A bottom–up model of spatial attention predicts human error patterns in rapid scene recognition

Wolfgang Einhäuser
Division of Biology, California Institute of Technology,
Pasadena CA, USA

T. Nathan Mundhenk
Department of Computer Science,
University of Southern California,
Los Angeles, CA, USA

Pierre Baldi
School of Information and Computer Sciences,
University of California in Irvine,
Irvine, CA, USA

Christof Koch
Division of Biology, California Institute of Technology,
Pasadena CA, USA

Laurent Itti
Department of Computer Science,
University of Southern California,
Los Angeles, CA, USA

Humans demonstrate a peculiar ability to detect complex targets in rapidly presented natural scenes. Recent studies suggest that (nearly) no focal attention is required for overall performance in such tasks. Little is known, however, of how detection performance varies from trial to trial and which stages in the processing hierarchy limit performance: bottom–up visual processing (attentional selection and/or recognition) or top–down factors (e.g., decision-making, memory, or alertness fluctuations)? To investigate the relative contribution of these factors, eight human observers performed an animal detection task in natural scenes presented at 20 Hz. Trial-by-trial performance was highly consistent across observers, far exceeding the prediction of independent errors. This consistency demonstrates that performance is not primarily limited by idiosyncratic factors but by visual processing. Two statistical stimulus properties, contrast variation in the target image and the information-theoretical measure of "surprise" in adjacent images, predict performance on a trial-by-trial basis. These measures are tightly related to spatial attention, demonstrating that spatial attention and rapid target detection share common mechanisms. To isolate the causal contribution of the surprise measure, eight additional observers performed the animal detection task in sequences that were reordered versions of those all subjects had correctly recognized in the first experiment. Reordering increased surprise before and/or after the target while keeping the target and distractors themselves unchanged. Surprise enhancement impaired target detection in all observers. Consequently, and contrary to several previously published findings, our results demonstrate that attentional limitations, rather than target recognition alone, affect the detection of targets in rapidly presented visual sequences.

Keywords: psychophysics, modeling, attention, saliency, RSVP

Citation: Einhäuser, W., Mundhenk, T. N., Baldi, P., Koch, C., & Itti, L. (2007). A bottom–up model of spatial attention predicts human error patterns in rapid scene recognition. *Journal of Vision, 7*(10):6, 1–13, http://journalofvision.org/7/10/6/, doi:10.1167/7.10.6.

Introduction

Humans and other primates grasp the "gist" of a complex natural scene even when presented for only a few tens of milliseconds (Biederman, 1981; Evans & Treisman, 2005; Fabre-Thorpe, Richard, & Thorpe, 1998; Li, VanRullen, Koch, & Perona, 2002; Potter & Levy, 1969; Rousselet, Fabre-Thorpe, & Thorpe, 2002; Thorpe, Fize, & Marlot, 1996; VanRullen & Thorpe, 2001). Furthermore, observers can detect with above-chance performance complex target items (such as an animal) in rapidly presented image sequences (rapid serial visual presentation [RSVP]; Evans & Treisman, 2005; Potter & Levy, 1969). Such performance is typically seen as evidence for a rapid, sensory-driven ("bottom–up") mode of processing, primarily driven by the visual stimulus. This leads to the hypothesis that properties of the stimulus, rather than observer-specific and possibly more idiosyncratic top–down processes, may, to a large extent, determine performance in RSVP. If so, what are these statistical properties?

doi: 10.1167/7.10.6 Received December 3, 2006; published June 20, 2007 ISSN 1534-7362 © ARVO

It has been argued that rapid recognition requires little or no focal *spatial* attention (Li et al., 2002; Rousselet et al., 2002). According to this view, bottom–up attention does not constitute the primary limit for rapid visual processing, but rather, such a limit is found in a later target recognition stage. Indeed, some aspects of overall performance can be captured by models of object recognition; for example, animals that appear farther away are more difficult to detect on average (Serre, Oliva, & Poggio, 2006). However, these studies typically use isolated images followed by masking stimuli. Contrary to these results, when using a stream of images, some of which are targets and most of which act as distractors, one finds two attentional phenomena that limit rapid processing: When two identical items are presented in direct succession, often only one is detected ("repetition blindness"; Kanswisher, 1987), and when a second target item is presented shortly—but not immediately—after a first one, its processing is also impaired ("attentional blink"; Raymond, Shapiro, & Arnell, 1992). Although repetition blindness and attentional blink are distinct phenomena (Chun, 1997), models of such attentional impairments are typically variants of an attentional gating model, as first formalized by Reeves and Sperling (1986): In this view, a salient item (e.g., a target) opens an "attentional gate" for its and subsequent items' access to visual short-term memory. Failure to quickly reopen the gate impairs the detection of the second target in attentional blink; furthermore, integration of information according to order and strength within an open gate epoch leads to the loss of order information, a potential cause for repetition blindness. In attentional blink, the saliency of an item to open a gate arises from its property of being a target or semantically related to the target (Barnard, Scott, Taylor, May, & Knightley, 2004). Items that attract attention because of their emotional content can also lead to an attentional-blink-like recognition impairment, which some, but not all, observers can overcome through volitional control (Most, Chun, Widders, & Zald, 2005). Similarly, odd items (e.g., the rare occurrence of a face in a letter task or vice versa) can impair subsequent processing (Marois, Todd, & Gilbert, 2003), as can items that are visually similar to the target but appear at peripheral locations (Folk, Leber, & Egeth, 2002). However, very little is known quantitatively of the neural mechanisms by which some items may strongly capture attention and create an attentional-blink-like effect.

We hypothesize that high stimulus-driven saliency can impair the detection of subsequent targets. In the view of attentional gating, the high-saliency item triggers gate opening and therefore impairs subsequent detection. Similarly, a highly salient item can compete with a target if the target itself triggered gate opening. This predicts that distractors of high bottom–up saliency can cause attentional impairments in the detection of adjacent targets in RSVP, akin to attentional blink or repetition blindness. This implies that a model of spatial attention ought to predict human error patterns, in a trial-by-trial manner, in such RSVP sequences. Furthermore, we hypothesize that target detection is primarily impaired by these attentional limits rather than by target recognition itself. This hypothesis predicts differential detection performance for identical targets embedded in differently ordered RSVP sequences.

To test these hypotheses, we measure human performance in a 20-Hz RSVP animal/no-animal detection task. We first test whether observers' success and error patterns are idiosyncratic or stereotypical, that is, whether different subjects tend to make their errors in the same sequences. To test whether attentional mechanisms underlie consistent performance, we use a model of Bayesian "surprise" (Baldi, 2005; Itti & Baldi, 2005), which has previously been shown to model the distribution of spatial attention (Itti & Baldi, 2006), to predict human performance in the RSVP task. To distinguish whether attention or recognition mechanisms primarily determine detection performance, we use the model's prediction to design a second experiment: We reorder sequences, in which target detection was successful, as to increase attention load by placing distractors adjacent to the target that create higher surprise according to the model. As this procedure keeps target and distractor items themselves unchanged and only modifies surprise, it tests the causal effect of surprise on recognition performance. Using this setting, we demonstrate that surprise causally impairs recognition performance, and attentional limitations, therefore, are a major cause of human errors in RSVP.

