Bonferroni corrected.

proportion correct decisions than an ideal observer of the STSa single cell activity. Nevertheless, the ideal observer of STSa activity shows the *same pattern* of performance across conditions as that of the human observers: Performance in the gap conditions *does* differ from that in the short-nogap conditions but performance in the gap conditions does *not* differ from that in the long-nogap conditions. This effect was tested statistically using an ANOVA due to the parametric nature of the ROC values. The ANOVA (2 groups \times 3 conditions \times 21 neurones) indicated a main effect for group, $F(1, 20) = 58, p < .001$, and condition, $F(1, 20) = 61, p < .001$. The same results were found when all 34 neurones tested in the slow group were analysed only for the slow group condition. For both the fast and the slow groups, a Newman-Keuls post hoc analysis (criterion $p < .05$) indicated no difference between the gap and long-nogap conditions with equal SOA (S102G9 vs. S18G93 and S4G51 vs. S46G9, respectively). By contrast, for both the fast and the slow groups there was a significant difference between the gap condition and short-nogap conditions with equal stimulus duration but different SOAs (S18G9 vs. S18G93 and S4G9 vs. S4G51, respectively).

Population response

Single cell responses to a particular stimulus are assessed by aligning responses to the onset of each occurrence of the stimulus and creating an average spike density function (SDF) of the response around the onset of the stimulus.

A second form of analysis was based more directly on these SDFs (Figure 1b). Since the response onset latency of the 34 neurones varied considerably, responses were latency-aligned. The peak activity also varied between neurones. All SDFs of a given neurone were therefore normalised by dividing SDF values by the peak amplitude of the SDF of each neurone in the S102G9 condition. The average over all available neurones ($n = 34$ in the slow and $n = 21$ in the fast group) of these normalised and latency-aligned SDFs are shown in Figure 1b, bottom traces. As can be seen, the population response to gap sequences (S18G93

and S4G51; Figure 1b dashed/red lines) is very similar to the response to long-nogap sequences (S102G9 and S46G9, respectively, thick lines), while being much stronger than that to short-nogap sequences (S18G9 and S4G9, respectively, thin/green lines). To test this effect statistically, using the same time window of interest as in the ROC analysis, the surface under the normalised SDF to the best stimulus was integrated for each neurone, yielding one surface area value for each neurone in each of the six conditions. Within-subject F -tests comparing this surface over all available neurones confirmed that there was no difference between the responses during the gap and long-nogap sequence, but that the responses during the gap sequence differed from those during the short-nogap sequence. Slow group H_0 : S102G9 = S18G93, $F(1, 33) = 0.16, p > .69$; H_0 : S18G93 = S18G9, $F(1, 33) = 51, p < 10^{-7}$; fast group, H_0 : S46G9 = S4G51, $F(1, 20) = 0.09, p > .76$; H_0 : S46G9 = S4G51, $F(1, 20) = 15, p < .001$. These effects were also present in a majority of single neurones if tested separately.

It is also apparent that, in the slow group, the initial peak of the response appears larger in the S18G93 compared with the S102G9 condition. Again, we tested this effect statistically by comparing the surface under the curves, but this time looking at the first 60 ms of the response. This analysis showed that in the slow group the initial peak of the response was significantly higher in the gap condition, while in the fast group, only a trend in the same direction was visible: slow group, $F(1, 33) = 16.8, p < .0003$ and $F(1, 21) = 2.9, p > .09$. This effect then inverts in the second part of the responses, with the gap condition leading to slightly smaller responses. Overall the gap and long-nogap conditions then do not differ significantly.

We have repeated these analyses with both shorter (e.g., $SD = 5$ ms), and longer (e.g., $SD = 20$ ms) convolution windows for the SDF, and found virtually identical results. Indeed, even the 10 ms convolution used in the presented data is relatively short compared to the duration of the responses (~ 116 ms and 157 ms for the fast and slow group respectively).

