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

of images in a sequence. Here we follow the hypothesis that temporally flanking stimuli interfere with target detection, if they "parasitically" capture attention so that the target may fall within an attentional blink period (e.g. target in frame A and C may interfere with target in frame B). The challenge is then to devise measures of such attentional masking for natural images using stimulus statistics. We herein derive and combine several such measures, based on a previously suggested model of spatial attention that uses Bayesian surprise. Note that this "surprise masking" hypothesis does *not* deny other sources of recognition impairment, e.g. some targets are simply difficult to identify even in spatio-temporal isolation. Instead "surprise masking" addresses recall impairment arising from the sequential ordering of stimuli, rather than the inherent difficulty of identifying the target itself.

If surprise gives a good metric of interference and masking between temporally offset targets in RSVP, we should expect that by measuring surprise between images by its spatial, temporal and feature offsets we will obtain good prediction of how well subjects will detect and recall target images. To this end, we first present evidence that surprise gives a good indication of the difficulty for detecting targets in RSVP sequences and that it elegantly reveals almost classic intrinsic masking effects in natural images. We then show for the first time a neural network model that predicts subject performance using surprise information. Additionally, we discuss the surprise masking observed between images in an RSVP sequence within the framework of a two-stage model of visual processing (Chun & Potter, 1995).

2. Methods

2.1. Surprise in brief

Although the mathematical details of the surprise model are given in Appendix A, we provide here an intuitive example for illustration. The complete surprise *system* subdivides into individual surprise *models*. Each of these individual models is defined by its feature (color opponencies, Gabor wavelet derived orientations, intensity), location in the image and scale. The respective feature values will be maintained over time, in the current state, at each location in the image, by feeding the observed value to a surprise model at that location. That is, every image location, feature and scale has its own surprise model. New data coming into surprise models can then by compared with past observations in a reentrant manner (Di Lollo, Enns, & Rensink, 2000) whereby observed normalcy by the models can guide attention when there is a statistical violation of it. Different frequencies of surprising events are accounted for with several time scales by feeding the results of successive surprise models forward.

Imagine that we present the system initially with the image of a yellow cat. If the system is continuously shown pictures of yellow cats, many surprise models will begin to develop a belief that what they should expect to see are features related to yellow cats. However, not all models will develop the same expectation. Models in the periphery will observe the corners or edges of the image, which in a general sample would be more likely to contain background material such as grass or sky. If after several pictures of yellow cats a blue chair is shown, surprise will be largest with models which had been locally exposed more often to the constant stimuli of the yellow cats. Additionally, surprise should be enhanced more for models which measure color, than for models which measure intensity or line orientations since, in this example, color was more constant prior to the observation of the blue chair. Finally, surprise can be combined across different types of features (channels), scales and locations into a single surprise metric. This combination can either be hardwired (Einhäuser et al., 2007b) or optimized for the prediction task (the present paper).

As is typical for Bayesian models, there is an initial assumption about the underlying statistical process of the observations. Since we wish to model a process of neural events, which should bring us closer to a neurobiological *like* event of surprise, a Poisson/Gamma process is used as the basis. The Gamma process is used since it gives the probability distribution of observing Poisson distributed events with refractoriness such as neural spike trains (Ricciardi, 1995). Surprise is then extended into a visually relevant tool by combining its functioning with an existing model of visual saliency [see also another saliency model: (Li, 2002)]. As a less abstract example, the Gamma process can be used to model the expected waiting time to observe event B after event A has been observed. If events A and B are neural spikes, then we can see how it is useful in gauging the change in firing rates related to alterations in visual features.