Methods

Observers

Eight volunteers from the Caltech community participated in each experiment (age range: Experiment 1, 18–26 years; Experiment 2, 18–23 years). All participants had normal or corrected-to-normal vision and gave written informed consent to participation. All procedures conformed to National and Institutional Guidelines for experiments in human subjects and with the Declaration of Helsinki.

Stimuli and setup

All stimuli (targets and distractors) were based on a data set used previously for RSVP tasks (Evans & Treisman, 2005; Li et al., 2002), in the form provided on the Web page of Li et al. (2002; http://visionlab.ece.uiuc.edu/datasets.html). The target set consists of 1,323 pictures of a variety of animals occurring at different scales, viewing angles, and positions within the images, of which we

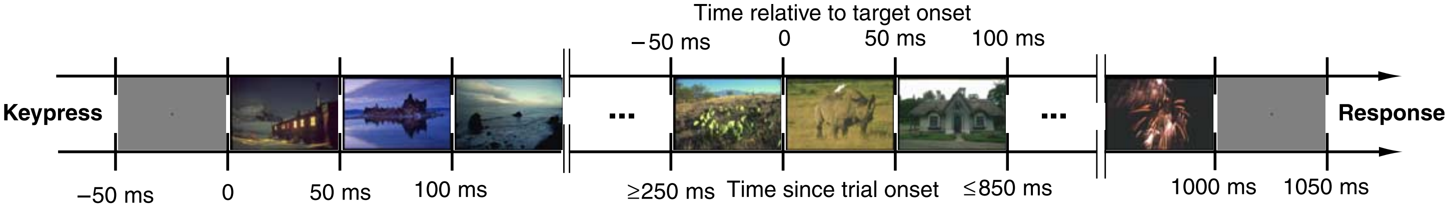


Figure 1. Stimuli and paradigm. Sample sequence. An observer starts a trial with a key press. In target-present sequences, the target (animal) can occur between the 6th and the 15th frame, that is, between 250 and 850 ms after trial onset.

randomly selected 500. The distractor set consists of 1,123 various outdoor scenes, none of which contains an animal. (Observers were explicitly instructed that humans do not constitute an "animal" in the context of the experiment.)

Stimuli were presented on a 19-in. computer screen of resolution 1,024 × 768 at 120 Hz using MatLab (Mathworks, Natick, MA) and its Psychophysics Toolbox extension (Brainard, 1997; Pelli, 1997; http://psychtoolbox.org/). Ambient light level was below 0.01 cd/m^2, with peak screen luminance at 29 cd/m^2. A chin rest stabilized viewing distance at 80 cm. Stimuli spanned 384 × 256 pixels, corresponding to 11° × 7° of visual angle.

Experiment 1

We created 1,000 image sequences of 20 frames each. Of those sequences, 500 ("target-present" sequences) contained a single target image (animal) between Frames 6 and 15, with 50 sequences having the target at Frame 6, 50 at Frame 7, and so forth (Figure 1). The remaining 500 sequences did not contain animal images but only distractor images. No target stimulus was used in more than one sequence, and no distractor was used more than once in the same sequence. The same sequences were used for all observers, whereas the order of sequences was randomized individually. In each trial, a single sequence was presented at 20 Hz (1 s total duration). Each sequence was preceded and followed by a black fixation cross on a gray rectangle presented for 50 ms. Observers started each trial by a button press and were allowed to take breaks as needed. In each trial, we asked observers to respond "as accurately as possible and as fast as possible without sacrificing accuracy" after the end of the sequence, whether an animal had been presented or not. Immediate feedback was provided on the correctness of the decision.

Experiment 2

For Experiment 2, we selected the 122 target-present sequences that were reported correctly by all eight observers in Experiment 1 ("easy" sequences). As detailed below, we rearranged the order of frames in each sequence to increase the surprise of the frame preceding the target, succeeding the target, or both, resulting in four conditions per sequence ("original", "pre", "post", and "both"). The target frame remained in the same temporal location in all conditions. In addition to those 488 (4 × 122) sequences, we used 488 "target-absent" sequences, which we selected at random from the 500 target-absent sequences. As in Experiment 1, the order of sequences was randomized for each individual subject. All instructions and setup were also identical to those of Experiment 1.

Behavioral analysis

Prediction for independent observers

Any statistical measure that does not take idiosyncratic information into account can maximally predict performance patterns that are consistent across different observers. At the other extreme, for any given target-present sequence, an observer's probability to correctly report the target would be independent of the properties of the sequence and, thus, of the other observers. If this were the case and all n observers had the same probability p_{hit} to correctly report the target, the probability $P_{\text{hit}}(k)$ that k observers correctly report a given sequence would be given by the binomial probability distribution $P_{\text{hit}}(k) = \binom{n}{k} p_{\text{hit}}^{k}(1 - p_{\text{hit}})^{(n-k)}$. Because—due to the "yes/no" experimental design—an individual's response depends on his or her criterion to trade off misses versus false alarms, p_{hit} will, in general, take different values for different observers. Denoting the probability that an observer j correctly reports a target by $p_{\text{hit},j}$ and assuming errors to be independent across observers, the probability of a sequence to be reported correctly by exactly k out of n observers is given by

$$\begin{aligned} P_{\text{hit}}(k) &= \sum_{V\in\Omega_k^n}\prod_{j\in V} p_{\text{hit},j}\prod_{j\in\bar{V}} p_{\text{miss},j} \\ &= \sum_{V\in\Omega_k^n}\prod_{j\in V} p_{\text{hit},j}\prod_{j\in\bar{V}}\left(1 - p_{\text{hit},j}\right), \end{aligned} \tag{1}$$

where Ω_k^n denotes the set of all k-element subsets of $\{1, \ldots, n\}$, V is one particular subset, and $\bar{V}$ is the complement of V with respect to $\{1, \ldots, n\}$. (Obviously, this expression collapses to the binomial distribution if all $p_{\mathrm{hit},j}$ were equal.) As we directly measure $p_{\mathrm{hit},j}$ as the fraction of correct responses in target-present trials for each individual, we can explicitly compute the prediction $P_{\mathrm{hit}}(k)$ for the assumption of independent observers. Multiplying $P_{\mathrm{hit}}(k)$ with the number of trials (500) results in the prediction depicted in gray in Figure 3.

Contrast variation

As a static correlate of spatial attention, we measured contrast as typically used in saliency-map algorithms (Itti & Koch, 2000) and eye-movement studies of natural scenes (Reinagel & Zador, 1999): For each 16 × 16 patch of the stimulus, we define luminance contrast as the standard deviation of pixel luminance values within this patch. The standard deviation over the resulting 384 patches (384/16 × 256/16) then provides a global scalar measure of contrast variation within a single image.