Discussion

Overall, STSa activity was indistinguishable during gap conditions and long-nogap conditions with matched SOAs. This shows that neurones in STSa, unlike neurones in V1 (Gur & Snodderly, 1997), do not represent the duration of brief stimuli. Instead, neurones in STSa appear to process stimuli for as long as they can—unless a new stimulus is presented. This is compatible with the idea that the ventral processing stream extracts information about the shape of an object: Information about shape accumulates progressively over a period of at least 150 ms (Gershon et al., 1998), with less than 50% of the information present in the first 50 ms. If neurones responded only for the duration of brief stimuli (e.g., 18 ms), much less information about the stimulus would be extracted. While the persistence of responses in STSa therefore successfully maximises shape processing in STSa, other aspects of the stimulus, such as its duration, are lost for STSa neurones. Other cortical areas will need to provide this lost information if it is required.

The long persistence of the responses in STSa may also have another function (e.g., Földiák, 1991). When we observe an object or person rotating we have no problem understanding that the different views belong to the same object or to the same person. At a neural level, we know that most neurones in the temporal cortex respond only to certain views of faces (Perrett et al., 1991) or objects (Logothetis, Pauls, & Poggio, 1995). But how are these single views combined to enable us to understand that they all belong to a single object? The prolonged responses we observe in the STSa might be a key to this question. Our measurements showed STSa neuronal responses lasted ~60 ms longer than the SOA. This creates a temporal overlap between the neural responses of consecutive frames in the RSVP sequences. The rotation of an object on a screen can be seen as the consecutive presentation of the different views of that object. Neighbouring views will systematically occur one after another, and there will thus be overlap between the activity of neurones representing adjacent object views. This overlap could

encourage Hebbian associations between the neurones responding to the different views of the same object. This scenario would predict that some neurones should then start to respond to all the views of an object. This implication has been confirmed for both faces (Perrett et al., 1991) and objects (Logothetis et al., 1995). In addition to linking the different views of an object, persistence could be a general way for the brain to connect visual events that systematically occur one after another. Connecting the different events in an action sequence or snapshots of biological actions might be interesting examples (Giese & Poggio, 2003; Jellema & Perrett, 2003). Future investigations will be needed to determine the duration of neural responses in RSVP sequences that resemble natural sequences such as walking or rotating, and compare these with sequences containing the same frames but in random order.

In addition, the responses of the STSa neurones recorded in the present study enable an ideal observer to perform as well in gap conditions as in long-nogap conditions. Hence, the “icon” of the stimulus used during the gap period to continue processing the stimulus is as informative as a continued retinal input. The retinal input would have stopped towards the second half of the response observed in STSa during the slow gap conditions: Retinal ganglion cells respond for only about 70 ms to stimuli lasting 18 ms without any masking stimulus (Levick & Sacks, 1970). In contrast, the STSa neurones reported here responded for at least 170 ms to these stimuli. The response of V1 neurones, too, would be different for 18 ms and 102 ms stimuli (Duysens et al., 1985; Gur & Snodderly, 1997). Hence, while early areas may contribute in part to the observed information persistence, some of the persistence has to derive from the response of neurones higher than V1 in the ventral visual processing stream. Whether such long persistence is only observed in STSa, or whether earlier areas occurring between V1 and STSa already show similarly long persistence remains for future investigations to explore. It should be kept in mind that, unlike our human observers, the monkey had no task to look for a particular target in the stream of images. The duration of neural