Here the Gamma process is given with the standard α and β hyperparameters which are updated with each new frame and are part of a conjugate prior over beliefs (Fig. 1). Abstractly, α corresponds to the expected value of visual features while β is the variance we have over the features as we gather them. Since we have a conjugate prior, the old α and β values are fed back into the model as the current best estimate when computing the new α and β (denoted α' and β') values for the next frame. In a sense, the new expected sample value and variance are computed by updating the old expected sample value and variance with new observations. This gives us a notion of belief as we gather more samples since we can have an idea of confidence in the estimate of our parameters. Additionally, as observations are added, so long as our beliefs are not dramatically violated, surprise will tend to relax over time. To put this another way, metaphorically speaking, the more things a surprise model sees, the less it tends to be surprised by what it sees.

2.2. Using surprise to extract image statistics from sequences

With RSVP image sequences, each image is fed like a movie into the complete surprise system one at a time. As the system observes the images, it formulates expectations about the values of each feature channel over the image frames. Like with the yellow cat example, images presented in series that are similar, tend to produce low levels of surprise. High peaks in surprise then tend to be observed when a homogeneous sequence of images gives way to a rather heterogeneous one. However, the term homogeneous is used loosely since repetitions of heterogeneous images are in terms of surprise homogeneous. That is, certain patterns of change can be expected, and believed to be normal. Even patterns of 1/f noise have an expected normalcy and as such should not necessarily be found to be surprising.

It should also be noted that surprise is measured not just between images, but within images as well. We suggest that an image can stand on its own with its inherent surprise. As a result, surprise is a combination of the spatial surprise within an image and the temporal surprise between images in a sequence. From the surprise system we can analyze surprise for different types of features as well. This is possible since the surprise model extracts initial feature maps in a manner identical to the saliency model of Itti and Koch (2001). So for instance, for any given frame, we can tell how surprising red/green color opponents are because we have a map of the red/green[1] color opponency for each frame. A surprise model is then created just to analyze the surprise for the red/green color opponents. The entire surprise system thus contains models of several feature channels which can be taken as a composite in

[1] For interpretation of color in Figs. 1–10, the reader is referred to the web version of this article.