Surprise

Our definition of surprise relies on a recently proposed Bayesian definition (Baldi, 2005). It quantifies how observing new data (here, each successive image) affects the internal beliefs that a Bayesian observer may have over a set of hypotheses or models of the world. Data observations that leave the observers' beliefs unaffected carry no surprise and, hence, elicit no response from this model, whereas data observations that cause the observers to significantly revise their beliefs elicit surprise. When applied to predicting the deployment of spatial attention, an early visual form of surprise may be computed over the instantaneous responses of low-level visual feature detectors analyzing an image. This low-level surprise determines that an image patch becomes surprising when its appearance changes abruptly and causes an internal reevaluation of beliefs about the nature of the visual stimulus depicted by the image patch (Baldi, 2005, and below). Regions in video sequences determined to be surprising at such low level by this theory and by the associated computational model have been previously shown to significantly attract human gaze, well above chance (Itti & Baldi, 2006). Here, we processed the image sequences through the model, yielding one topographic "surprise map" for each image in the sequence. This map encodes, for every location in the image, a prediction of how surprising the visual appearance of this location is likely to be to a human observer and, from the aforementioned human gaze tracking results, how likely the location is to capture the attention of the observer.

The implementation of the model used here has been previously described (Baldi, 2005; Itti & Baldi, 2005, 2006) and is publicly available on the World Wide Web (http://ilab.usc.edu/). Briefly, surprise is computed in small image patches over the entire image, along several feature dimensions (color, intensity, orientation, etc.) and at several spatial and temporal scales. At every image patch, a set of beliefs about the visual properties of the world at the corresponding visual location is iteratively established over time. In the current implementation, these beliefs are over low-level hypotheses about the visual world: for example, how much green color or horizontal orientation may be contained in the physical stimulus that gave rise to the observed image patch. The model then uses Bayes' theorem as the basic engine for transitioning from a prior probability distribution $\{P(M)\}_{M\in\mathcal{M}}$ over a set of hypotheses or models M in a model space $\mathcal{M}$ to a posterior distribution $\{P(M|D)\}_{M\in\mathcal{M}}$ after each data observation D:

$$\forall M\in\mathcal{M},\quad P(M|D)=\frac{P(D|M)}{P(D)}P(M). \tag{2}$$

In this framework, the new data observation D (here, an image) carries no surprise if it leaves the observer's beliefs unaffected, that is, if the posterior is identical to the prior; conversely, D is surprising if the posterior distribution resulting from observing D significantly differs from the prior distribution. Therefore, we formally measure surprise by quantifying the distance (or dissimilarity) between the posterior and prior distributions. This is best done using the relative entropy or Kullback–Leibler (KL) divergence (Kullback, 1959). Thus, surprise is defined by the average of the log-odd ratio:

$$\begin{aligned} S(D,\mathcal{M}) &= \mathrm{KL}(P(M|D),P(M)) \\ &= \int_{\mathcal{M}} P(M|D)\log\frac{P(M|D)}{P(M)}\,\mathrm{d}M \end{aligned} \tag{3}$$

taken with respect to the posterior distribution over the model space $\mathcal{M}$ (Baldi, 2005). In the present study, we analyze the average of the surprise values obtained over all the image patches in the visual field, that is, one scalar summary surprise value for each image.

Intuitively, surprise as implemented in our computational model may be viewed as an extension to the concept of saliency but computed simultaneously over space and time at several scales and including an adaptation quality. A spatial oddball (e.g., a red item in an array of green items) or a temporal oddball (e.g., a red item in a sequence of green items) will typically elicit surprise when first presented to the observer. However, because the observer is constantly learning through Bayes' rule and using every data observation to adjust its internal beliefs about the world, surprise elicited by repeated presentations of such oddballs decreases with every presentation (Baldi, 2005). A striking example of this is provided by the observation that white snow, as one would observe on a malfunctioning television set and where every successive frame is entirely unpredictable, is only surprising at its onset but quickly becomes boring as the observer adjusts its internal beliefs toward favoring a random pixel model (Itti & Baldi, 2005). While, in many simple situations, surprise correlates with statistical uniqueness or "unlikelihood" (i.e., how much of an outlier are the observed data given a learned distribution of previously observed data values, or, equivalently, how informative are the data in Shannon's sense given that learned distribution of values), one should be aware that, in a number of cases, the two notions can make opposite predictions: A rare target is not necessarily surprising, as exemplified by white snow where every snow image is unique and extremely rare, and a surprising stimulus is not always an oddball (Itti & Baldi, 2006). This discrepancy stems from the fact that surprise essentially is a measure of the evolving state of the observer in response to successive data observations, rather than solely of the statistics of the observed data. In previous experiments, we found that surprise significantly better predicts the deployment of human gaze over natural video stimuli than either standard saliency or an outlier-based metric (Itti & Baldi, 2006).

Surprise modification in Experiment 2

For Experiment 2, we generated new sequences from the 122 target-present sequences that were correctly reported by all observers (easy sequences) in Experiment 1. For each sequence, we randomly reordered the five distractors prior to and following the target while keeping the target in place. Because the target frame can come as early as Frame 6 and as late as Frame 15, we were limited to a scope of five frames, to ensure similar treatment of all sequences. The same random reordering was tried on each of the easy conditions. That is, we created a template of 100 different random reorderings so that reordering and testing were identical throughout the 122 easy sequences. Of those, we picked the one with highest surprise before ("pre") and after ("post") the target and the one with highest average surprise before and after the target ("both"): Each of the 100 new random sequences for each of the 122 easy sequences was measured for average surprise for each frame. A difficulty score was assigned to each random sequence based on the difference in surprise between the target image frame and the flankers in the condition. Thus, the most difficult "pre" condition would be seen if the image just prior to the target was as surprising as possible compared with the target image itself. It should be noted that this method does not guarantee that sequences picked as "hardest" were the hardest possible; they were the hardest of the 100 random sequences tried. As such, we do not assert that the set is optimal, but it is sufficient. The generated sequences were used together with the original sequences, yielding 488 target-present sequences for Experiment 2. We randomly selected 488 of the 500 target-absent sequences of Experiment 1 to be used in Experiment 2.

