Vision Research 49 (2009) 1620–1637

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Automatic computation of an image's statistical surprise predicts performance of human observers on a natural image detection task

T. Nathan Mundhenk [a,*], Wolfgang Einhäuser [b], Laurent Itti [a]

[a] Department of Computer Science, University of Southern California, Hedco Neuroscience Building, HNB 10 Los Angeles, CA 90089-2520, USA
[b] Department of Neurophysics, Phiilipps-University Marburg Renthof 7, 35032 Marburg, Germany

ARTICLE INFO

Article history:
Received 24 July 2008
Received in revised form 28 March 2009

Keywords:
Surprise
Attention
RSVP
Detection
Modeling
Saliency
Natural image
Computer vision
Masking
Statistics

ABSTRACT

To understand the neural mechanisms underlying humans' exquisite ability at processing briefly flashed visual scenes, we present a computer model that predicts human performance in a Rapid Serial Visual Presentation (RSVP) task. The model processes streams of natural scene images presented at a rate of 20 Hz to human observers, and attempts to predict when subjects will correctly detect if one of the presented images contains an animal (target). We find that metrics of Bayesian surprise, which models both spatial and temporal aspects of human attention, differ significantly between RSVP sequences on which subjects will detect the target (easy) and those on which subjects miss the target (hard). Extending beyond previous studies, we here assess the contribution of individual image features including color opponencies and Gabor edges. We also investigate the effects of the spatial location of surprise in the visual field, rather than only using a single aggregate measure. A physiologically plausible feed-forward system, which optimally combines spatial and temporal surprise metrics for all features, predicts performance in 79.5% of human trials correctly. This is significantly better than a baseline maximum likelihood Bayesian model (71.7%). We can see that attention as measured by surprise, accounts for a large proportion of observer performance in RSVP. The time course of surprise in different feature types (channels) provides additional quantitative insight in rapid bottom-up processes of human visual attention and recognition, and illuminates the phenomenon of attentional blink and lag-1 sparing. Surprise also reveals classical Type-B like masking effects intrinsic in natural image RSVP sequences. We summarize these with the discussion of a multistage model of visual attention.

1. Introduction

What are the mechanisms underlying human target detection in RSVP (Rapid Serial Visual Presentation) streams of images, and can they be modeled in such a way as to allow prediction of subject performance? This question is of particular interest since, when images are presented at high speed, humans can detect some but not all images of a particular type (target images; e.g. images containing an animal) which they would be able to detect with far greater accuracy at a slower rate of presentation. Our answer to this question is that two primary forces are at work related to attention and are part of a two or more stage model (Chun & Potter, 1995; Reeves, 1982; Sperling, Reeves, Blaser, Lu, & Weichselgartner, 2001). Here we will suggest that the *first stage* is purely an attentional mask with the blocking strength of attention given by image features which have already been observed. The *second stage* on the other hand can block the perception of the image if another image is already being processed and is monopolizing its limited resources.

We consider here the metric of Bayesian surprise (Itti & Baldi, 2005, 2006) to predict how easily a target image containing an animal may be found among 19 frames of other natural images (distractors) presented at 20 Hz. In a *first* experiment, we show that surprise measures are significantly different for target images which subjects find easy to detect in the RSVP sequences vs. those which are hard. We then present a *second* experiment which attempts to predict subject performance by utilizing the surprising features to determine the strength of attentional capture and masking. This is done using a back-propagation neural network whose inputs are the features of surprise and whose output is a prediction about the difficulty of a given RSVP sequence.

1.1. Overview of attention and target detection

It has long been argued that attention plays a crucial role in short term visual detection and recall (Duncan, 1984; Hoffman, Nelson, & Houck, 1983; Mack & Rock, 1998; Neisser & Becklen, 1975; Sperling et al., 2001; Tanaka & Sagi, 2000; VanRullen & Koch, 2003a; Wolfe, Horowitz, & Michod, 2007). This also applies to detection of targets when images are displayed, one after another,

* Corresponding author.
E-mail address: Nathan@mundhenk.com (T.N. Mundhenk).

0042-6989/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2009.03.025

in a serial fashion (Duncan, Ward, & Shapiro, 1994; Raymond, Shapiro, & Arnell, 1992). Many studies have demonstrated that a distracting target image presented before another target image blocks its detection in a phenomenon known as the attentional blink (Einhäuser, Koch, & Makeig, 2007a; Einhäuser, Mundhenk, Baldi, Koch, & Itti, 2007b; Evans & Treisman, 2005; Maki & Mebane, 2006; Marois, Yi, & Chun, 2004; Raymond et al., 1992; Sergent, Baillet, & Dehaene, 2005). Thus, one image presented to the visual stream can interfere with another image that quickly follows, essentially acting as a forward mask (e.g. target in frame A blocks target in frame B).

Additionally, with attentional blink, interference follows a time course, whereby optimal degradation of detection and recall performance for a second target image can occur when it follows the first target image from 200–400 ms (Einhäuser et al., 2007a), which is evidence of a second stage processing bottleneck. In most settings, an intermediate distractor between the first and second targets is needed to induce an attentional blink, a phenomenon known as lag-1 sparing (Raymond et al., 1992). However, for some types of stimuli, such as a strong contrast mask with varying frequencies and naturalistic colors superimposed with white noise, interference can occur very quickly (VanRullen & Koch, 2003b). This may create a situation whereby for some types of stimuli, a target is blocked by a prior target with a very recent onset (<50 ms prior), or by contrast, a much earlier onset (>150 ms prior). As such, there seems to be a short critical period with a U-shaped performance curve where interference is reduced against the second target. That is, interference is reduced if the preceding distractor comes in a critical period of approximately a 50–150 ms window before the target, but is larger otherwise. This interval we will generically refer to as the sparing interval.

