Short temporal asynchrony disrupts visual object recognition

Jedediah M. Singer

Department of Ophthalmology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

Gabriel Kreiman

Department of Ophthalmology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA
Center for Brain Science, Harvard University, Cambridge, MA, USA
Swartz Center for Theoretical Neuroscience, Harvard University, Cambridge, MA, USA

Humans can recognize objects and scenes in a small fraction of a second. The cascade of signals underlying rapid recognition might be disrupted by temporally jittering different parts of complex objects. Here we investigated the time course over which shape information can be integrated to allow for recognition of complex objects. We presented fragments of object images in an asynchronous fashion and behaviorally evaluated categorization performance. We observed that visual recognition was significantly disrupted by asynchronies of approximately 30 ms, suggesting that spatiotemporal integration begins to break down with even small deviations from simultaneity. However, moderate temporal asynchrony did not completely obliterate recognition; in fact, integration of visual shape information persisted even with an asynchrony of 100 ms. We describe the data with a concise model based on the dynamic reduction of uncertainty about what image was presented. These results emphasize the importance of timing in visual processing and provide strong constraints for the development of dynamical models of visual shape recognition.

Introduction

Humans and other primates can recognize complex objects and scenes in a glimpse. Psychophysics data suggest rapid processing of visual information within approximately 150 ms of stimulus presentation (Kirchner & Thorpe, 2006; Potter & Levy, 1969). The ability to recognize complex shapes is instantiated by the cascade of processes along the ventral visual stream (Connor, Brincat, & Pasupathy, 2007; Logothetis & Sheinberg, 1996; Serre et al., 2007; Tanaka, 1996). Consistent with the behavioral measures, single neuron recordings in macaque inferior temporal cortex (Hung, Kreiman, Poggio, & DiCarlo, 2005; Richmond, Optican, & Spitzer, 1990; Rolls, 1991), electroencephalographic signals from the human scalp (Johnson & Olshausen, 2003; Thorpe, Fize, & Marlot, 1996), and intracranial field potentials from the human occipital and inferior temporal cortex (Liu, Agam, Madsen, & Kreiman, 2009) have revealed image-specific responses as early as 100–150 ms after stimulus onset. While these neurophysiological and behavioral observations suggest that information can quickly propagate through the visual system, it is not clear what proportion of human performance this initial wave of activity can account for; recognition in the natural world may require significant temporal integration.

Given that there are at minimum approximately eight synapses between photoreceptors in the retina and high-level visual neurons in the inferior temporal cortex, several investigators have argued that, to a first approximation, rapid recognition can be reasonably described by a mostly bottom-up hierarchy of transformations along the ventral visual stream (Deco & Rolls, 2004; DiCarlo, Zoccolan, & Rust, 2012; Fukushima, 1980; Riesenhuber & Poggio, 2000; Rolls, 1991; Serre et al., 2007; vanRullen & Thorpe, 2002). These so-called feed-forward architectures represent a major simplification of the complex organization of neocortex, which includes ubiquitous top-down signals and horizontal connections in addition to bottom-up synapses (Douglas & Martin, 2004; Felleman & Van Essen, 1991). Independently of the relative contributions of bottom-up and top-down signals, the short latencies observed in the human and macaque inferior temporal cortex impose a maximum limit of 10 to 20 ms on the amount of processing that can take place at

Citation: Singer, J. M., & Kreiman, G. (2014). Short temporal asynchrony disrupts visual object recognition. *Journal of Vision*, *14*(5):7, 1–14, http://www.journalofvision.org/content/14/5/7, doi:10.1167/14.5.7.

doi: 10.1167/14.5.7 Received November 15, 2013; published May 12, 2014 ISSN 1534-7362

each stage before the first bits of information are passed on to the next stage.

Despite the rapid progression of this initial wave of information, response durations can extend from a few tens of ms in V1 (Ringach, Hawken, & Shapley, 2003) to approximately 70 ms in MT (Bair & Movshon, 2004) to 100 ms or more in inferior temporal cortex (De Baene, Premereur, & Vogels, 2007). Spatiotemporal integration is critical for the definition of receptive fields in early visual areas as well as for motion detection signals along the dorsal stream. The contributions of spatiotemporal integration along the ventral stream in regions involved in high-level shape recognition such as inferior temporal cortex are less clearly understood. The accumulation of visual shape information over time may improve recognition performance beyond what could be achieved by independent processing of sequential "snapshots."

Psychophysics studies have shown that subjects can hold a representation of basic visual information such as letters (Sperling, 1960) or flashing arrays of squares or dots (Brockmole, Wang, & Irwin, 2002; Hogben & Di Lollo, 1974) in "iconic memory" for a short time after presentation. The recognition of a whole object can be facilitated when presented in close temporal contiguity and spatial register with one of its parts (Sanocki, 2001), demonstrating that varying the dynamics with which object information is presented can influence perception. Prior studies of temporal integration in terms of interference between parts of different faces (Anaki, Boyd, & Moscovitch, 2007; Cheung, Richler, Phillips, & Gauthier, 2011; Singer & Sheinberg, 2006), emotion recognition (Schyns, Petro, & Smith, 2007), and extraction of low-level shape features (Aspell, Wattam-Bell, & Braddick, 2006; Clifford, Holcombe, & Pearson, 2004) have found relevant time scales in the range of several tens to a few hundred milliseconds. Longer integration windows, up to several seconds, have been reported in studies of motion discrimination (Burr & Santoro, 2001) and biological motion discrimination (Neri, Morrone, & Burr, 1998).

There is thus a range of temporal integration windows that could play a role in recognition of complex objects. Observations of rapid processing suggest that even small disruptions to simultaneity might carry large consequences, while observations of long temporal receptive fields and integration or persistence of low-level visual stimuli suggest that object recognition might be robust to temporal asynchrony. Here we used psychophysics experiments in which images were broken into asynchronously presented parts (Figure 1) to evaluate the time course over which asynchronous information can be integrated together and lead to the recognition of complex objects.

Methods

All procedures were carried out with subjects' informed consent and in accordance with protocols approved by the Boston Children's Hospital Institutional Review Board.

Apparatus

Stimuli were presented on a Sony Multiscan G520 21-in. cathode-ray tube monitor (Sony Corporation, Tokyo, Japan), running at a 170 Hz refresh rate and 872 × 654 pixel resolution. The experiment was run on an Apple MacBook Pro computer (Apple Computer, Cupertino, CA) running MATLAB software (MathWorks, Natick, MA) with the Psychophysics Toolbox and Eyelink Toolbox extensions (Brainard, 1997; Cornelissen, Peters, & Palmer, 2002; Pelli, 1997). For 37 of the 50 subjects across all experiments, we obtained reliable measurements of subjects' eye movements using the EyeLink 1000 system (using infrared corneal reflection and pupil location, with nine-point calibration) running in remote mode (SR Research, Mississauga, Ontario). Subjects were seated in a dimly lit windowless room approximately 53 cm from the eye tracking camera, and approximately 71 cm from the display monitor. Subjects performed a four-alternative forced choice task (described below) and indicated their responses using a Logitech Cordless RumblePad 2 game controller (Logitech, Fremont, CA). Trials in which the requested times for stimulus onset were missed by more than one screen refresh (5.9 ms) were discarded. In separate tests, we used a photodiode to independently verify the accuracy of the image presentation times.