Results

Experiment 1—Surprise and contrast variation correlate to performance

Basic behavioral data

In the first experiment, we presented 1,000 sequences, each containing 20 frames of natural scenes, to eight human observers at a presentation rate of 20 Hz (Figure 1). In half of the sequences, exactly one of the pictures (between Frames 6 and 15) depicted one or more nonhuman animal (s), and observers had to report after each sequence whether or not an animal had been presented. To assess interobserver consistency, we used the same 1,000 image sequences in all observers, with the order of sequences randomized for each observer. All observers performed this task above chance level of 50% correct (minimum observer performance: 62.2% correct), but observers were far from perfect (maximum observer performance: 73.3% correct). All but one observer (M.M.) employed a conservative criterion (Figures 2A–2H), that is, exhibited more misses (reporting no target when a target was present) than false alarms (reporting a target when none was present). This trend is reflected in the average performance: In 77.4% of target-absent trials, observers correctly reported the absence of a target (Figure 2I, left), whereas only 59.9% of the targets presented were correctly reported ("hits"). Because we were primarily concerned with sequence statistics relative to target frames, we only analyzed the target-present trials further. Partly owing to our criterion-dependent yes/no design (as compared to forced choice), we observed

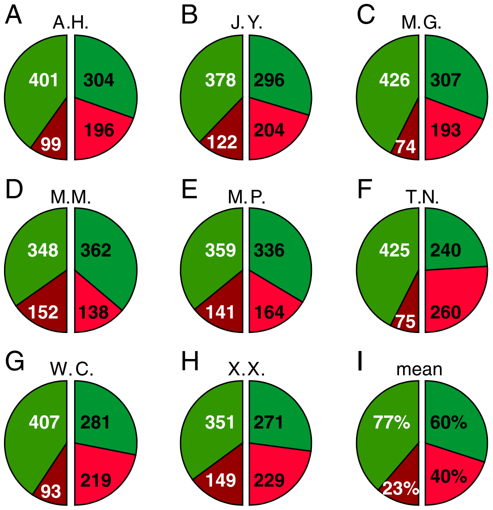


Figure 2. Performance. Pie diagrams for recognition performance of the eight individual subjects (A–H) and mean over all subjects (I). Target-absent trials are shown to the left, whereas target-present trials are shown to the right. Correct responses (correct rejects and hits) are plotted in green, whereas incorrect responses (false alarms and misses) are plotted in red.

substantial interindividual variability in hit rates (correct responses in target-present trials), ranging from 48.0% (T.N., Figure 2F) to 72.4% (M.M., Figure 2D). This variability and the generally high error rates allowed us to have sufficient error trials for statistical analysis.

Subjects' performance is not idiosyncratic

Are error patterns consistent across observers? That is, do different individuals fail or succeed for the same precise image sequences? If so, can we find statistical properties of sequences that predict whether observers fail or succeed to detect a target? If we assume that all subjects commit independent errors, we can represent, a priori, whether or not an observer j correctly detects the target in any given target-present sequence by a probability $p_{\mathrm{hit},j}$, which depends on the individual criterion of the observer. The probability that the target in a given sequence is detected by all observers would then be the product of all $p_{\mathrm{hit},j}$. Given the individually measured values of $p_{\mathrm{hit},j}$ of our eight observers, we obtained a predicted probability of 1.6% for all observers being correct for a sequence. Hence, if observers were independent in their errors, we would expect that in about 8 (1.6% × 500) target-present image sequences, all observers would correctly detect the target (Figure 3, gray). In fact, in 122 of the 500 target-present sequences (24.4%), all eight subjects found the target (Figure 3, black) 15 times higher than expected by chance when subjects make independent errors. Similarly, there were more sequences for which all observers failed to detect the target (29) than the independent observer assumption would predict (0.3). Figure 3 depicts the independence prediction (see the Methods section) for all possible numbers of correct observers together with the data: Toward the extremes (0, 1, and 2 or 7 and 8 correct observers), the data vastly exceed the independence prediction (and consequently undershoots it for intermediate values as the integral must sum to the total number of trials—500—in both cases). This demonstrates that observers were not independent in their error patterns. There are sequences that were easier and there are sequences that were harder than others for all observers. Hence, it is unlikely that high-level processes alone, for example, generic limitations of memory or fluctuating levels of alertness (concentration), are responsible for observer errors, as errors due to these high-level processes would have no reason to be correlated among observers. Instead, our results indicate that there is something common about individual sequences that makes them inherently more easy or more difficult to process.

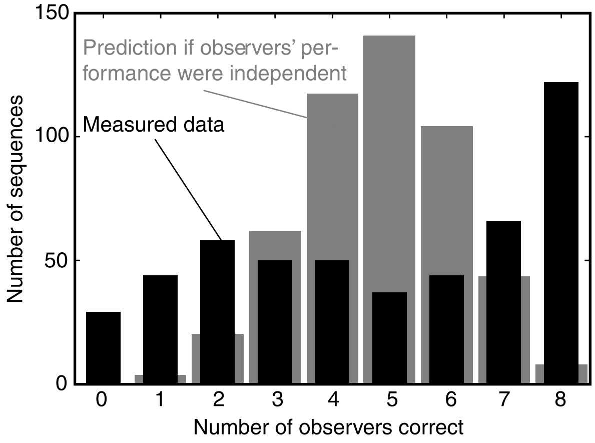


Figure 3. Errors are not independent across observers and sequences. Black bars: Number of target-present sequences correctly reported by any given number of subjects (0–8). Gray bars: prediction, if all subjects were independent from each other. For example, the same 50 image sequences (out of 500) were correctly identified by four observers to contain an animal. If we assume that each observer makes his or her own pattern of errors, we predict 110 such sequences. The data are clearly incompatible with the independence assumption.

Contrast variation in target frame is correlated to performance

We hypothesize that attentional mechanisms contribute to target detection performance in RSVP, that is, render a target easy or hard to detect. Luminance contrast is one of the simplest image properties correlated with fixation probability in natural scenes (Reinagel & Zador, 1999). Luminance contrast is also exploited by saliency-map models to predict gaze allocation (Itti & Koch, 2000), which suggests that a high variation of contrast in the target frame should allow certain features to quickly capture attention, akin to stimulus-driven (bottom–up) "pop-out" of targets defined by elementary features (Treisman & Gelade, 1980). We found the measure of local contrast variation across each target image (see the Methods section) to be significantly increased for the target frame in easy sequences (those in which all observers correctly detected the target) compared to the target frame in all other sequences (p = .0003, t test), as well as compared to the target frame in hard sequences (those in which no subject correctly detected the target, p = .001, Figure 4A). In addition, we observed a small but significant correlation between contrast variation in the target frame and the number of observers correctly reporting the target (r = .14, p = .002). These data demonstrate that easily detectable targets differ—on average—from others in at least one static measure typically associated with spatial attention, their contrast variation.

"Surprising" events before and after the target masks its detection

Although the described correlation of contrast variation to detection may be suggestive of an attentional mechanism modulating detection performance, a property of the target frame itself, in principle, cannot distinguish between attentional and recognition mechanisms. In