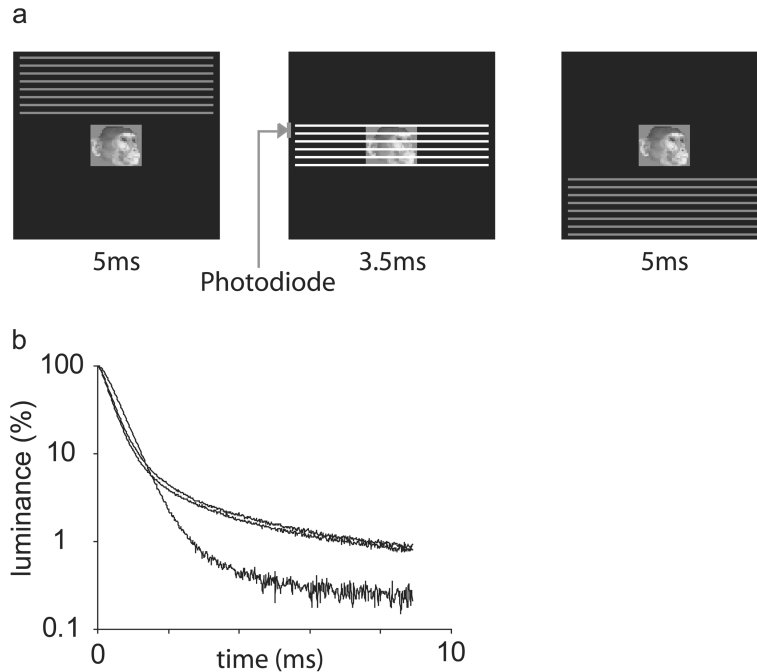


Figure 5. Actual duration of a stimulus. (a) The stimuli occupied only the centre $\frac{1}{4}$ of the screen. Of the 14 ms of total frame duration, 5.25 ms are spent sweeping the empty top and 5.25 ms the empty bottom of the screen. The actual image is swept for only 3.5 ms. Going back to the beginning of the next frame is virtually instantaneous. A photodiode on the side of the screen measured the actual beginning of the stimulus presentation. (b) After a given pixel has been illuminated on the screen (time 0 ms), the three coloured P22 phosphors of the screen progressively fade. The 10% luminance mark is reached after ~ 1 ms, the 1% luminance mark after ~ 8 ms (Graph: courtesy of the Sony Corporation). Physiological experiments in V1 suggest that at 10% contrast most, and at 1% all, neurones stop responding (Müller et al., 2001) to their preferred gratings. It is therefore reasonable to assume that the image is virtually “gone” ~ 4 ms after the beginning of its presentation (i.e., 3.5 ms of sweep time + 1 ms of decay time). This represents the timing from the beginning of the top left to the end of the bottom right corner of the stimulus. Indeed, the top left pixels are at less than 5% of their initial luminance by the time the bottom right pixel of the screen is illuminated. To view this figure in colour, please see the online issue of the Journal.

responses may look slightly different if the monkey is looking for a particular target in the sequence. If responses do indeed last longer for a target stimulus in a sequence, this prolonged response could alter the processing of subsequent stimuli in the sequence and could be at the basis of the attentional blink observed in psychophysical investigations (Keyers & Perrett, 2002; Raymond, Shapiro, & Arnell, 1992). Physiological measurements are needed to answer this issue.

It should be noted, however, that in RSVP any stimulus is subjected both to forward and backward masking: Every stimulus is both followed and preceded by other stimuli in the sequence. Forward

masking has been shown to depend both on the onset and the offset of the mask preceding the target stimulus (Crawford, 1947; Macknik & Livingstone, 1998). In the gap condition, the offset of the previous stimulus happened earlier than in the no-gap condition, and one might therefore expect less forward masking in gap conditions. The initially stronger response in the S18G93 condition seems to support this hypothesis. Pure backward masking experiments are needed to tell us if gap and long-gap conditions produce equal responses even in the absence of a forward-masking advantage for the gap condition. Unfortunately, no such experiments have been done; neither Rolls and Tovee (1994) nor

Kovács et al. (1995) ever compared the responses of gap and no-gap conditions directly.

OVERALL CONCLUSIONS

STSa and iconic memory

Overall, our data demonstrates that when the task is to decide if a particular image is in a sequence of images, human subjects perform as well to a brief stimulus followed by a gap as to a much longer stimulus followed immediately by a mask. The icon underlying the continued processing of the stimulus is thus as informative as a continued retinal input.

Recordings from single cells show that, unlike the retina (Levick & Sacks, 1970) and V1 (Duysens et al., 1985; Gur & Snodderly, 1997), the STSa provides responses that would be adequate for supporting human psychophysical decisions in our detection task. That is to say, STSa neurones respond as well during the gap condition as during the long-nogap conditions. An ideal observer of STSa activity would perform as well in these two conditions. The monkeys from which STSa activity was recorded had never been trained for a memory task, and were not required to remember stimuli during the physiological recordings. The persistence of the responses of the neurones therefore reflects an automatic persistence that is characteristic of iconic memory. Further investigations will be needed to discover if the decay rate of the response of neurones in the STSa corresponds to the decay of information observed in humans (Loftus et al., 1992).