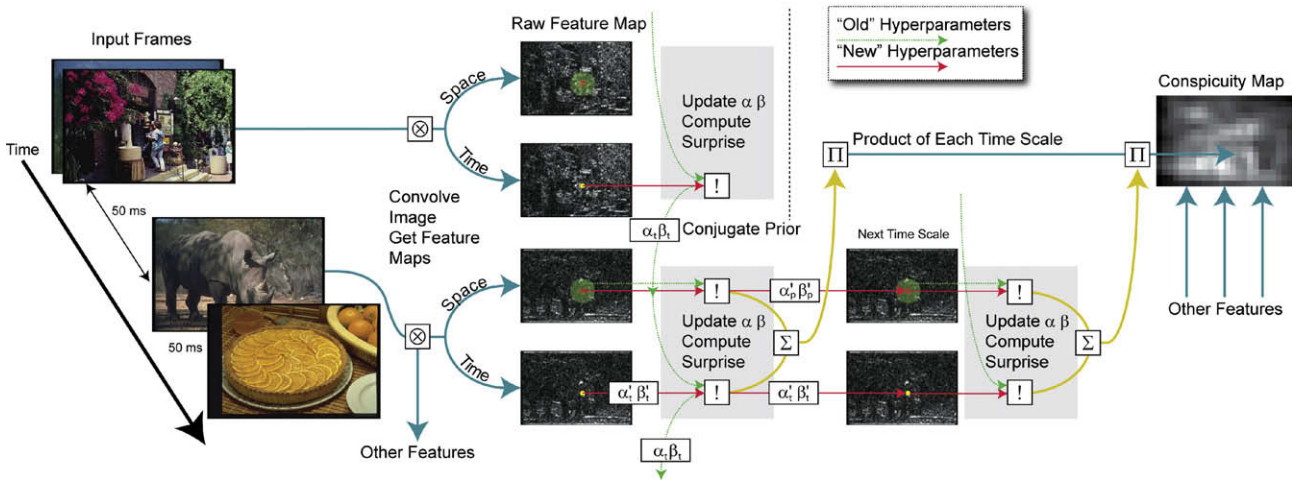


Fig. 1. With an image sequence, we run each frame through the surprise system. Feature maps such as orientation and color are extracted from the original images in the sequence (noted with "⊗"). Each feature map is then processed by surprise for both space and time. In *spatial* surprise, a locations' beliefs about what its value should be are described in the Gamma probability process with the hyperparmeters α'_p, β'_p. This is compared with its surround (α_p, β_p) within the current frame in the surprise model (noted with the box "!") (for details see Appendix A Eqs. (16)–(18)). In *temporal* surprise, each new location (α'_t, β'_t) is compared with the same location in a previous old frame (α_t, β_t) using the same update and surprise computation as space. Both space and time have their own α and β hyperparameters. Spatial and temporal surprise for each location within the same frame are merged in a weighted sum "Σ". Additionally, the α'_t, β'_t and α'_p, β'_p are successively fed forward into models several times (six times in the current model) to give it sensitivity to surprise at different time scales. Here the time scales are used to create sensitivity to recurring events at different frequencies. The complete surprise for each frame of the same feature is the product (shown with the box "Π") of merged spatial/temporal surprise for each time scale.

much the same way as the Itti and Koch saliency model does. A large number of features are supported within the surprise system but we use here intensity, red/green, blue/yellow color opponencies and Gabor orientations at 0°, 45°, 90° and 135°.

Once surprise has been determined for each model (color, Gabor orientations, etc.) in the surprise system, each image in the RSVP sequence is automatically labeled in terms of surprise statistics (Fig. 2). If desired, the resolution of surprise can be as fine grained as individual pixels. This can, as mentioned, be further parsed into the surprise system along each feature type. However, advantage may be gained in using broader aggregate descriptions of surprise. That is, for each RSVP image, we here obtain a single maximum or minimum value for surprise. This would simply be the maximum or minimum value of surprise given all the surprise values over all the model results for a frame.

Maximum surprise is a useful metric if attention uses a maximum selector method (Didday, 1976) based on surprise statistics. The maximum valued location of surprise should be the location in the image most likely to be attended. By using this metric, we can assess the likelihood that attended regions will overlap in space from frame to frame increasing their potential interaction (e.g. masking or enhancement). The overlap from this metric can