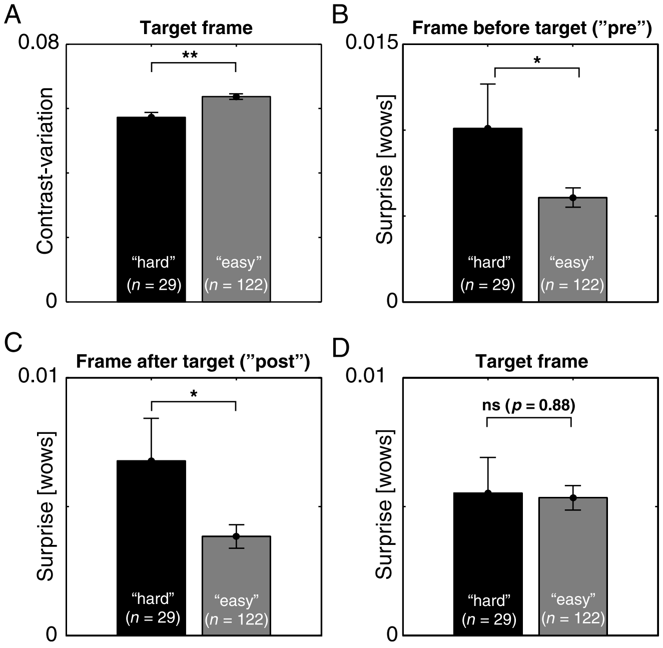


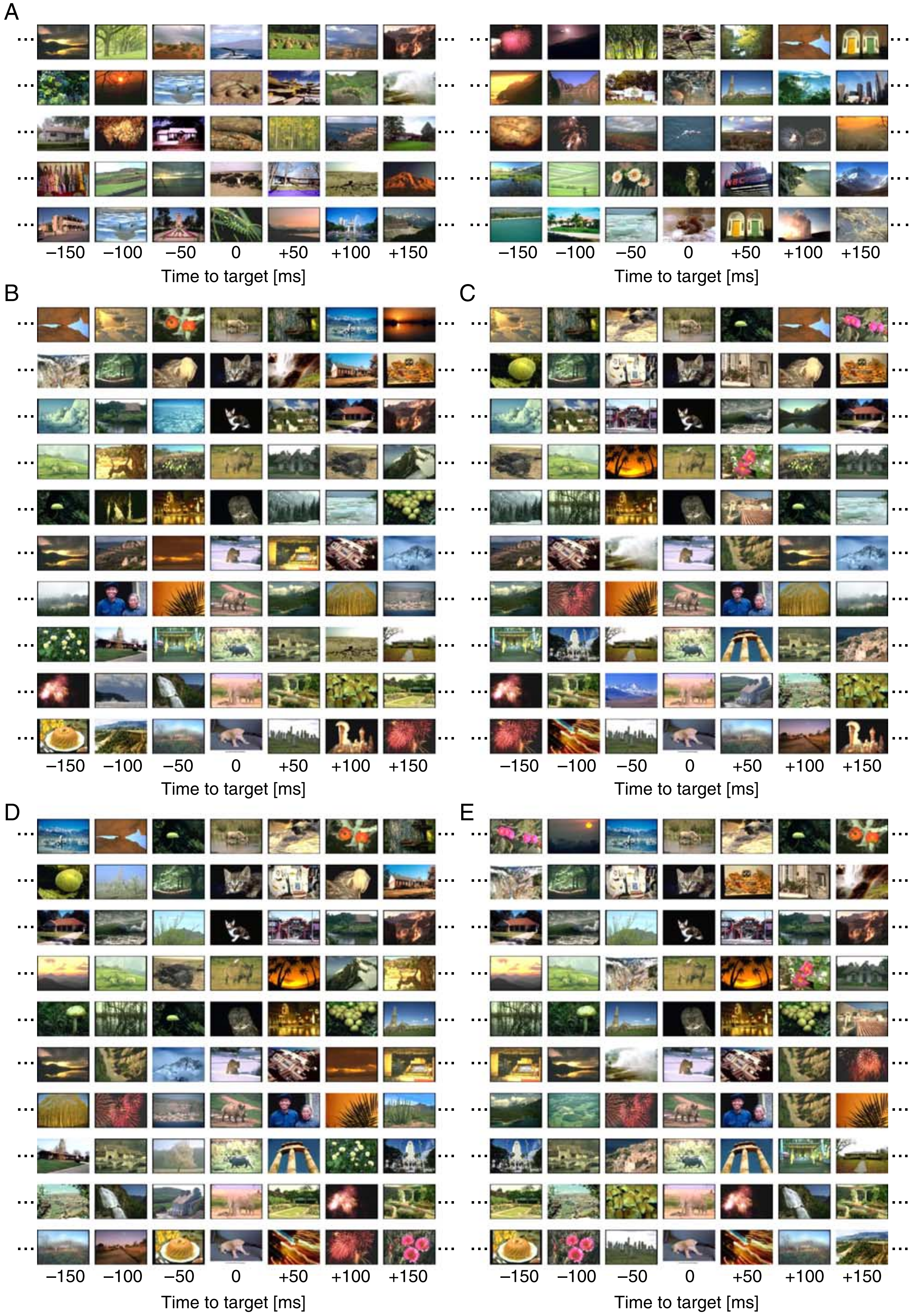
Figure 4. Contrast variation and surprise predict easy from hard sequences. Easy sequences (all eight observers correctly detected the target) differ from hard sequences (none of the observers detected the target) in the contrast variation of the target frame (A) and in the surprise of the frames preceding (B) and succeeding (C) the target frame but not in the surprise of the target frame itself (D). All panels depict *M* ± *SEM* over sequences for the 29 hard (left, black bar) and the 122 easy (right, gray bar) sequences. Significance markers refer to *t* tests (**p* < .05, ***p* < .01).

contrast, if properties of adjacent frames, not correlated to target properties, affect target detection, this would be evidence that attentional mechanisms (and not recognition mechanisms) mediate target detection performance. Hence, we tested under which circumstances temporally adjacent items effectively mask the target, impairing its detection. We hypothesized that if attention were drawn toward the item immediately preceding or succeeding the target, target detection would be impaired. As spatial attention in image sequences is recruited to locations that exhibit high Bayesian surprise (Itti & Baldi, 2005), this hypothesis predicts that high surprise in frames adjacent to the target will impair its detection. Hence, we tested whether low-level visual surprise is increased in frames adjacent to the target. We found that mean surprise was significantly higher for the 29 "hard" sequences (Figure 5A) than for the 122 easy sequences (Figure 5B), in the frames before (p = .02, t test, Figure 4B) and after (p = .02, Figure 4C) the target frame, but not for the target itself (p = .88, Figure 4D). Furthermore, surprise was not correlated to contrast variation in the target frame, neither for the preceding (r = .01, p = .76), the succeeding (r = .03, p = .56), nor the target frames (r = .0006, p = .99). Hence, the effects of surprise did not result from any correlation to contrast variation in the target frame. This implied that increased surprise, before or after the target, itself is involved in impairing target detection. In summary, Experiment 1 demonstrated that error patterns in RSVP are consistent across individuals and can be predicted by at least two independent statistical measures of attention: contrast variation of the target and surprise of adjacent items. Target images with higher contrast variation were more often detected, while frames exhibiting higher surprise adjacent to a target frame effectively masked the target and impaired detection. Hence, target detection performance is significantly correlated with simple measures of statistical image properties thought to mediate spatial attention; in Experiment 2, we test whether the correlation observed for the surprise measure reflects a causal effect of low-level visual surprise on target detection.