A multiplayer account of iconic memory

There has been a long-lasting debate on the exact information contained in iconic memory (see Coltheart, 1983, for a review). Is the icon like a slowly fading photograph of the stimulus, containing only pre-categorical information, or is it a description of the image containin post-categorical information (e.g., “a monkey face in the middle of the screen looking to the right,” albeit in a nonverbal form)? Historically, answers to this question

were limited to psychophysical considerations. In particular, two findings challenged the idea that there is a simple answer to this question. First, if subjects are shown a matrix containing digits and letters, and later asked to report all digits, they perform better than if asked to report all items (Duncan, 1983). Hence, the category of the stimulus (letter vs. digit) can be a cue for retrieval from iconic memory, and iconic memory must therefore contain some post-categorical (letter vs. digit) information. Second, when a spatial cue is used (i.e., “Tell me which item was present here in the matrix”) subjects make two types of error: misidentification error, i.e., reporting an item that does not exist in the matrix; and mislocation errors, i.e., reporting a number that is in the matrix at a location that was close to but not exactly at the cued location (Townsend, 1973). Misidentification errors are compatible with the idea of a fading photograph that renders identification increasingly difficult; mislocation errors on the other hand are not: They require that the subject has identified the stimulus but has no clear information about the location of the item. Mislocation errors appear to occur predominantly if the cue is given relatively late after the original stimulus, suggesting that space and category information may fade at different rates in iconic memory.

Taking a physiological perspective on the issue may help the understanding of these two phenomena. The visual system can be described as being organised as a hierarchy of areas, with “early” areas such as V1 having small receptive fields and responding to simple features such as lines, and “later” visual areas having much larger receptive fields and much more complex visual properties, with single neurones representing objects as complex as faces. Hence early areas have precise spatial information due to their small receptive fields, but no information about the high-level category of an image (e.g., is it a face?). Late areas, on the other hand, lack the spatial information because of their large receptive fields, but have information about the high-level category of a stimulus. Indeed, given that between V1 and STSa the complexity of a stimulus required to drive cells appears to increase progressively (from simple edges to face

patterns), the concept of pre- and post-categorical information used by psychophysicists should be replaced with the idea of a continuum of degrees of shape categorisation. The fact that responses in the later visual areas of the ventral stream have longer latencies and longer persistence compared to the earlier areas has implications for the processing of a brief stimulus. Directly after stimulus onset, only very elementary feature (pre-categorical) information about the stimulus is present, due to the longer latencies of neurones higher up in the hierarchy. Shortly thereafter, progressively more high-level information is added by higher visual cortex while the pre-categorical image is still available from the persisting responses of earlier neurones. At this time, objects can be identified based on their category. Some time later, the pre-categorical information represented in early visual cortex fades away leaving only the high-level, categorical, but poorly localised information embodied in the responses of higher visual cortex neurones with longer persistence and latencies. At this time, mislocation errors will be likely. Finally, after several hundred milliseconds, even this post-categorical information stored in higher visual cortex will fade away. Then only that information selected for working memory (Fuster & Jervey, 1981; Naya et al., 1996) will be kept alive, with weaker delay activity responses and limited memory capacity. Iconic memory in this neurophysiological perspective is thus a multilayered process, composed of a spectrum of levels of categorisation with different time courses. In addition to the layers of the ventral visual processing stream, the dorsal processing stream adds additional layers to the picture. Given that different areas have different persistence durations, the information available from different "layers" comes and goes with different time courses.

Comparative measurements of the rate of decay in the response to brief stimuli in the different areas along the ventral visual processing stream would be needed to make precise hypotheses about the time course of the different information layers that could underlie iconic memory.