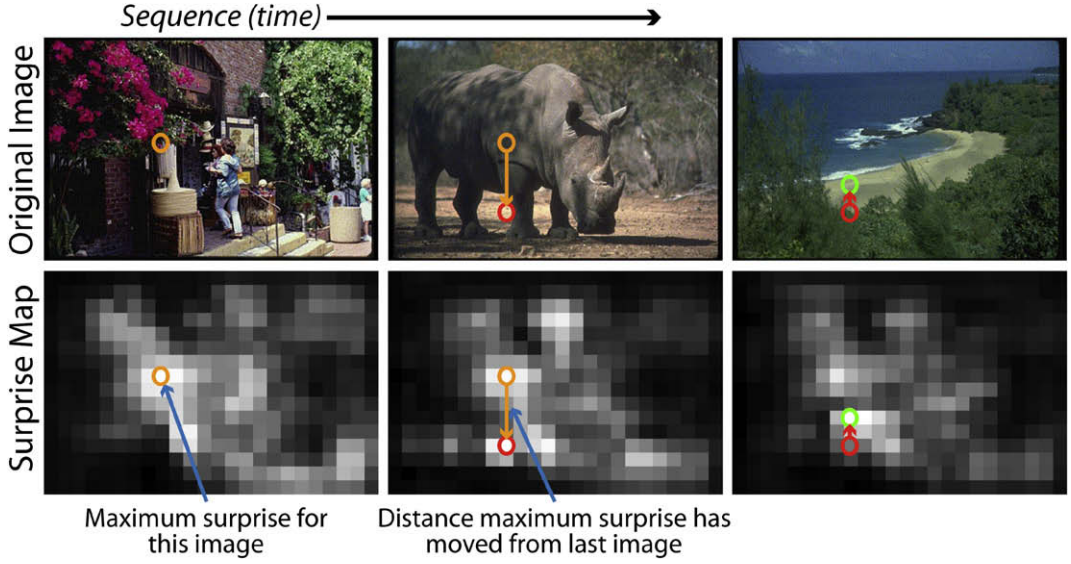


Fig. 2. The top row of images are three from a sequence of 20 that were fed into the system one at a time. The bottom row shows the derived combined surprise maps, which gives an indication of the statistical surprise the models find for each image in the sequence. Each map is a low-resolution image, which has a surprise value at each pixel, which corresponds to a location in the original image. Basic statistics are derived from the maximum, minimum, mean and standard deviation of the surprise values in each surprise map as well as the distance between maximum surprise among images giving us five basic statistic types per image. As mentioned the final surprise map is computed from each feature type independently, so while computing surprise we have access to the surprise for each feature. As an added note, the surprise for each independent image as expressed in the surprise map may look like a poor quality saliency map. However, what is surprising in each image frame is strongly affected by the images that precede it. This causes the spatial surprise component to be less visible in the output shown.

only be assumed based on proximity. That is, the closer a maximum location is between frames, the more likely two surprising locations are to overlap. However, as a spatial metric this is somewhat limiting as it does not give us a metric of the true spatial extent and distribution of surprise within a frame, for that we use the mean and standard deviation of surprise. So, within a frame, we use the point of maximum surprise to figure the attended location, but mean and standard deviation are used as a between image statistic to show the power and tendency to overlap that indicates to what extent natural images are competing (or co-operating) with each other.

The three statistics (maximum, mean and standard deviation) combined; ideally will yield a basic metric of the attentional capture an image in a sequence exhibits. Additionally, it will allow us to consider whether attention capture is competitive (Keysers & Perrett, 2002). It is then useful to analyze surprise by its overlap between images. We do this because attention capture may be more parasitic along similar feature dimensions or along more proximal spatial alignment. Put another way, interaction or interference between locations in image frames may be stronger if they are spatially overlapping or their feature statistics are highly similar. Closer spatial, temporal and feature proximity between two locations naturally lends itself to more interaction.

2.3. Performance on RSVP and its relationship with surprise

The psychophysical experiments used here to exercise the system were described in detail in (Einhäuser et al., 2007b). In brief: for the first experiment, eight human observers were shown 1000 RSVP sequences of 20 natural images. Half of the sequences contained a single target image, which was a picture of an animal. The observers' task was to indicate whether a sequence contained an animal or not. The speed of presentation at 20 Hz, yielded a comparably low performance, which was, however, still far above chance with a desirable abundance of misses (no target reported despite being present) as compared to false alarms (target reported despite not being present). This large number of missed target sequences is particularly convenient when we account for the sequences in which no subject is able to spot the target, 29 in all. This is a large enough set to allow for statistical contrast with sequences which subjects perform superior on.

Here we refine the utility of observer responses to assign a difficulty to all of the 500 target-present sequences. If one or none of the eight observers detected the target in a particular sequence, then it is classified as *hard*. If seven or eight of the human observers correctly spotted the target, the sequence is classified as *easy*. All other sequences are classified as *intermediate*. That is, if the target in a sequence of images can be spotted more often, then something about that sequence makes it easier. Note that this definition is different from (Einhäuser et al., 2007b), where only targets detected by either all or none of the observers were analyzed. By the present definition, there were 73 hard, 188 easy, and 239 intermediate sequences. As will be seen, partitioning image sequences in such a way yields a visible contrast between sequences labeled as easy and those labeled as hard. For visibility and clarity of illustration, intermediate sequences will be omitted from the graphs of the initial analysis, but will be included again in the neural network model that follows. This is reasonable since the intermediate sequences yield no additional insight into the workings of surprise, but instead make the important aspects illustrated more difficult to visually discern for the reader.