Experiment 2—Increasing surprise renders easy sequences hard

In a second experiment, we tested whether the observed effect of surprise on target detection is causal and independent of static image properties. To this end, we used predictions from the surprise measure to reorder the distractor frames in the 122 sequences in which all subjects had successfully detected the target. For each of these easy sequences, we generated four different conditions (see the Methods section) by reshuffling distractor images while maintaining the target image: the order in Experiment 1 ("original," Figure 5B), an order that exhibited increased surprise before the target ("pre", Figure 5C), after the target ("post", Figure 5D), or both before and after the target ("both", Figure 5E). In all cases, the target frame remained at the same temporal location and the same 19 distractors were used. Eight volunteers participated in Experiment 2, none of whom had participated in Experiment 1 or in any other experiment using the same set of stimuli. In all of these individuals (Figures 6A–6H), the hit rate for the original sequences was higher than for any of the surprise-modified conditions, indicating that increased surprise indeed causally impaired target detection. The "pre" and "post" conditions fell between original and "both" in all observers with no consistent difference between the "pre"

Figure 5. Sample sequences. (A) Ten examples of hard sequences, that is, those for which no observer reported the target. (B) Ten examples out of 122 easy sequences. (C–E) Reordered versions of sequences of Panel B: (C) "pre" condition, (D) "post" condition, (E) "both" condition. For all sequences, only the three frames adjacent to the target frame are depicted. Resolution of images was reduced for the figure. All sequences and corresponding results for all observers are available at http://n.ethz.ch/~einhaeuw/download/ (jovSuppl.tar.gz).


A
−150
−100
−50
0
+50
+100
+150
Time to target [ms]
−150
−100
−50
0
+50
+100
+150
Time to target [ms]
B
−150
−100
−50
0
+50
+100
+150
Time to target [ms]
C
−150
−100
−50
0
+50
+100
+150
Time to target [ms]
D
−150
−100
−50
0
+50
+100
+150
Time to target [ms]
E
−150
−100
−50
0
+50
+100
+150
Time to target [ms]

and “post” conditions: Five observers had higher hit rates in “pre”, and two had higher hit rates in “post”. Performance in the “both” condition was worse than in any other condition for each individual, indicating that increasing surprise before and after the target causes a more effective impairment than either alone.

While the consistency of effects in eight out of eight subjects—performance is best in the original sequences, worse for “pre” and “post”, and worst for “both”—had already been a strong indication that increased surprise adjacent to a target frame causally impairs target detection, we also analyzed the mean hit rates across subjects. As expected, analysis of variance revealed that there was a significant main effect of condition on hit rates, $F(3, 28) = 12.7$, $p = 2 \times 10^{-5}$ (Figure 6I). Post hoc *t* tests revealed highly significant differences between the original and all the other conditions ($p = .008$ to “pre”, $p = 2 \times 10^{-4}$ to “post”, and $p = 7 \times 10^{-5}$ to “both”) as well as a significant difference between “pre” and “both” ($p = .01$). No significant effect was found between “post” and “both” ($p = .09$) and between “pre” and “post” ($p = .18$). Because we were interested in the relative effect of manipulating surprise rather than the absolute performance, we normalized each observer’s performance to his or her performance for the original sequences (Figure 6K). Mean normalized performance for the “pre”, “post”, and “both” conditions was significantly different from the nominal 100% of the original (one-sample *t* tests: $p = 7 \times 10^{-4}$, $p = 8 \times 10^{-6}$, and $p = 3 \times 10^{-5}$, respectively). Significant differences were also found for all pairwise *t*-test comparisons between conditions: between “pre” and “both” ($p = 7 \times 10^{-4}$), between “post” and “both” ($p = .01$), and between “pre” and “post” ($p = .04$). In summary, for all individuals, increasing surprise before and/or after the target frame impaired performance. This impairment was stronger if surprise was increased both before and after the target than for either modification alone. These data demonstrate that surprising events significantly impaired the detection of target items that had been readily detected when identical distractors were presented in a different order. Hence, low-level visual surprise—or some dynamic measure closely related to it—causally modulates the human ability to detect temporally adjacent items. Because only the spatiotemporal context, and not the targets themselves, was different between conditions, we conclude that an attentional mechanism (surprise) causally modulates target detection. Consequently, our findings demonstrate that target detection in RSVP streams is not dependent on target recognition alone but, to a large extent, on attentional mechanisms.

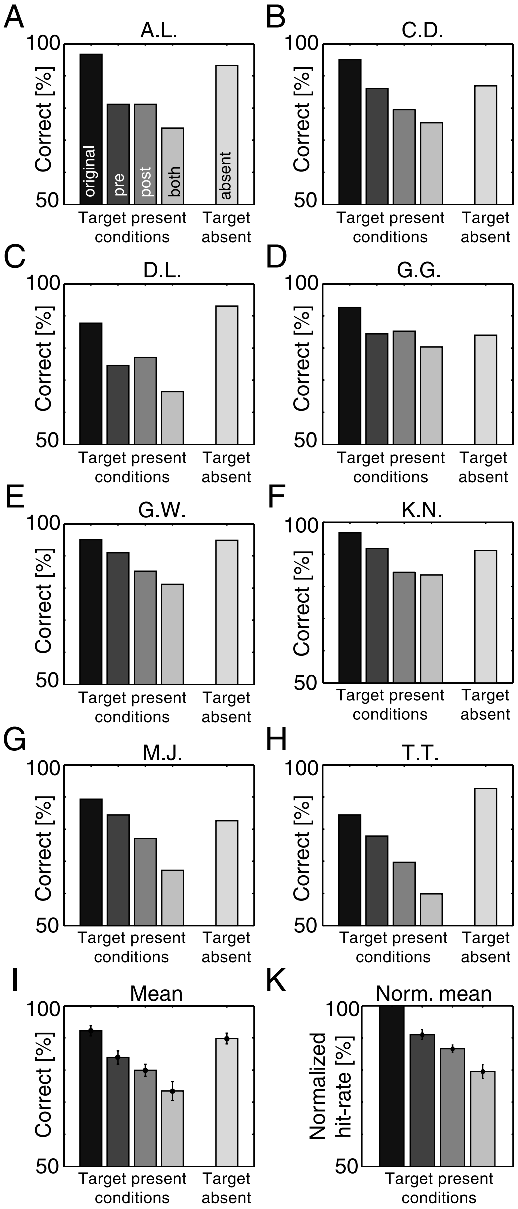


Figure 6. Surprise makes easy sequences hard. Each panel from left to right: Detection performance (hits) on original sequences (easy sequences of Experiment 1), on the same sequences reordered to have increased surprise in the frame preceding the target (“pre”), succeeding the target (“post”), or both; correct rejects in target-absent trials. Individual subjects (A–H), *M* ± *SEM* across all subjects (I), and normalized performance (K). The *y*-axis has been truncated at 50% for all plots.

Discussion

In this study, we show that performance in an RSVP task is predicted by statistical properties of the stimulus sequence. In particular, contrast variation in the target frame and the information-theoretical measure of surprise in adjacent frames are correlated to performance. Increased surprise of flanking images impairs performance for targets that were readily detected when a different temporal order of distractors elicited less surprise in frames adjacent to the target. Hence, attentional mechanisms, to a large extent, determine human performance in a rapid detection task.