Stimulus presentation, Experiment 1

Sixteen subjects participated in Experiment 1. Images were drawn randomly with replacement from a library of 428 grayscale silhouettes of animals, people, plants, and vehicles, with each category equally represented. These images were obtained from several freely accessible sources on the Internet, resized to occupy most of a 256 × 256 pixel square (which sometimes involved clipping of object edges), superimposed on a noise background, and adjusted using Photoshop (Adobe, San Jose, CA) to have flat intensity histograms. The power spectrum of each image was then calculated using a Fast Fourier Transform (FFT), along with the average power spectrum across all images. The final set of images was generated by, for each image, taking the inverse FFT of the population

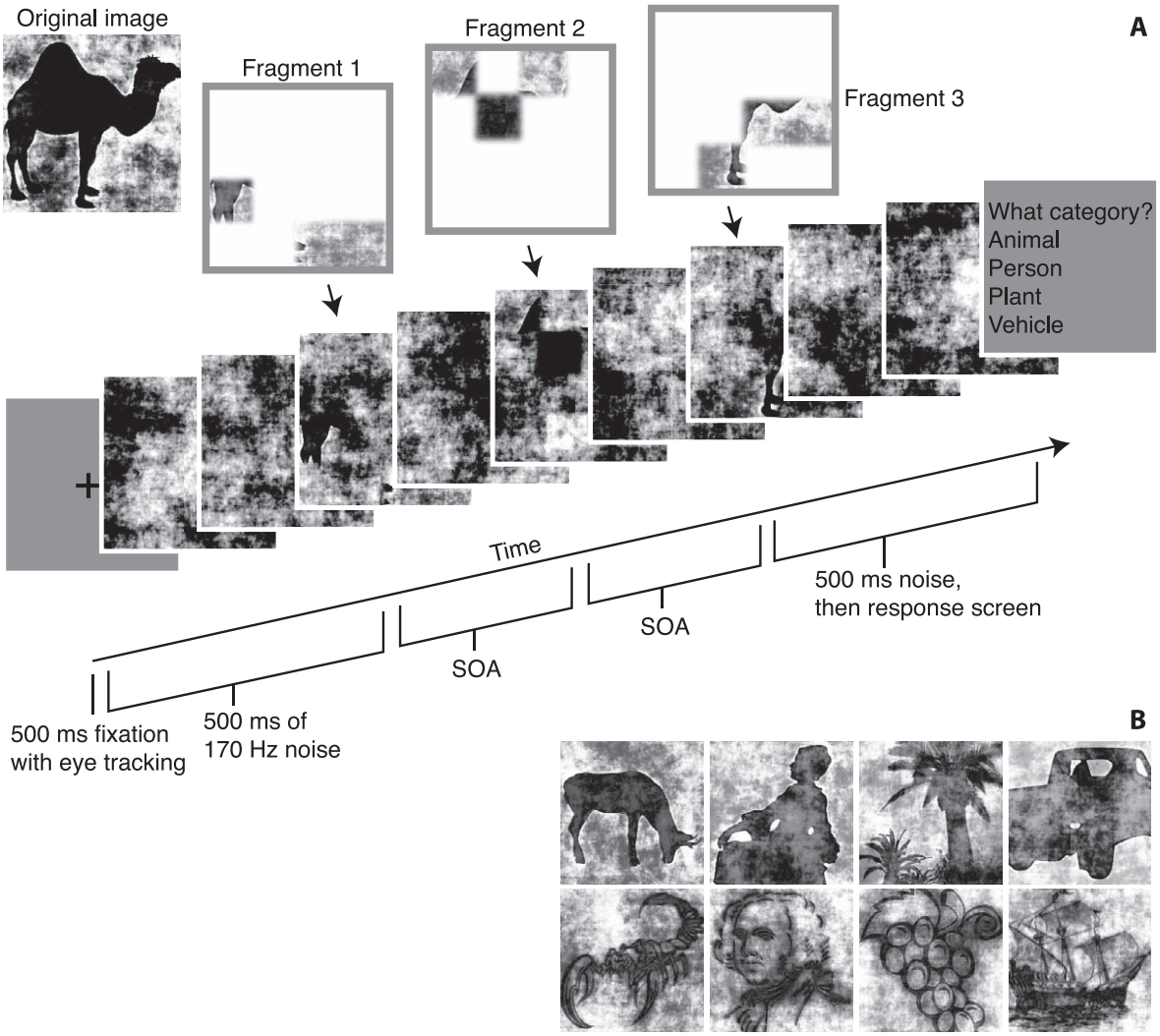

Figure 1. Schematic of the organization of the experiments. A. After 500 ms fixation, between one and four fragments, each consisting of 3/16 of a source image (upper left), were briefly shown embedded in a stream of flickering 170 Hz phase-scrambled visual noise. Fragments lasted 11.8 ms and were presented with a stimulus onset asynchrony (SOA, constant within a given trial) that took 1 of 10 possible values from 0 ms (synchronous presentation) to 294.1 ms (Methods). The source image, the number of fragments, and the SOA value were randomly chosen in each trial. Noise continued for 500 ms after the onset of the final fragment, after which subjects were presented with a four-alternative forced choice categorization task. B. Examples of other images used in the experiments.

average power spectrum with that image's respective unique phase spectrum. This resulted in a set of images with consistent power at each spatial frequency, and backgrounds filled with cloud-like noise. Example images are shown in Figure 1B (the line drawings on the second row were only used in Experiment 2); all the images used in these experiments can be obtained from http://klab.tch.harvard.edu/resources/singer_asynchrony.html.

Each of 16 subjects (13 female, age 19–35, mean age 25) completed 50 blocks of 46 trials each. The sequence of events in each trial is illustrated in Figure 1A. The image selected for each trial was drawn with equal probability from one of the four categories. Each trial began with a 14-pixel (0.5°) fixation cross presented for 500 ms. In those subjects for which eye tracking was available (14 of 16 subjects), we required 500 ms of fixation within 3° of the center of the cross. Following fixation, a 256 × 256 pixel (9.3° × 9.3°) square of flickering (170 Hz) noise appeared in the center of the screen. This noise was generated by randomizing the phase of the mean power spectrum of the library of images. The purpose of the noise was to eliminate the possible effects of apparent motion between two successive fragments (e.g., Cavanagh, Holcombe, & Chou, 2008). We considered a 4 × 4 square tessellation of each image; edge alpha masks for these squares were blurred with a Gaussian blur of radius 4 pixels, using Photoshop. We defined a "fragment" as 3 out of the 16 squares (not necessarily spatially contiguous) randomly

chosen in each trial. Depending on the configuration of the three squares, each fragment spanned a minimum of 4.7° and a maximum of 9.3°. Spatial integration of all of the information in four fragments required integrating over 9.3°. Each fragment was present for 11.8 ms (two monitor refreshes). Trials containing one, three, or four fragments were randomly interleaved. When more than one fragment was used in a trial, they contained nonoverlapping parts of the source image. Fragments were sequentially presented, with the first fragment appearing 500 ms after the noise began. Successive fragments were presented with a stimulus onset asynchrony (SOA) selected at random from the following list: 0 ms (synchronous presentation), 11.8 ms (one fragment's offset coincided with the subsequent fragment's onset), 17.7 ms, 29.4 ms, 47.1 ms, 70.6 ms, 100 ms, 147.1 ms, 205.9 ms, 294.1 ms. The SOA value was constant within a trial and the different SOAs were randomly interleaved across trials (Movie 1). We also included randomly interleaved catch trials, in which 15/16 of an image was shown for 58.8 ms. The catch trials served to ensure that subjects were engaged in the task and understood the instructions; subjects who scored below 80% on catch trials, or below 1/3 overall, were excluded (criteria established a priori, all subjects passed in Experiment 1).

The flickering noise continued while the fragments were displayed and for 500 ms after the last fragment appeared, at which point it was replaced by a blank gray screen containing the text "Please indicate which kind of picture was hidden in the clouds:" and four options were shown, corresponding to the four gamepad buttons used, for "Animals," "People," "Vehicles," and "Plants." This choice screen remained until the subject pressed a button. Performance is reported as mean ± *SEM* throughout the manuscript. A high-contrast mask (one of the noise images, selected at random, with its contrast maximized) then appeared for one monitor refresh, followed by the fixation cross for the next trial. Trials in which the subject looked away from the stimulus or blinked were excluded from analyses. No feedback was given, except between blocks, when the overall performance for that block was displayed.