In addition to interference with the second target, detection of the first target itself can be blocked by backward masking (e.g. target in frame B blocks target in frame A) (Breitmeyer, 1984; Breitmeyer & Öğmen, 2006; Hogben & Di Lollo, 1972; Raab, 1963; Reeves, 1980; VanRullen & Koch, 2003b; Weisstein & Haber, 1965). However, in natural scene RSVP, backward masking occurs at a very short time interval, <50 ms without a good ability to dwell in time (Potter, Staub & O'Conner, 2002) That is, interference is not U-shaped in the same way as with forward masking. As we will mention later, longer intervals (>150 ms) may conversely enhance detection of the first target (e.g. target in frame B enhances target in frame A). However, backwards masking in the case of RSVP still retains a U *like* shape as the effects of the mask peak and decrease. The difference is that it does not have a second almost discrete episode of new masking following a short interval of sparing as forward masking does. That is, once the backwards mask fades in effect the first time, it is finished masking. The forward mask on the other hand has the ability mask twice.

Putting these pieces together, there is a lack of literature showing a strong reverse attentional blink which would be produced by a second interval of backwards masking. With different stimuli, both forward and backward masking can be observed over very short time periods by flanking images in a sequence. However, only forward masking interferes with targets observed several frames apart, following a sparing lag with a much higher target onset latency. It should be noted that these masking effects are not universally observed in all experiments. As such, the mechanisms responsible for masking are dependant to some degree on both masking and target stimuli, which may result in a target being spared or completely blocked.

Much like masking from temporal offsets, if the first target is displayed *spatially offset* from the second target, interference is decreased for recall of the first target (Shih, 2000). As an example, a large spatial offset would occur if a target in frame A is in the upper right hand corner while a target in frame B is in the lower left hand corner. Thus, if we overlapped the frames, the targets themselves would not overlap. As a result, recognition of individual target images allows for some parallel spatial attention (Li, VanRullen, Koch, & Perona, 2002; McMains & Somers, 2004; Rousselet, Fabre-Thorpe, & Thorpe, 2002). This is also seen if priming signals for target and distractor are spatially offset (Mounts & Gavett, 2004). However, at intervals over 100 ms, spatial overlap may actually prime targets in an attentional blink task (Visser, Zuvic, Bischof, & Di Lollo, 1999). We then should gather that objects offset in space lack some power to mask each other, but in contrast, may at longer time intervals lack the ability to prime each other. Additionally it has been found that more than one object at different temporal offsets can be present in memory pre-attentively, at the same time, but only if they do not overlap at critical *temporal, spatial* and even *feature* offsets (VanRullen, Reddy, & Koch, 2004). However, there is reduced performance as more items, such as natural images, are added in parallel (Rousselet, Thorpe, & Fabre-Thorpe, 2004). Thus, even if objects do not interfere along critical dimensions, performance may degrade as a function of the number of complex distractors added.

1.2. Surprise and attention capture

Prediction of which flanking images or objects will interfere with detection of a target might be accounted for in a Bayesian metric of statistical surprise. Such a metric can be derived for attention based on measuring how a new data sample may affect prior beliefs of an observer. Here, surprise (Itti & Baldi, 2005, 2006), is based on the *conjugate prior* information about observations combined with a *belief* about the reliability of each observation (For mathematical details, see Appendix A). Surprise is strong when a new observation causes a Bayesian learner to substantially adjust its beliefs about the world. This is encountered when the distribution of posterior beliefs highly differs from the prior. The present paper extends our previous work (Einhäuser et al., 2007b), by optimally combining surprise measures from different low-level features. The contribution of different low-level features to "surprise masking", and thus their role in attention, can be individually assessed. Additionally, we will demonstrate how we have extended on this work by creating a non-relative metric that can compare difficulty for RSVP sequences with disjoint sets of target and distractor images. That is, our original work was only able to tell us if a new ordered sequence was relatively more difficult than its original ordering. The current work will focus on giving us a parametric and absolute measure based on how many observers should be able to spot a target image in a given sequence set.

While surprise has been shown to affect RSVP performance, it remains to be seen how surprise from different types of image *features* interacts with recall. Importantly, critical peaks of surprise, along specific feature dimensions, can be measured and used to assess the degree to which flanking images may block one another. For instance, should an image with surprising horizontal lines have more power in masking a target than an image with surprising vertical lines? This is important since some features may be more or less informative, for instance if they have a low signal to noise ratio (SNR) between the target and distractors (Navalpakkam & Itti, 2006). Additionally, some features may be primed in human observers, making them more powerful. As an example, if features can align and enhance along temporal dimensions (Lee & Blake, 2001, Mundhenk, Landauer, Bellman, Arbib, & Itti, 2004; Mundhenk, Everist, Landauer, Itti, & Bellman, 2005) in much the same way they do spatially (Li, 1998; Li & Gilbert, 2002; Mundhenk & Itti, 2005; Yen & Fenkel, 1998), then some features that appear dominant may have a fortunate higher incidence of temporal and/or spatial colinearity in image sequences.

In order to predict and eventually augment detection of target images, a metric is needed that measures the degree of interference