We deliberately chose a database that has been widely used in rapid scene recognition experiments (Evans & Treisman, 2005; Li et al., 2002). Because these studies did not find anything specific to detecting the category of "animals" (e.g., compared to vehicles), we join a large body of previous studies in employing an animal/no-animal detection task (Evans & Treisman, 2005; Fabre-Thorpe et al., 1998; Li et al., 2002; Rousselet et al., 2002; Thorpe et al., 1996; VanRullen & Thorpe, 2001). Because our study, furthermore, does not test for differences between targets and nontargets per se but focuses on performance differences between different targets or between different adjacent distractors, it seems very unlikely that the choice of this common task and data set affects our results. For a sound analysis of error patterns, we needed observers to perform below ceiling. Because previous RSVP studies found high performance for animal detection at rates up to 13.3 Hz (Evans & Treisman, 2005), we chose an even higher rate (20 Hz). As expected, performance was far below ceiling (100%), but all subjects performed the overall task above chance (50%). Because most observers employed a conservative criterion for trading off false alarms versus misses, we nevertheless obtained a sufficient number of error trials (on average, 40% of all target-present trials). Given this, the interobserver consistency is even more remarkable than if there were only very few hard trials interspersed among otherwise very easy trials. We chose a yes/no protocol rather than a forced-choice design. Although forced-choice designs are preferable in many situations, in which results should not depend on individual criteria, they would be suboptimal in the present context: Assume we would show two sequences, in one of which there is an animal and there is none in the other, and subjects had to make a forced decision, the question then is which is which? In this scenario, even if the target-present sequence were so hard that no observer could detect the target, we would expect half of the observers to pick this sequence. Consequently, this possible forced-choice design (and similarly other forced-choice detection designs) would complicate comparing the same sequence across different observers, which is key to this study. Hence, for the purpose of this study, a yes/no design was preferable. As a final note on the design, it should be emphasized that our study deals with target detection, rather than identification. Because a previous study on the same data set (Evans & Treisman, 2005) showed that attentional demands between these two tasks differ, it might be interesting for further research to also investigate (subcategorical) target identification.

The measures that we found to predict RSVP performance are widely used for modeling spatial attention (Baldi, 2005; Itti & Baldi, 2005, 2006; Itti & Koch, 2000). Under most experimental conditions, luminance contrast is correlated to fixation probability in natural scenes (Einhäuser & König, 2003; Mannan, Ruddock, & Wooding, 1996; Reinagel & Zador, 1999; Tatler, Baddeley, & Gilchrist, 2005), and fixations constitute a close correlate of spatial attention in natural viewing (Rizzolatti, Riggio, Dascola, & Umilta, 1987). In turn, spatial attention modulates the gain of visual neurons as if the attended stimulus' luminance contrast had increased (Reynolds & Desimone, 2003). Our measure of contrast variation describes a second-order contrast property, which has also been suggested to effectively drive spatial attention (Einhäuser & König, 2003; Parkhurst & Niebur, 2004). While we do not claim that contrast variation is the *only* static feature that distinguishes easy from hard targets, its correlation to detection performance remains striking in the light of its role in bottom–up-driven spatial attention.

Like contrast and contrast variation, the measure of surprise was originally introduced to model the sensory-driven (bottom–up) guidance of spatial attention, in particular for dynamic stimuli (Baldi, 2005; Itti & Baldi, 2005, 2006). While the exact mechanisms by which surprise impairs target detection remain to be revealed, it seems likely that surprising events "parasitically" capture attention and thereby block resources that would be needed for the target either to be detected or to be consolidated into visual short-term memory. Such an occupation of resources is typically assumed in models of attentional blink, in which processing of a second target occurring shortly after a first one is impaired (Raymond et al., 1992). In contrast to attentional blink, surprising events also impair detection of the following item rather than sparing this direct successor. In this respect, our observations are more reminiscent of another attentional impairment, repetition blindness (Kanwisher, 1987), which is related to attentional blink, but distinct from it (Chun, 1997). Nevertheless, some previously described higher level attentional-blink-like phenomena do not exhibit such "lag-1 sparing," most notably cross-modal attentional impairments (Jolicoeur, 1999). Rather than dealing with the specific situations of attentional blink or repetition blindness, which require the repetition of a particular exemplar or close succession of items sticking out by experimental design (being defined as target), we here model attentional impairment that is caused by specific statistical properties of generic distractor items. Our results are consistent with an attentional gating model (Reeves & Sperling, 1986), variants of which underlie most accounts of attentional impairments. In this view, a surprising event preceding the target opens the attentional gate and enters the same processing epoch as the target, a surprising event succeeding the target slips in the epoch opened by the target. In both cases, the "surprising" distractor competes with the target for access to visual short-term memory. A distractor of sufficiently strong surprise overcomes the target's top–down salience (as being a target) and impairs its report. Even if alternative explanations to this gating model may exist, the fact that a model of spatial attention predicts RSVP performance clearly demonstrates an involvement of attentional limitations. As surprise impairs the detection of otherwise readily detectable targets, the limitation of attention resources, rather than target recognition mechanisms alone, is likely to be the primary source for detection errors in RSVP.

By design, our surprise model in its current implementation only accounts for stimulus-driven (bottom–up) factors on attention while neglecting task-related (top–down) components. Including task-related factors in the model will be an interesting issue for future research: One may, for example, include the instruction to ignore a specific item by biasing the prior distribution accordingly. Furthermore, surprise captures a subject's low-level response to observed stimulus statistics but does not, at this stage, *explicitly* model the semantic content of an image. Most et al. (2005) show that emotional images (e.g., a murder scene) impair recognition of subsequent items akin to the attentional blink between two target items. Only some observers can overcome this impairment through attentional strategies, whereas those with high "harm avoidance" cannot do so. This indicates that such "semantic surprise" is not only task dependent but also personality dependent. Marois et al. (2003) demonstrated that odd items in an RSVP sequence impair subsequent recognition and dubbed this phenomenon "surprise blindness." In their experiments, the oddity (face vs. letter) is at least also semantic. This semantic difference is likely to be correlated to statistical differences between odd item and sequence items. Consequently, our model would most likely also correctly predict the observed performance impairment in response to these statistical differences. Intriguingly, even without knowledge of semantics or task demands, our model predicts attentional impairments of recognition. Extensions of our model that include instructions and scene semantics are, however, conceivable and likely to further improve our model's predictions.