Stimulus presentation, Experiment 2, variant A

Experiment 2A addressed two questions. First, can the uncertainty reduction model accurately describe the simplest possible asynchronous trials, containing only two fragments? Second, while the performance curves from Experiment 1 appeared to be near asymptote by 294.1 ms SOA, do the results hold at very long SOA? Images were drawn from a library of 364 line drawings and silhouettes, which partially overlapped with that

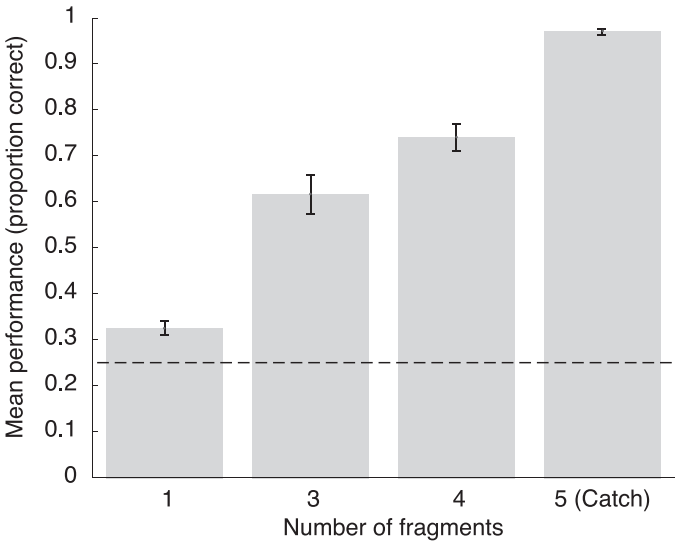

Figure 2. Mean performance across all 16 subjects in Experiment 1, in trials with zero SOA. Error bars indicate standard errors of the mean (*SEM*). All presentations lasted for 11.8 ms, except the catch trials, which lasted 58.8 ms. The dashed line indicates chance performance.

used in Experiment 1. Aside from the different image set and the additional conditions, the experiment parameters were identical to those in Experiment 1. Twenty-one new subjects (13 female, age 19–51, mean age 27) participated in this experiment. We excluded from analyses one subject due to low performance on catch trials (< 0.8 correct) and three subjects due to low overall performance ($< 1/3$ correct); this exclusion did not significantly change the overall results. In this variant, we had eye tracking available for 10 of the 17 subjects whose performance exceeded threshold (the eye tracker was unavailable for five subjects and calibration was unreliable for two subjects). Catch trials were included in six subjects. There were 36 blocks of 36 or 37 trials, depending on whether catch trials were included.

Stimulus presentation, Experiment 2, variant B

In Experiment 1, longer SOAs and more fragments both led to longer trials. These differences might have influenced the observed results (e.g., forgetting information over the course of a second or more may lead to lower performance in longer trials). To control for this possibility, we fixed the total trial duration in a second variant experiment. Moreover, during the approximately 2-hr duration of each run of Experiment 1, each 1/16 of each image was presented an average of 3.4 times. To ensure that this repetition did not lead to rote memorization or long-term learning effects, this second variant also included a larger image library, so that

each 1/16 of each image was presented to each subject an average of 0.5 times. There were 36 blocks of 34 trials. SOA values were drawn from the same list as in Experiment 1. Two-fragment trials were also presented but not included in the reported analyses to make the results more directly comparable to those in Experiment 1. Adding the data from the two-fragment trials did not change the conclusions. One of the two key differences in this variant is that the stream of flickering noise in each trial lasted for 1882.4 ms (for all SOA values and number of fragments)—long enough to accommodate a trial with four fragments at the longest asynchrony used (294.1 ms), followed by 500 ms of noise. Trials with shorter asynchronies or fewer fragments concluded with a longer period of uninformative noise. In this way the time interval between trial onset and behavioral response was identical in all trials. The second key difference in this variant is that images were drawn at random from a library of 1,200 line drawings and silhouettes, a superset of those used in Experiments 1 and 2A. This means that each 1/16 of each image had an approximately 50% chance of being presented to a given subject. Eighteen new subjects (13 female, age 18–35, mean age 24) participated in Experiment 2B; one subject was excluded from analyses due to low performance on catch trials (< 0.8 correct). We had reliable eye tracking in 13 of the 17 subjects whose performance exceeded threshold.

Data analyses

Diagnostic image parts

Even though we took several precautions to make images homogeneous in terms of basic low-level properties (see above), some images were easier to categorize than others and parts of some images were more informative than others. If part of an image is reliably diagnostic, subjects may be able to perform the task based only on that part, regardless of any other information presented. This would confound the study of temporal integration; there might be no contribution from the other fragments presented, and thus SOA would be irrelevant. To mitigate this effect, we investigated how informative each part of each image was. Each square j ($j = 1, \ldots, 16$ for each image) was assigned a "diagnosticity" score DS_j defined as:

$$DS_j = \frac{\sum_t \frac{c(t) \cdot s(t) \cdot p(t)}{a(t)}}{\sum_t \frac{s(t) \cdot p(t)}{a(t)}}. \quad (1)$$

Here, t ranges over all trials, for all subjects and across all experiments in which square j was present. If the category reported in a given trial t was correct, $c(t)$ is 1; otherwise $c(t)$ is 0. The overall performance of the subject who performed trial t is $p(t)$, and that subject's overall performance on trials with the same number of fragments and the same asynchrony as trial t is given by $a(t)$. Finally, $s(t)$ is the number of 1/16 squares of the image present overall in trial t (that is, three times the number of fragments in trial t). This score indicates how reliably each particular 1/16 square led to correct classification across all subjects. We excluded trials in which any of the 1/16 squares shown had a diagnosticity score greater than 0.8. This excluded an average of 33.7% ± 0.3% of trials from Experiment 1 and 17.2% ± 0.3% and 23.1% ± 0.3% from Experiments 2A and 2B, respectively. Including these trials does not change the conclusions reported here. The uncertainty reduction model's window of integration for individual subjects (see below) when including all trials changed by only 13% ± 3%. Including trials containing diagnostic squares does however raise performance disproportionally at longer asynchronies due to the greater number of trials in which a single fragment is sufficient for categorization. In decreasing the dynamic range of performance, the variances of the model fits are increased.

Probability summation

Assuming independent responses to each fragment, one may predict performance in trials with multiple fragments from the performance observed in single-fragment trials, y_1. We assume that the subject actually knew the correct image category in some fraction i of the trials, and guessed at chance in the rest, so that $y_1 = i + (1 - i)/4$ ($y_1 = 1/4$ if $i = 0$ and $y_1 = 1$ if $i = 1$). The chance that the subject fails to discern the category from a trial with n fragments is $(1 - i)^n$, and the final performance predicted by probability summation is then

$$1 - (1 - i)^n + \frac{(1 - i)^n}{4} = 1 - \frac{3(1 - i)^n}{4}. \quad (2)$$

Note that even for small values of i, this expression can be appreciably higher than chance (0.25). For example, a single fragment performance of $y_1 = 0.30$ is obtained from $i = 0.0667$ and this leads to an overall performance of 0.43 when $n = 4$ fragments. We cannot assume independence, however, because multiple fragments from a single image contain correlated information, and so this independent probability summation prediction overestimates expected performance. Observing the opposite pattern, i.e., exceeding the performance predicted by independent probability summation, is thus a particularly strong indicator of synergistic spatiotemporal integration of information from multiple fragments.