PrEview proof published online 17 January 2005

REFERENCES

- Bridgeman, B. (1998). Durations of stimuli displayed on video display terminals: $(n-1)/f$ plus persistence. *Psychological Science*, *9*, 232–233.
- Celebrini, S., & Newsome, W. T. (1994). Neuronal and psychophysical sensitivity to motion signals in extrastriate area MST of the macaque monkey. *Journal of Neuroscience*, *14*, 4109–4124.
- Coltheart, M. (1980). The persistences of vision. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *290*, 57–69.
- Coltheart, M. (1983). Iconic memory. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *302*, 283–294.
- Crawford, B. H. (1947). Visual adaptation in relation to brief conditioning stimuli. *Proceedings of the Royal Society of London*, *134b*, 283–302.
- Duncan, J. (1983). Perceptual selection based on alphanumeric class: Evidence from partial reports. *Perception and Psychophysics*, *33*, 533–547.
- Duysens, J., Orban, G. A., Cremieux, J., & Maes, H. (1985). Visual cortical correlates of visible persistence. *Vision Research*, *25*, 171–178.
- Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, *3*, 194–200.
- Fuster, J. M., & Jervey, J. P. (1981). Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. *Science*, *212*, 952–955.
- Gershon, E. D., Wiener, M. C., Latham, P. E., & Richmond, B. J. (1998). Coding strategies in monkey V1 and inferior temporal cortices. *Journal of Neurophysiology*, *79*, 1135–1144.
- Giese, M. A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *National Review of Neuroscience*, *4*, 179–192.
- Gur, M., & Snodderly, D. M. (1997). A dissociation between brain activity and perception: Chromatically opponent cortical neurons signal chromatic flicker that is not perceived. *Vision Research*, *37*, 377–382.
- Jellema, T., & Perrett, D. I. (2003). Perceptual history influences neural responses to face and body postures. *Journal of Cognitive Neuroscience*, *15*, 1–11.
- Keyers, C., & Perrett, D. I. (2002). Visual masking and RSVP reveal neural competition. *Trends in Cognitive Science*, *6*, 120–125.
- Keyers, C., Xiao, D. K., Földiák, P., & Perrett, D. I. (2001). The speed of sight. *Journal of Cognitive Neuroscience*, *13*, 90–101.

- Kovács, G., Vogels, R., & Orban, G. A. (1995). Cortical correlate of pattern backward masking. *Proceedings of the National Academy of Sciences USA*, *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.
- Loftus, G. R., Duncan, J., & Gehrig, P. (1992). On the time course of perceptual information that results from a brief visual presentation. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 530–549 [discussion, 550–561].
- Loftus, G. R., & Irwin, D. E. (1998). On the relations among different measures of visible and informational persistence. *Cognitive Psychology*, *35*, 135–199.
- Logothetis, N. K. (1998). Single units and conscious vision. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *353*, 1801–1818.
- Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, *5*, 552–563.
- Macknik, S. L., & Livingstone, M. S. (1998). Neural correlates of visibility and invisibility in the primate visual system. *Nature Neuroscience*, *1*, 144–149.
- Muller, J. R., Metha, A. B., Krauskopf, J., & Lennie, P. (2001). Information conveyed by onset transients in responses of striate cortical neurons. *Journal of Neuroscience*, *21*, 6978–6990.
- Naya, Y., Sakai, K., & Miyashita, Y. (1996). Activity of primate inferotemporal neurons related to a sought target in pair-association task. *Proceedings of the National Academy of Sciences USA*, *93*, 2664–2669.
- 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., Harries, M. H., Bevan, R., Hietanen, J. K., Benson, P. J., & Thomas, S. (1991). Viewer-centred and object-centred coding of heads in the macaque temporal cortex. *Experimental Brain Research*, *86*, 159–173.
- Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink? *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 849–860.
- Rolls, E. T., & Tovee, M. J. (1994). Processing speed in the cerebral cortex and the neurophysiology of visual masking. *Proceedings of the Royal Society of London B: Biological Sciences*, *257*, 9–15.
- 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.
- Sperling, G. (1960). The information available in brief visual presentation. *Psychological Monographs*, *74*, 1–29.
- Townsend, V. M. (1973). Loss of spatial and identity information following a tachistoscopic exposure. *Journal of Experimental Psychology*, *98*, 113–118.