3. Results

Initial experiments showed a cause and effect relationship between surprise levels in image sequences and performance of observers on the RSVP task. Specifically, as can be seen in the example in Fig. 3A, flanking a target image with images that register higher surprise make the task more difficult for observers. Experimental support for a causal role of surprise is provided by experiment 2 in (Einhäuser et al., 2007b), in which re-ordering image sequences according to simple global surprise statistics significantly affects recall performance. To improve the power of prediction and augmentation in an RSVP task, we need to analyze the individual features to see if they yield helpful information beyond the combined surprise values such as those seen in Fig. 3B. This is because it has remained open whether prediction performance could be improved further by extracting refined spatio-temporal measures from the surprise data. In particular, does a feature-specific surprise system rather than an aggregate measure over all features further enhance prediction? Optimal feature biasing in a search (Navalpakkam & Itti, 2007) or feature priming might lend more power to certain types of features.

Analysis was carried out on all image sets which contained an animal target ("target-present" sequences). From each image in each sequence, surprise statistics were extracted for each feature

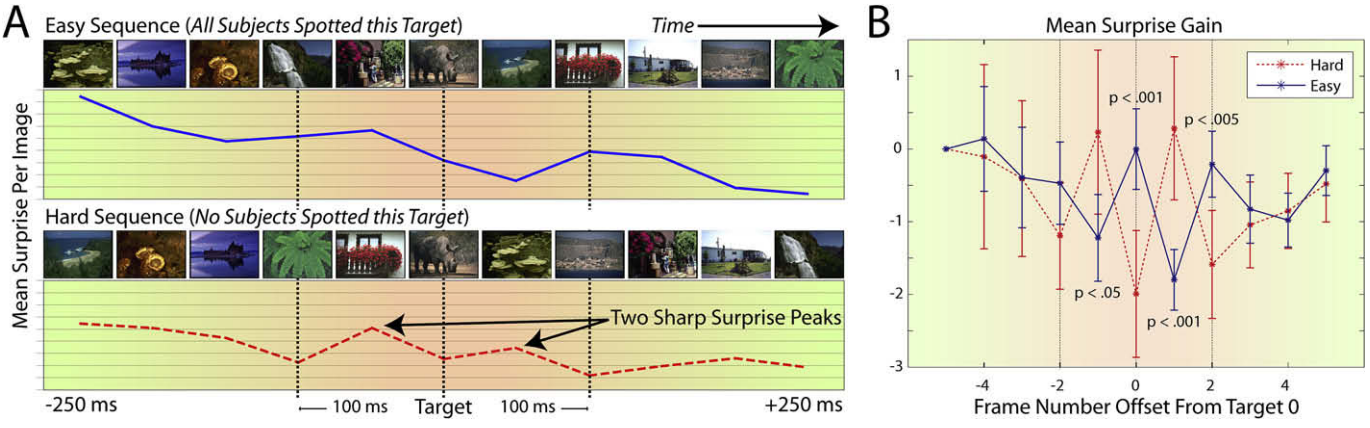


Fig. 3. (A) Mean surprise produced by each image in an example RSVP sequence is shown. The sequence on the top is easy and all 8 of 8 observers manage to spot the animal target. By rearranging the image order, we can create a hard condition in which none of the eight observers spot the animal target. This was achieved by placing images in an order which produced surprise peaks both before and after the target image at ±50 ms with relative surprise dips at ±100 ms creating a strong "M" shape in the plot centered at the target (we also refer to this as a "W" shape in the case of easier sequences or M–W for both). Note that the strength in the graphs shown is in the gain and not just the absolute value. A slight downward slope is observed since surprise can tend to relax over time given similar images. (B) The M–W shape can be seen more clearly when the gain of mean surprise for all hard and easy sequences combined is observed. In this illustration, the average over all gains for mean surprise is shown.