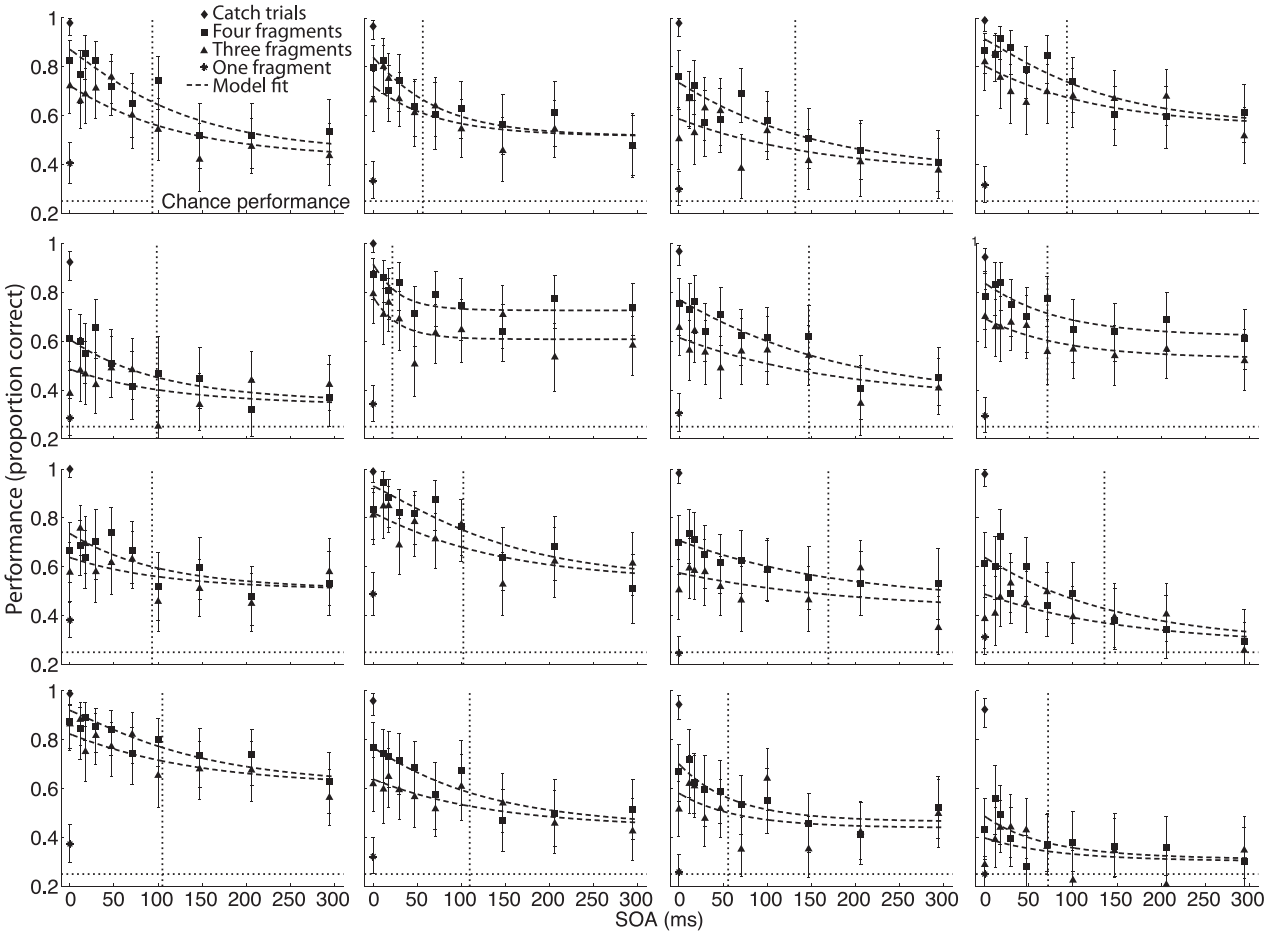

Figure 3. Individual subjects' performance in Experiment 1. Error bars indicate Clopper-Pearson 95% confidence intervals. The *x* axis indicates SOA and the *y* axis indicates the average fraction correct performance for the three-fragment condition (triangles), four-fragment condition (squares), and single fragment condition (*). Chance performance is 0.25 and the performance in catch trials was 0.97 ± 0.01. The dashed lines show the fits from the uncertainty reduction model (Text). The mean value of the integration window indicated by the uncertainty reduction mode is indicated by a vertical dotted line.

Results

We evaluated the sensitivity to temporal imprecision during visual object recognition by quantifying categorization of asynchronously presented object fragments (Methods, Figure 1).

Experiment 1

Overall performance in the task was above chance (chance = 0.25) for all subjects, numbers of fragments, and SOA values. Averaged across all SOA values and numbers of fragments, performance ranged from 0.37 to 0.77 (0.60 ± 0.03, mean ± *SEM*). Performance increased as the number of fragments shown increased (Figure 2). All 16 subjects' performance was better for the four-fragment condition than the three-fragment condition, significantly so for 13 subjects (chi-squared test, $p < 0.05$, compare triangles versus squares in Figure 3). Catch trial performance was essentially at ceiling (0.97 ± 0.01). Single-fragment performance was slightly above chance levels (0.33 ± 0.01). Overall performance varied slightly but significantly by category (chi-squared test, $p < 10^{-8}$; animals: 0.59, people: 0.60, plants: 0.54, vehicles: 0.57). In sum, all subjects were able to correctly perform the task in spite of the rapid presentations and fragmented objects.

If subjects performed the task by independently evaluating each fragment, the performance in single-fragment trials would be predictive of overall performance in trials with multiple fragments via independent probability summation (Methods). Different fragments from a given image contain redundant information, which would bias an independent probability summation prediction towards higher values. Yet performance at all SOAs was higher than that predicted by probability summation, dropping at the longest SOA to values consistent with probability summation (at SOA

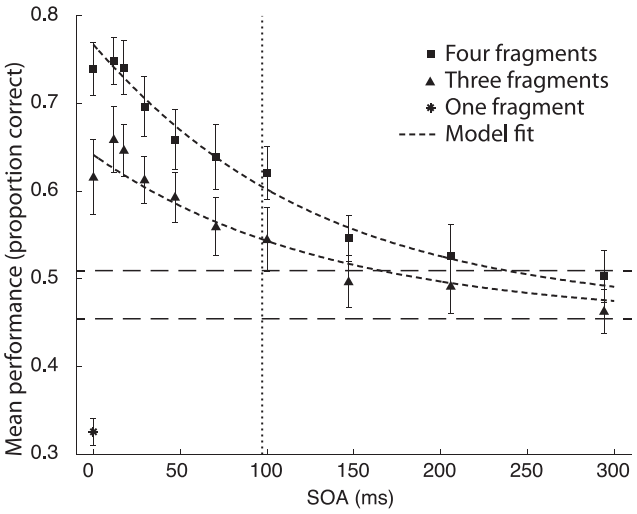

Figure 4. Summary of results from Experiment 1. Performance is averaged across $n = 16$ subjects (individual performance shown in Figure 3). Conventions are the same as in Figure 3; horizontal dashed lines indicate the performance level predicted by probability summation, based on single-fragment performance. Error bars show *SEM* across subjects.

$= 294.1$ ms: 0.46 ± 0.03 observed versus 0.44 ± 0.04 predicted, and 0.50 ± 0.03 observed versus 0.49 ± 0.04 predicted, for three- and four-fragment conditions, respectively, Figure 4). Hence, independent probability summation is not a good model to describe overall performance in this task. That performance at all but the longest asynchronies exceeded the independent probability summation prediction despite the redundancy of the fragments within each trial suggests visual integration, which we quantify below.

Categorization performance showed a strong decrease with increasing values of SOA both for three-fragment trials and four-fragment trials (Figures 3–4). Performance declined with increasing temporal asynchrony for all 16 subjects (Figure 3). This change in performance was significant for 11 and 14 subjects in three-fragment and four-fragment trials, respectively (chi-squared test, $p < 0.05$). The performance decrement between simultaneous presentations and asynchronous trials first reached significance at an SOA of 29.4 ms (sign test, $p = 0.05$), with a difference of 2.3% ± 1.4% (Figure 4). While a slight improvement in performance was apparent in the mean values at 11.8 ms SOA relative to simultaneous presentations, it was not statistically significant (sign test, $p = 0.60$). These observations show that temporal imprecision of even 30 ms disrupts recognition performance.

To quantify performance (y) as a function of SOA (denoted A below), we started by performing a least-squares fit of an exponential curve to the data from each subject for each number of fragments:

$$\hat{y}(A) = B + R \cdot 2^{-A/\tau}. \quad (3)$$

The three parameters were the time constant τ, the baseline value $B = \hat{y}(\infty)$, and the range in performance $R = \hat{y}(0) - \hat{y}(\infty)$, where $\hat{y}(0)$ represents the performance at 0 ms SOA and $\hat{y}(\infty)$ denotes the asymptotic performance at infinite SOA. We adjusted these three parameters (B, R, τ) to minimize the squared error between $\hat{y}(A)$ and the observed data $y(A)$. The exponential curves provided a reasonable fit to the data (root-mean-squared error, *RMSE*, of 0.058 ± 0.004 and 0.045 ± 0.003, for three-fragment and four-fragment data, respectively). As a measure of the window of integration, we considered τ, the SOA at which the exponential curve achieved its half height. The mean value of τ across 16 subjects was 120 ± 24 ms and 104 ± 19 ms, in three-fragment and four-fragment trials, respectively. Thus, shape integration at spatial scales of at least approximately 5° remains evident over a time scale on the order of 100 ms.