Our results show that attentional resources are needed for rapid detection of animals in sequences of natural scenes. This is in conflict with earlier studies, which demonstrated that rapid recognition could be performed in the "near" absence of attention (Li et al., 2002; Rousselet et al., 2002). The most important difference between these studies and ours is with regard to how attentional load is generated. Both aforementioned studies use dual-task paradigms: Observers perform concurrent recognition at different spatial locations. In contrast, we here use sequences of stimuli at the same location. Recent preliminary data on famous face recognition indicate a difference between attentional impairments in RSVP sequences as compared to dual-task paradigms (Reddy, VanRullen, Koch, & Perona, unpublished observations). One explanation states that, in dual-task paradigms, some residual attention remains at the "unattended" location, as spatial attention, to some extent, can be divided between multiple locations (Kramer & Hahn, 1995; McMains & Somers, 2004). However, Li et al. (2002) have carefully controlled for this possibility, which renders such an explanation unlikely. It is nevertheless possible that the mechanisms underlying spatial attention and attentional limitations in rapidly presented sequences do not overlap in full. In particular, in dual-task paradigms, attention is pinned top–down onto specific locations in the visual field, which are determined by the task design. In contrast, in an RSVP paradigm, covert spatial attention is free to move within the extent of the image stream. Our data suggest that, indeed, bottom–up saliency/surprise, typically associated with rapid shifts of spatial attention, significantly impacts RSVP recognition performance. Integrating task-related factors into our surprise model will be an interesting aspect of future research and might also help to unveil the differences of dual-task and RSVP paradigms.

As humans by far outperform contemporary artificial systems in scene recognition, several computational approaches try to exploit human-like strategies. Recent successful approaches include global features, that is, features that do not require prior image segmentation (Oliva & Torralba, 2006), as well as texture-based models (Renninger & Malik, 2004). Unlike in these previous computational studies, the main aim of our investigation was *not* to optimize performance of such a recognition system. Rather, we model human error patterns in an experimental setting in which a high presentation rate ensures comparably low performance. Besides the insight in physiological mechanisms, our findings may aid machine-assisted human operation, in applications of high visual throughput, such as surveillance, luggage screening, or imagery analysis[1] (Clapper, 2004). Even if such applications usually operate at lower presentation rates, comparable to the 3–5 Hz of saccadic eye movements, and therefore at considerably lower error rates, understanding human error patterns is crucial, as the consequences of even a single missed target might come at an extremely high cost.

Acknowledgments

This work was supported by the Swiss National Science Foundation (W.E., PA00A-111447), DARPA, NGA, NSF, ONR, the NIMH, and HFSP.

Commercial relationships: none.

Corresponding author: Wolfgang Einhäuser.
Email: wolfgang.einhaeuser@inf.ethz.ch.
Address: ETH Zürich, Institute of Computational Science, CAB G 82.2, Universitätstrasse 6, 8092 Zürich, Switzerland.

Footnote

[1] The need for such applications is obvious from several governmentally funded programs in the United States; see, for example, http://www.darpa.mil/dso/thrusts/trainhu/nia/index.htm, which also, in part, funds some of the authors.

References

Baldi, P. (2005). Surprise: A shortcut for attention? In L. Itti, G. Rees, & J. Tsotsos (Eds.), *Neurobiology of attention* (pp. 24–28). San Diego, CA: Elsevier.

Barnard, P. J., Scott, S., Taylor, J., May, J., & Knightley, W. (2004). Paying attention to meaning. *Psychological Science, 15,* 179–186. [PubMed]

Biederman, I. (1981). On the semantics of a glance at a scene. In M. Kuboby & J. R. Pomerantz (Eds.), *Perceptual organisation* (pp. 213–253). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Chun, M. M. (1997). Types and tokens in visual processing: A double dissociation between the attentional blink and repetition blindness. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 738–755. [PubMed]

Clapper, J. R. (2004). *Geospatial-Intelligence (GEOINT) basic doctrine*. Washington, DC: National Geospatial Intelligence Agency.

Einhäuser, W., & König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience, 17,* 1089–1097. [PubMed]

Evans, K. K., & Treisman, A. (2005). Perception of objects in natural scenes: Is it really attention free? *Journal of Experimental Psychology: Human Perception and Performance, 31,* 1476–1492. [PubMed]

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

Folk, C. L., Leber, A. B., & Egeth, H. E. (2002). Made you blink! Contingent attentional capture produces a spatial blink. *Perception & Psychophysics, 64,* 741–753. [PubMed] [Article]

Itti, L., & Baldi, P. (2005). A principled approach to detecting surprising events in video. *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 1,* 631–637.

Itti, L., & Baldi, P. (2006). Bayesian surprise attracts human attention. *Advances in Neural Information Processing Systems (NIPS 2005), 19,* 1–8.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40,* 1489–1506. [PubMed]

Jolicoeur, P. (1999). Restricted attentional capacity between sensory modalities. *Psychonomic Bulletin & Review, 6,* 87–92. [PubMed]

Kanwisher, N. G. (1987). Repetition blindness: Type recognition without token individuation. *Cognition, 27,* 117–143. [PubMed]

Kramer, A. F., & Hahn, S. (1995). Distribution of attention over noncontiguous regions of the visual field. *Psychological Science, 6,* 381–386.

Kullback, S. (1959). *Information theory and statistics*. New York: Wiley.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences of the United States of America, 99,* 9596–9601. [PubMed] [Article]

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision, 10,* 165–188. [PubMed]

Marois, R., Todd, J. J., & Gilbert, C. M. (2003). Surprise blindness: A distinct form of attentional limit to explicit perception [Abstract]? *Journal of Vision, 3*(9):738, 738a, http://journalofvision.org/3/9/738/, doi:10.1167/3.9.738.

McMains, S. A., & Somers, D. C. (2004). Multiple spotlights of attentional selection in human visual cortex. *Neuron, 42,* 677–686. [PubMed] [Article]

Most, S. B., Chun, M. M., Widders, D. M., & Zald, D. H. (2005). Attentional rubbernecking: Cognitive control and personality in emotion-induced blindness. *Psychonomic Bulletin & Review, 12,* 654–661. [PubMed]

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research, 155,* 23–36. [PubMed]

Parkhurst, D. J., & Niebur, E. (2004). Texture contrast attracts overt visual attention in natural scenes. *European Journal of Neuroscience, 19,* 783–789. [PubMed]

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

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

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. [PubMed]

Reeves, A., & Sperling, G. (1986). Attention gating in short-term visual memory. *Psychological Review, 93,* 180–206. [PubMed]

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network, 10,* 341–350. [PubMed]

Renninger, L. W., & Malik, J. (2004). When is scene identification just texture recognition? *Vision Research, 44,* 2301–2311. [PubMed]

Reynolds, J. H., & Desimone, R. (2003). Interacting roles of attention and visual salience in V4. *Neuron, 37,* 853–863. [PubMed] [Article]

Rizzolatti, G., Riggio, L., Dascola, I., & Umiltá, C. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia, 25,* 31–40. [PubMed]

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience, 5,* 629–630. [PubMed] [Article]

Serre, T., Oliva, A., & Poggio, T. (2006). Feedforward theories of visual cortex predict human performance in rapid categorization [Abstract]. *Journal of Vision, 6*(6):615, 615a, http://journalofvision.org/6/6/615/, doi:10.1167/6.6.615.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45,* 643–659. [PubMed]

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

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12,* 97–136. [PubMed]

VanRullen, R., & Thorpe, S. J. (2001). The time course of visual processing: From early perception to decision-making. *Journal of Cognitive Neuroscience, 13,* 454–461. [PubMed]