channel, such as color opponencies and Gabor wavelet line responses. A mean surprise is computed by taking the average of the surprise values from all locations within a frame for any given feature channel. Mean surprise of individual Gabor wavelet channels exhibit the "W" or "M" shaped interference from flanking images to the target (Fig. 4) that was also characteristic for the full system. The M–W shape is itself very similar to forms seen in masking studies as a bi-modal enhancement of Type-B masking functions (Breitmeyer & Öğmen, 2006; Cox & Dember, 1972; Reeves, 1982; Weisstein & Haber, 1965). The important thing to notice is that the "W" pattern is seen with target enhancement while the "M" pattern is seen with target masking. The significance

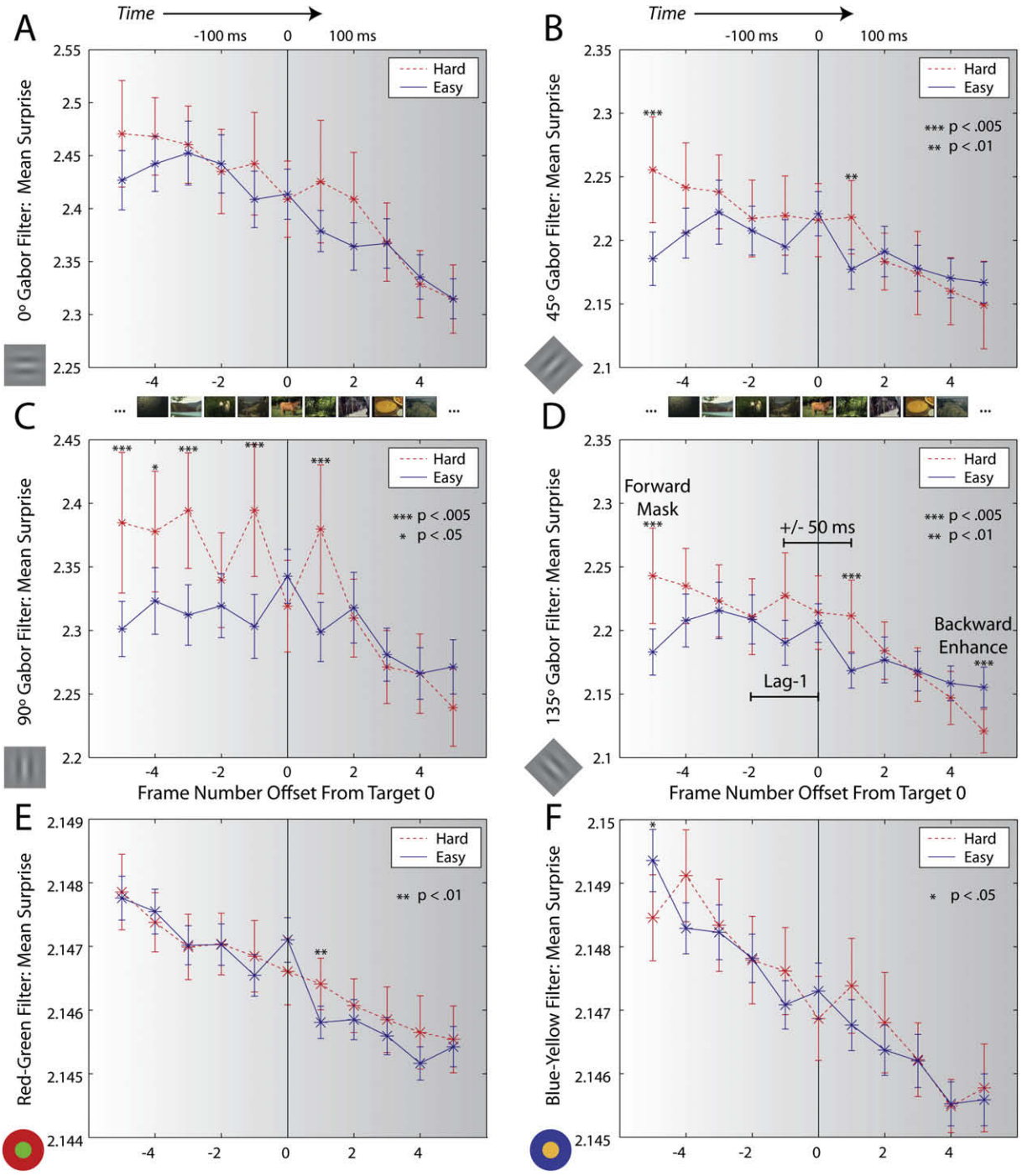


Fig. 4. Mean surprise for each frame per feature type is shown for RSVP sequences which observers found easy vs. sequences observers found hard. All sequences have been aligned similar to Fig. 3 for analysis so that the animal target offset is frame 0. (D) is labeled to show the properties of all the sub-graphs in Fig. 4. At −250 ms we see strong masking within this feature since surprise is very high for difficult sequences. −100 ms is the lag-1 sparing interval and as such we should expect to see no effect at this frame. We see at ±50 ms strong masking by elevated surprise. The effect at this interval is the most common. At +250 we see forward priming or enhancement since surprise is higher for this frame in easy sequences. All plots show the characteristic "W" and "M" shapes centered at the target, but are most visible in (C) and (D). However, there is obviously more power for vertical orientation features. Additionally, the red/green opponent channel gives M–W results like the Gabor feature channels. However, blue/yellow does not exhibit the M–W pattern (at least not significantly). Oddly, it is notable that surprise is significantly different at both ±250 ms (five frames), far beyond what would be expected for simple bottom-up feature effects. Also, the lag sparing is seen in all sub-graphs at −100 ms, even in C highlighting the strength of the effect. Error bars have been Bonferroni corrected for multiple *t*-tests.