The exponential fits described the shapes of each of the response curves and experimental parameters separately. We sought a unifying model that could capture all the different experimental conditions and could be realized by the firing rates of neurons. As discussed above, probability summation failed in assuming that multiple fragments contribute to perception independently. To include the interdependence among different fragments and to account for the increasing number of possible two-fragment interactions as fragment count increases, we developed a model based on the notion that pairs of fragments interact to reduce uncertainty about the image as an exponential function of their SOA. In this model, we described a subject's observed performance in terms of an underlying uncertainty u, which is reduced by interactions between presented fragments. As above, we assume that the subject actually knew the correct image category in some fraction i of the trials, and guessed at chance in the rest, so that $y(A) = i + [(1 - i)/4]$. The empirical uncertainty is then $1–i$, the chance that the subject did not know the image category. We assume that a second fragment reduces the underlying uncertainty u by some factor $f(A)$ that depends on the asynchrony A with which it is presented relative to the first. A third fragment reduces uncertainty by another factor of $f(A)$, and also by a factor of $f(2A)$, because it is presented A ms after the second fragment and $2A$ ms after the first fragment. We assume independence, so the total uncertainty with three fragments is $u \cdot f^2(A)$ $f(2A)$. Similarly, the total uncertainty in a four-fragment trial is $u \cdot f^3(A) \cdot f^2(2A) \cdot f(3A)$. If we denote the total uncertainty with a particular number of fragments as u', the predicted performance is then $1 - u' + (u'/4)$. Note that u is conceptually related to performance in single fragment trials but is not directly mathematically relatable: The value of u also incorpo-

rates the fact that different fragments from a single image are not independent but that this dependency is factored out in the consideration of pairs of fragments via $f(A)$.

For each subject, we used their empirical performance to calculate the empirical uncertainty $1–i$ at each SOA and each number of fragments. We then found the least-squares best fits for the uncertainty reduction function $f(A)$ and the underlying uncertainty u that was modified by $f(A)$. We considered for $f(A)$ exponential functions of the form $f(A) = B' - R' \cdot 2^{-A/\tau'}$. Here $B' = f(\infty)$ is the asymptote, which determines what might be called "cognitive" or "memory-based" integration, the ability to combine information from multiple fragments at an arbitrarily long SOA. $R' = f(\infty) - f(0)$ is the range of uncertainty reduction spanned by f, which describes the benefit of presenting visual information simultaneously rather than with very large SOA. Finally, τ' is the time constant, which describes the SOA at the uncertainty reduction function's half-height. Given the exponential nature of this model, there is some small amount of integration that persists at arbitrarily long asynchronies, decreasing asymptotically towards zero. Along with the underlying uncertainty u, this makes four total parameters to describe a subject's performance across all SOAs.

The uncertainty reduction model fit the data quite well: root-mean-square errors (*RMSE*) for the model (0.061 ± 0.004 and 0.046 ± 0.002, for three-fragment and four-fragment trials, respectively) were not significantly higher than the corresponding *RMSE* values for exponential fits to the raw data discussed above (0.058 ± 0.004 and 0.045 ± 0.003, respectively; Wilcoxon rank sum test comparing the two models, $p = 0.42$ and 0.61, respectively), even though the exponential fits to the raw data have two more free parameters. Given those extra parameters, the corrected Akaike Information Criterion (Akaike, 1974; Burnham & Anderson, 2002) is lower for the uncertainty reduction model in 15 out of 16 subjects; on average, the likelihood of the uncertainty reduction model is 3.74 ± 0.51 times higher than that of the exponential fits.

We described the window of integration with the uncertainty reduction model exponential's half height parameter τ'. When we fit the model to the data from individual subjects, we obtained a mean τ' of 97 ± 9 ms (Figure 3). In Figure 4, we also made use of the same framework, considering all subjects within the experiment together. Rather than fitting to each subject's performance separately, we fit one uncertainty reduction model to all subjects' data. The values of the best fit parameters for the uncertainty reduction model were $u = 0.43 \pm 0.02$, $B' = 0.99 \pm 0.01$, $R' = 0.15 \pm 0.02$, and $\tau' = 109$ ms. R^2 values between the data and the predictions for the population uncertainty reduction model were 0.92 for three-fragment trials and 0.97 for four-fragment trials. This model succinctly and accurately described a subject's performance across all conditions based solely on the reduction of uncertainty as a function of SOA. The model used fewer parameters to explain a subject's performance than the set of exponential fits, it could generalize across different numbers of fragments, and it suggested an underlying psychological interpretation.

We further considered a model based on signal detection theory, which assumes a probabilistic framework to describe each subject's performance in each trial. Previous analyses of psychophysics data (Swets, 1961) have shown that high-threshold models like the uncertainty reduction model just described sometimes do not capture subjects' behavior as well as models based on signal detection theory (SDT). We therefore constructed a SDT model to test whether it would give different results from the uncertainty reduction model. We used a Markov chain Monte Carlo (MCMC) method for calculating d′ (a SDT measure of sensitivity to stimulus identity) with bias in four-alternative forced choice tasks (DeCarlo, 2012). Models were fit using JAGS (http://sourceforge.net/projects/mcmc-jags/) with 10,000 iterations of burn-in followed by 20,000 iterations to approximate d′ distributions. Having calculated d′ for each subject, each asynchrony, and each number of fragments, we proceeded to fit a four-parameter model to each subject's data. Let g be a scaled base-two exponential function with parameter τ'' and scale factor R'', $g(A) = R'' \cdot 2^{-A/\tau''}$. Then g describes the increase in sensitivity brought about by having two fragments with an asynchrony of A. The third parameter is v'', the d′ value in single-fragment trials (constrained to lie within the 95% confidence interval estimated from the MCMC fit). Finally, we assumed that adding additional fragments to a trial might not aid sensitivity in a perfectly efficient fashion (because fragments contain correlated information) and we described this inefficiency with a multiplicative factor c'' that progressively reduced the contributions of successive fragments. The model predicts d′ for n fragments with asynchrony A as follows:

$$\mathrm{d}' = \sum_{i=1}^{n} \left(v'' \cdot c''^{i-1} + (i-1) \cdot g\Big(A \cdot (n-i+1)\Big)\right). \tag{4}$$

We fit the four parameters of this model using a least-squares procedure and compared the time constant values (τ'') obtained with the SDT model against those obtained from the uncertainty reduction model (τ'). The values of the best fits for the SDT model were $R'' = 0.21 \pm 0.02$, $v'' = 0.42 \pm 0.06$, $c'' = 0.47 \pm 0.09$, and a cross-subject mean τ'' of 101 ± 8 ms. Given that the SDT results were similar to those of the uncertainty reduction model and the higher computational re-

quirements for the SDT model, we used only the uncertainty reduction model for the remainder of our analyses.

As noted above, performance differed slightly between categories. To evaluate whether such inhomogeneity contributed to the observed integration time constants, we repeated the analyses after matching performance across categories by random subsampling. Under these performance-matched conditions, the mean τ' was 95 ± 9 ms, and the population τ' was 105 ms. We also analyzed each category and each subject separately after matching performance. The mean τ' for the four categories were 125 ± 23 ms, 144 ± 33 ms, 89 ± 21 ms, and 131 ± 26 ms, respectively, for animals, people, plants, and vehicles. Note that these fits were noisier due to having fewer data. While subjects could use different features to recognize images belonging to different categories, the results were similar across categories.

Experiment 2

We extended the first experiment in two variants aimed at evaluating two-fragment images and very long SOAs (Experiment 2A) and constant trial lengths and no fragment repetition (Experiment 2B). These experiment variants are described under Methods.

In Experiment 2A, the window of integration, τ', was 101 ± 14 ms when two-fragment trials were included in the fit, and 91 ± 13 ms when they were not included (Figure 5A). The other model parameters were $u = 0.33$ ± 0.02, $B' = 0.990$ ± 0.003, and $R' = 0.069$ ± 0.006. These values were not significantly different from each other or from the results in Experiment 1 (Wilcoxon rank sum test, $p = 0.55$ comparing these two fits, $p =$ 0.99 and 0.68, respectively, comparing Experiment 1 to these results with and without two-fragment trials). Furthermore, the model fit using only three- and four-fragment trials was a good fit for the two-fragment data (*RMSE* of 0.091 ± 0.006 using only three- and four-fragment trials versus 0.081 ± 0.004 using all trials; Wilcoxon rank sum test, $p = 0.20$). This demonstrates that the uncertainty reduction model was able to generalize to fragment counts that were not used to fit the model. While performance at 294.1 ms was slightly better than performance at 705.9 ms (differing by 3.4% ± 1.7%), this difference was not significant (sign test, $p = 0.16$). *RMSE* and τ' values were unaffected by including the data from 705.9 ms trials (Wilcoxon rank sum tests, $p = 0.97$ and 0.92, respectively). R^2 values were 0.30, 0.75, and 0.78 for two-, three-, and four-fragment trials, respectively. The two-fragment R^2 value was low because the plot of performance against asynchrony was almost flat. The results of this experimental variant suggest that performance at long

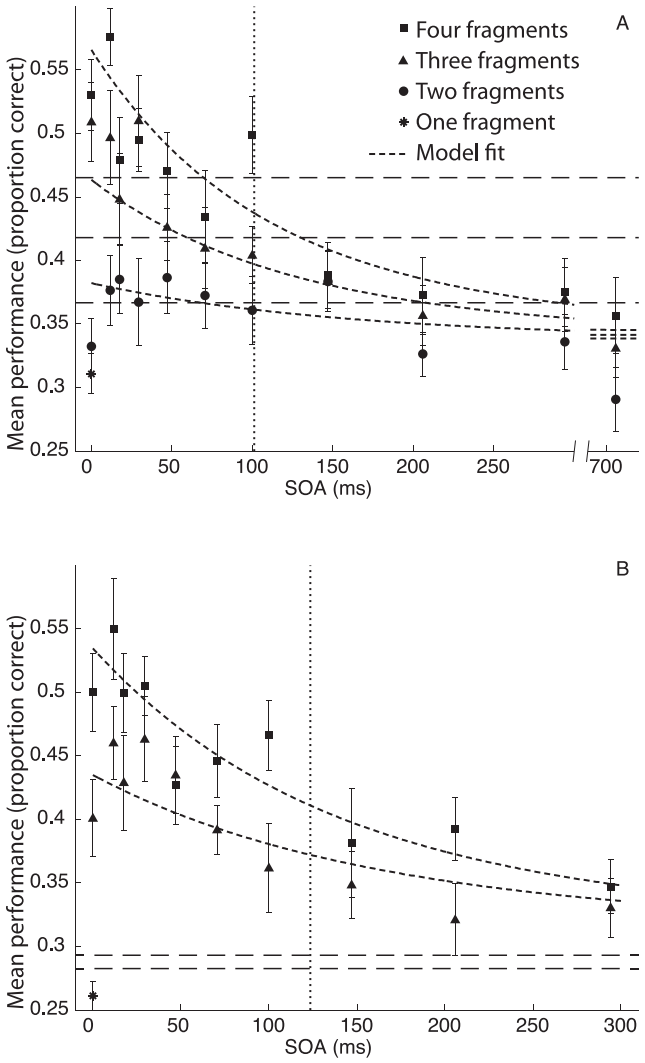

Figure 5. Mean performance across all subjects in Experiment 2. **A**. Variant A ($n = 17$ subjects), which added a two-fragment condition (circles) and data at 705.9 ms SOA. **B**. Variant B ($n =$ 17 subjects), which controlled for influence of trial duration and stimulus repetition (see Results and Methods). Conventions are otherwise as in Figure 4.

SOAs in Experiment 1 was close to asymptote and that the uncertainty reduction model can extrapolate to two-fragment trials.

The overall performance in Experiment 2B (as well as in the previous variant) was lower than in Experiment 1 (cf. Figure 5 vs. Figure 4) due to the inclusion of more difficult line drawings in the image sets (Figure 5B). Considering only the trials that included the same types of images across experiments (silhouette images) raised performance to levels comparable to the ones in Experiment 1 without significantly changing τ'. R^2 values for the uncertainty reduction model were 0.64 and 0.81 for three- and four-fragment trials, respectively. The integration window from the uncertainty reduction model in Experiment 2B

($\tau' = 123 \pm 20$ ms) was indistinguishable from that obtained in Experiment 1 (Wilcoxon rank sum test, $p = 0.79$). The other model parameters were $u = 0.30 \pm 0.02$, $B' = 0.990 \pm 0.003$, and $R' = 0.064 \pm 0.009$. The similar τ' values suggest that the results of Experiment 1 cannot be ascribed to differences in trial duration or slight familiarity with the images shown.

Discussion

We measured the performance of 50 subjects in different variants of an experiment to evaluate the effect of presenting object fragments with varying degrees of asynchrony. By applying this externally induced temporal jitter, we characterized the importance of temporal precision for visual object recognition.

Subjects might in principle have made educated guesses based on the content of individual fragments, or combined such guesses independently across multiple fragments. Instead, we found that performance was significantly higher than predicted by independent probability summation, indicating that information from different fragments was synergistically integrated. Recognition performance decreased with increasing SOA and was well characterized by a simple model based on the reduction of uncertainty as a function of the temporal proximity of the presented image parts. In principle, such a model could be realized by neurons whose firing rate increases with the arrival of each new burst of information and then relaxes more slowly towards baseline. Neurons showing such firing rate increases upon transient stimulus onset, and decay dynamics over hundreds of milliseconds, have been described in recordings in the lateral intraparietal cortex during a motion discrimination task (Huk & Shadlen, 2005).

While we focused on the uncertainty reduction model, we also fit the data from Experiment 1 with a model based on signal detection theory. These two models reflect different paradigms of psychological processing. In the former, each additional piece of information (provided by each successive fragment) multiplicatively reduces the subject's uncertainty about what was shown, increasing the probability of giving a correct answer. In the latter, each additional fragment adds to the distance between the actual category and the other categories in an implicit psychological space, again increasing the probability of discriminating the true category from the others. There are important differences between these paradigms (Swets, 1961), though either could be realized via the integrative neurons hypothesized above. The uncertainty reduction model might be implemented by a set of downstream neurons that act as thresholds on the evidence accumulated by the integrative neurons; the first downstream neuron to cross threshold could report the identity of the viewed image. Support for one or another image would manifest as increased firing rates in integrative neurons feeding into the associated downstream neurons, increasing its chance of being the first to cross threshold. The signal detection theory might instead reflect a set of downstream neurons whose activity reflected which integrative neuron was most active, in a winner-take-all fashion. The integrative neurons would then essentially be counting votes. Ultimately, given that the two models yielded indistinguishable estimates of the window of temporal integration, it is difficult to draw any conclusions about which paradigm is more likely to be correct based on the current data.

The results demonstrated that asynchronies of a few tens of milliseconds disrupted categorization performance. This disruption is consistent with a body of neurophysiological observations and computational models of object recognition that are based on a rapid cascade of processing events along the ventral visual stream (DiCarlo et al., 2012; Fukushima, 1980; Hung et al., 2005; Johnson & Olshausen, 2003; Kirchner & Thorpe, 2006; Liu et al., 2009; Potter & Levy, 1969; Richmond et al., 1990; Riesenhuber & Poggio, 2000; Rolls, 1991; Serre et al., 2007; Thorpe et al., 1996; vanRullen & Thorpe, 2002). The response latencies across the ventral stream consistently increase by about 10 to 20 ms at each stage from the retina to the thalamus to a cascade of cortical areas from primary visual cortex to inferior temporal cortex (Maunsell, 1987; Schmolesky et al., 1998). The rapidity with which visual shape information progresses through the ventral visual cortex and is converted into behavioral or perceptual output is consistent with the current observation that object recognition begins to suffer impairment with even minor disruptions of stimulus timing.