Table 1
The significance of the gain of surprise for the target frame (0 ms) over the flanking frames (−50 ms and 50 ms) is presented. For each feature type and the combined surprise map, the mean of the two flanking images are combined and compared against the mean of the target. The combined mean surprise map statistics are the same as the ones presented in Fig. 3B. From the two means for the target and flankers, the *t* value is presented. For *P* values which are non-significant, the field is left blank. It is notable that mean, standard deviation and spatial offset of surprise spikes for easy sequences at the target frame. The peak is strongest for easy sequences, but is also notable for hard sequences. While this helps to validate the presence of the M–W pattern, it also suggests that the "W" pattern for easy targets is stronger than the "M" pattern for hard targets.

	Feature	*t* Hard	*P*	*t* Easy	*P*
Mean	Combined	2.961297	0.005	3.747544	0.001
	Gabor 0°	1.595949	∼	2.08278	0.05
	Gabor 45°	0.290076	∼	4.720356	0.001
	Gabor 90°	3.785376	0.001	4.855656	0.001
	Gabor 135^{Q}	0.454917	∼	4.137201	0.001
	Red/green	0.2679	∼	2.815474	0.005
	Blue/yellow	1.152454	∼	1.311841	∼
Stdev	Combined	1.704888	∼	7.921573	0.001
	Gabor 0°	2.158931	0.05	2.709357	0.005
	Gabor 45°	0.486151	∼	8.054697	0.001
	Gabor 90°	4.216503	0.001	5.774448	0.001
	Gabor 135°	1.161105	∼	6.014868	0.001
	Red/green	0.341483	∼	6.308344	0.001
	Blue/yellow	1.251205	∼	2.577544	0.01
Space	Combined	0.385937	∼	1.170551	∼
	Gabor 0°	0.990172	∼	1.088986	∼
	Gabor 45°	0.281247	∼	4.254909	0.001
	Gabor 90°	1.056511	∼	2.018851	0.05
	Gabor 135°	0.162157	∼	2.526581	0.01
	Red/green	0.739866	∼	2.390607	0.05
	Blue/yellow	0.352024	∼	0.036651	∼

of the M–W jump pattern for each feature types is presented in Table 1. For most feature types, and in particular, for easy sequences, mean surprise jumps significantly for the target frame when compared with the flanking frames. For targets which are difficult, there is also a significant drop in surprise compared with the flanking frames, but it is less profound.