In addition to the impairment in shape recognition through spatial integration due to temporal asynchrony, timing judgments can be influenced by spatial cues. Spatial grouping cues that encourage the binding together of separate parts can interfere with simultaneity judgments, at a time scale similar to that reported here (Cheadle et al., 2008). Spatial judgments are influenced by temporal constraints and temporal judgments are influenced by spatial constraints.

While temporal asynchrony disrupted recognition, we observed that performance was well above asymptotic levels even at SOAs of approximately 100 ms. Integration at these SOAs cannot be ascribed to trial duration, eye movements, remembering the stimuli from previous trials, or forgetting the fragments shown within a trial (Figure 5). Performance values at SOAs

of approximately 100 ms are consistent with the extended durations of responses in inferior temporal cortex (De Baene et al., 2007). These extended response durations might instantiate a buffer, allowing the integration of visual shape information over time. The ability to integrate visual information over brief spans could underlie many critical visual functions. For example, percepts of camouflaged or partially occluded objects could be assembled over time as the objects moved relative to their background or occluders (Nishida, 2004) and information could be integrated across saccades (Irwin, Yantis, & Jonides, 1983; Jonides, Irwin, & Yantis, 1982; O'Regan & Levy-Schoen, 1983).

The observed integration of form information over < 100 ms is unlikely to reflect working memory processes, as the temporal scales involved in working memory span several seconds (Vogel, Woodman, & Luck, 2006). There is, however, sufficient time between stimulus presentation and the response period that the (integrated) image or category information could be held in working memory. The window of temporal integration reported here has been observed in other domains and with other experimental paradigms (Caudek, Domini, & Di Luca, 2002; Forget, Buiatti, & Dehaene, 2009; Nishida, 2004). In particular, performance that exceeds probability summation at short SOAs is consistent with the temporal scales involved in iconic memory (Coltheart, 1980; Di Lollo, 1977; Eriksen & Collins, 1968; Hogben & Di Lollo, 1974; Sperling, 1960). Prior work has shown that information about arrays of letters (Sperling, 1960) or flashing squares (Hogben & Di Lollo, 1974) remains accessible for a short time after presentation. It is possible that the integration we observed at short SOAs in the object recognition system arises from mechanisms similar to those underlying the previously reported persistence of simple visual "icons," though iconic memory has been shown to be disrupted by masks (Di Lollo, Clark, & Hogben, 1988). Visual persistence, as instantiated by the dynamics of retinal cells, is generally terminated by a mask; it is more likely that we are measuring neural or informational persistence at a higher level of the visual system (Coltheart, 1980). While the dynamic noise in the stimulus sequence could partly mask visual persistence, it certainly does not completely obliterate it. Hence, the integration constants reported here constitute an upper bound and it is conceivable that visual information could decay even more rapidly than reported here.

Two additional phenomena share similar characteristics with the dynamic rate of decay reported here. Integration masking, in which irrelevant information presented immediately before or after the presentation of a target inhibits perception of that target by adding noise to it, exhibits a similar time scale to that described here, and may reflect a similar process (Enns & Di Lollo, 2000). Priming (in which a stimulus facilitates the perception or naming of a later stimulus) might also play a part in these observations. Priming would likely increase performance at longer SOA; positive effects of priming typically are small or nonexistent at SOA close to zero and increase with increasing SOA (La Heij, Dirkx, & Kramer, 1990; Scharlau & Neumann, 2003).

A series of elegant studies has shown that detectability and perception can be influenced by the exact time at which a stimulus is presented with respect to ongoing endogenous oscillations (Rohenkohl & Nobre, 2011; Van Dijk, Schoffelen, Oostenveld, & Jensen, 2008). In the context of the task presented here, it is conceivable that the initial dynamic noise could reset such endogenous oscillations and that certain presentation times and SOAs could lead to enhanced recognition by virtue of their specific phase. The limited number and uneven sampling of SOAs precludes a systematic investigation of these effects here, but it will be interesting in future studies to examine the physiological signals underlying recognition of asynchronously presented objects.

Several studies have shown interactions between asynchronous parts of faces (Anaki et al., 2007; Cheung et al., 2011; Singer & Sheinberg, 2006); those findings are better described in terms of higher level phenomena than object recognition. Complex images have also been used in a study of subjective simultaneity (Loftus & Hanna, 1989), in which subjects were asked to rate how complete or integrated images appeared to be when divided into two parts presented with varying durations and asynchronies. Here we expand upon this body of previous work by showing that complex shape information can be combined across both space and time to enable object recognition. Our results were consistent across a battery of controls designed to account for eye movements, stimulus familiarity, and memory on longer time scales.

Conclusions

Integration across image fragments begins to break down when as little as 30 ms separates the fragments and decays with a time constant of approximately 100 ms. The current characterization of the time course with which visual object recognition breaks down as a function of temporal asynchrony provides information critical to building computer vision systems that operate beyond static snapshots, to modeling visual object recognition, to designing experiments that probe the performance over time of the human visual system, and to the construction of theories of perception in a dynamic world.

Keywords: object recognition, temporal integration, fragmented images, temporal sensitivity, visual dynamics

Acknowledgments

The authors thank Neal Dach and Joanna Li for assistance with stimulus preparation and data collection and John Maunsell, Thomas Miconi, and Hanlin Tang for their insightful comments on the manuscript. This work was supported by NSF grants 0954570 and CCF-1231216 (GK).

Commercial relationships: none.
Corresponding author: Jedediah Miller Singer.
Email: jedediah.singer@childrens.harvard.edu.
Address: Department of Ophthalmology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA.

References

Akaike, H. A. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–723.

Anaki, D., Boyd, J., & Moscovitch, M. (2007). Temporal integration in face perception: Evidence of configural processing of temporally separated face parts. *Journal of Experimental Psychology: Human Perception and Performance, 33*(1), 1–19. doi: 2007-01135-001 [pii] 10.1037/0096-1523.33.1.1.

Aspell, J. E., Wattam-Bell, J., & Braddick, O. (2006). Interaction of spatial and temporal integration in global form processing. *Vision Research*, *46*(18), 2834–2841. doi: S0042-6989(06)00118-0 [pii] 10.1016/j.visres.2006.02.018.

Bair, W., & Movshon, J. A. (2004). Adaptive temporal integration of motion in direction-selective neurons in macaque visual cortex. *Journal of Neuroscience*, *24*(33), 7305–7323. doi:10.1523/JNEUROSCI.0554-04.2004 24/33/9305 [pii].

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436.

Brockmole, J. R., Wang, R. F., & Irwin, D. E. (2002). Temporal integration between visual images and visual percepts. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(2), 315–334.

Burnham, K. P, & Anderson, D. (2002). *Model selection and multi-model inference* (2nd ed.). New York: Springer.

Burr, D. C., & Santoro, L. (2001). Temporal integration of optic flow, measured by contrast and coherence thresholds. *Vision Research, 41*(15), 1891–1899. doi: S0042-6989(01)00072-4 [pii].

Caudek, C., Domini, F., & Di Luca, M. (2002). Short-term temporal recruitment in structure from motion. *Vision Research, 42,* 10.

Cavanagh, P., Holcombe, A. O., & Chou, W. (2008). Mobile computation: Spatiotemporal integration of the properties of objects in motion. *Journal of Vision, 8*(12):1, 1–23, http://www.journalofvision.org/content/8/12/1, doi:10.1167/8.12.1. [PubMed] [Article]

Cheadle, S., Bauer, F., Parton, A., Müller, H., Bonneh, Y. S., & Usher, M. (2008). Spatial structure affects temporal judgments: Evidence for a synchrony binding code. *Journal of Vision, 8*(7):12, 1–12, http://journalofvision.org/content/8/7/12, doi: 10.1167/8.7.12. [PubMed] [Article]

Cheung, O. S., Richler, J. J., Phillips, W. S., & Gauthier, I. (2011). Does temporal integration of face parts reflect holistic processing? *Psychonomic Bulletin and Review, 18*(3), 476–483. doi:10.3758/s13423-011-0051-7.

Clifford, C. W., Holcombe, A. O., & Pearson, J. (2004). Rapid global form binding with loss of associated colors. *Journal of Vision*, *4*(12):8, 1090–1101, http://www.journalofvision.org/content/4/12/8, doi: 10.1167/4.12.8. [PubMed] [Article]

Coltheart, M. (1980). Iconic memory and visible persistence. *Attention, Perception, & Psychophysics, 27*(3), 183–228. doi:10.3758/bf03204258.

Connor, C. E., Brincat, S. L., & Pasupathy, A. (2007). Transformation of shape information in the ventral pathway. *Current Opinion in Neurobiology*, *17*(2), 140–147.

Cornelissen, F. W., Peters, E. M., & Palmer, J. (2002). The Eyelink Toolbox: Eye tracking with MATLAB and the Psychophysics Toolbox. *Behavioral Research Methods, Instruments, and Computers, 34,* 4.

De Baene, W., Premereur, E., & Vogels, R. (2007). Properties of shape tuning of macaque inferior temporal neurons examined using rapid serial visual presentation. *Journal of Neurophysiology*, *97*(4), 2900–2916. doi: 00741.2006 [pii] 10.1152/jn.00741.2006.

DeCarlo, L. T. (2012). On a signal detection approach to m-alternative forced choice with bias, with maximum likelihood and bayesian approaches to estimation. *Journal of Mathematical Psychology, 56*(3), 196–207.

Deco, G., & Rolls, E. T. (2004). *Computational*

neuroscience of vision. Oxford, UK: Oxford University Press.

Di Lollo, V. (1977). Temporal characteristics of iconic memory. *Nature*, *267*(5608), 241–243.

Di Lollo, V., Clark, C. D., & Hogben, J. H. (1988). Separating visible persistence from retinal afterimages. *Perception and Psychophysics*, *44*(4), 363–368.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434. doi: S0896-6273(12)00092-X [pii] 10.1016/j.neuron.2012.01.010.

Douglas, R. J., & Martin, K. A. (2004). Neuronal circuits of the neocortex. *Annual Review of Neuroscience, 27,* 419–451.

Enns, J. T., & Di Lollo, V. (2000). What's new in visual masking? *Trends in Cognitive Sciences, 4*(9), 345–352.

Eriksen, C. W., & Collins, J. F. (1968). Sensory traces versus the psychological moment in the temporal organization of form. *Journal of Experimental Psychology*, *77*(3), 376–382.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1,* 1–47.

Forget, J., Buiatti, M., & Dehaene, S. (2009). Temporal integration in visual word recognition. *Journal of Cognitive Neuroscience, 22*(5), 14.

Fukushima, K. (1980). Neocognitron: A self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36*(4), 193–202.

Hogben, J. H., & Di Lollo, V. (1974). Perceptual integration and perceptual segregation of brief visual stimuli. *Vision Research*, *14*(11), 1059–1069.

Huk, A. C., & Shadlen, M. N. (2005). Neural activity in macaque parietal cortex reflects temporal integration of visual motion signals during perceptual decision making. *The Journal of Neuroscience, 25*(45), 16.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast read-out of object identity from macaque inferior temporal cortex. *Science, 310,* 863–866.

Irwin, D. E., Yantis, S., & Jonides, J. (1983). Evidence against visual integration across saccadic eye movements. *Perception and Psychophysics*, *34*(1), 49–57.

Johnson, J. S., & Olshausen, B. A. (2003). Timecourse of neural signatures of object recognition. *Journal of Vision*, *3*(7):4, 499–512, http://www.journalofvision.org/content/3/7/4, doi:10.1167/3.7.4. [PubMed] [Article]

Jonides, J., Irwin, D. E., & Yantis, S. (1982). Integrating visual information from successive fixations. *Science*, *215*(4529), 192–194.

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research, 46*(11), 1762–1776. doi: S0042-6989(05)00511-0 [pii] 10.1016/j.visres.2005.10.002.

La Heij, W., Dirkx, J., & Kramer, P. (1990). Categorical interference and associative priming in picture naming. *British Journal of Psychology*, *81*(4), 511–525.

Liu, H., Agam, Y., Madsen, J. R., & Kreiman, G. (2009). Timing, timing, timing: Fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron, 62*(2), 281–290.

Loftus, G. R., & Hanna, A. M. (1989). The phenomenology of spatial integration: Data and models. *Cognitive Psychology, 21*(3), 363–397. doi: 0010-0285(89)90013-3 [pii].

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience, 19,* 577–621.

Maunsell, J. H. R. (1987). Physiological evidence for two visual subsystems. In L. Vaina (Ed.), *Matters of intelligence* (pp. 59–87). Dordrecht, The Netherlands: Reidel Press.

Neri, P., Morrone, M. C., & Burr, D. C. (1998). Seeing biological motion. *Nature, 395*(6705), 894–896. doi: 10.1038/27661.

Nishida, S. (2004). Motion-based analysis of spatial patterns by the human visual system. *Current Biology, 14,* 830–839.

O'Regan, J. K., & Levy-Schoen, A. (1983). Integrating visual information from successive fixations: Does trans-saccadic fusion exist? *Vision Research, 23*(8), 765–768.

Pasupathy, A., & Connor, C. E. (1999). Responses to contour features in macaque area V4. *Journal of Neurophysiology, 82*(5), 2490–2502.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442.

Potter, M. C., & Levy, E. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology*, *81*(1), 10–15.

Richmond, B. J., Optican, L. M., & Spitzer, H. (1990). Temporal encoding of two-dimensional patterns by single units in primate primary visual cortex. I. Stimulus-response relations. *Journal of Neurophysiology, 64*(2), 351–369.

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience*, *3*(Suppl), 1199–1204.

Ringach, D. L., Hawken, M. J., & Shapley, R. (2003). Dynamics of orientation tuning in macaque V1: The role of global and tuned suppression. *Journal of Neurophysiology, 90*(1), 342–352. doi:10.1152/jn.01018.2002 01018.2002 [pii].

Rohenkohl, G., & Nobre, A. C. (2011). Alpha oscillations related to anticipatory attention follow temporal expectations. *The Journal of Neuroscience*, *31*(40), 14076–14084.

Rolls, E. T. (1991). Neural organization of higher visual functions. *Current Opinion in Neurobiology, 1,* 274–278.

Sanocki, T. (2001). Interaction of scale and time during object identification. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(2), 290–302.

Scharlau, I., & Neumann, O. (2003). Temporal parameters and time course of perceptual latency priming. *Acta Psychologica, 113*(2), 185–203.

Schmolesky, M. T., Wang, Y. C., Hanes, D. P., Thompson, K. G., Leutgeb, S., Schall, J. D., ... Leventhal, A. G. (1998). Signal timing across the macaque visual system. *Journal of Neurophysiology, 79*(6), 3272–3278.

Schyns, P. G., Petro, L. S., & Smith, M. L. (2007). Dynamics of visual information integration in the brain for categorizing facial expressions. *Current Biology*, *17*(18), 1580–1585. doi: S0960-9822(07)01904-5 [pii] 10.1016/j.cub.2007.08.048.

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., & Poggio, T. (2007). A quantitative theory of immediate visual recognition. *Progress in Brain Research, 165C,* 33–56.

Singer, J. M., & Sheinberg, D. L. (2006). Holistic processing unites face parts across time. *Vision Research*, *46*(11), 1838–1847. doi: S0042-6989(05)00563-8 [pii] 10.1016/j.visres.2005.11.005.

Sperling, G. (1960). *The information available in brief visual presentations*. Washington, DC: American Psychological Association.

Swets, J. A. (1961). Is there a sensory threshold? *Science*, *134*(3473), 168–177.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience, 19,* 109–139.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381,* 520–522.

Van Dijk, H., Schoffelen, J. M., Oostenveld, R., & Jensen, O. (2008). Prestimulus oscillatory activity in the alpha band predicts visual discrimination ability. *The Journal of Neuroscience*, *28*(8), 1816–1823.

vanRullen, R., & Thorpe, S. J. (2002). Surfing a spike wave down the ventral stream. *Vision Research, 42,* 2593–2615.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2006). The time course of consolidation in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 32*(6), 1436–1451. doi:10.1037/0096-1523.32.6.1436.