As can be seen, Gabor surprise is increased before and after the target for "hard" as compared to "easy" sequences (Fig. 4). That is, surprise produced by edges in the image interferes with target detection. However, the difference in surprise between easy and hard targets is only significant for vertical orientations ($P < .005$ in both flanking images; Fig. 4C) and diagonal orientations (for the image following the target $P < .01$ and .005, respectively; Fig. 4B and D) at time ±1 frame (i.e. ±50 ms) from the target frame. The vertical orientation results clearly show a stronger significant probability. Horizontal orientations do not exhibit significant effects ($P \geqslant .05$ for any image in the sequence; Fig. 4A). We can also see that there is significance in the red–green color opponencies[2] (For the image following the target $P < .01$; Fig. 4E). It is notable that the effect for vertical lines not only holds for the immediate flankers of the target, but also for images that precede the target by as much as 250 ms ($P < .005$ for vertical and diagonal orientation). Conversely, there is evidence for long term backwards enhancement at 250 ms following the target which is significant for the 135° orientation channel ($P < .005$; Fig. 4D). Thus, within the sequences viewed by the observers, vertical line statistics when plotted and analyzed, seem to stand out more when observers are watching for animals among natural scenes. Interestingly, the significant jumps in surprise at ±250 ms illustrates that the effects of surprise seem too prolonged to be constrained to early visual cortex (Schmolesky et al., 1998).

Using the same analysis as for Fig. 4, a similar pattern can also be seen in Fig. 5 with the standard deviation of surprise within an image. However, the significance is concentrated on the target image itself. That is, if surprise values form a strong peak which gives a greater spread of surprise values in an image, then observers are more likely to experience that a target is easy to find. Again, the vertical line statistics yield a smaller probability value (For the target image, $P < .001$; Fig. 5C) than the horizontal line statistics (For the image after the target, $P < .05$; Fig. 5A) and red/green color opponents are significant as well ($P < .05$ for the target image and .01 for the image following the target; Fig. 5E).

That the target image benefits most from a high standard deviation of surprise is due to strong peaks in surprise which give a wider spread of values. That is, the target image may draw more attention by having basic large peaks in surprise, while masking images gain from having a broader (larger area) generally high value. In more basic terms, attention is drawn to stronger singular surprise values, while it is blocked more effectively by flanking images with a wider surprising area. One may speculate that this reflects the locality of attention: wider surprise surfaces have a better chance to overlap with the target and are thus more likely to mask it.

This hypothesis is borne out by observing the spatial information from the surprise metric. If the point of maximum surprise in the target image is offset more from the points of maximum surprise in the flanking images, observers find the target easier to spot (Fig. 6). Thus, to summarize, a more peaked surprise, with more spatial offset in the target image seems to aid in the RSVP task. Additionally, different types of features, within a given sequence, may have more power than others. However, the red/green color opponent acts differently in this particular metric. It appears that it exhibits some attentional capture within the target image itself directed away from the animal target. That is, the red/green channel shows distraction behavior within the target image itself if it is not blocked by the flanking frames.

4. A neural network model to predict RSVP performance

So far we have considered statistical measures pooled over many sequences and obtained a hypothesis on how surprise modulates recall. If the hypothesis holds true, a basic computational system should be able to predict observer performance on the RSVP task. Ideally, the system predicts, given a sequence of images,

[2] Color opponencies are derived using H2SV2 color space which is described in Mundhenk et al. (2005) with source code at: http://ilab.usc.edu/wiki/index.php/HSV_And_H2SV_Color_Space.

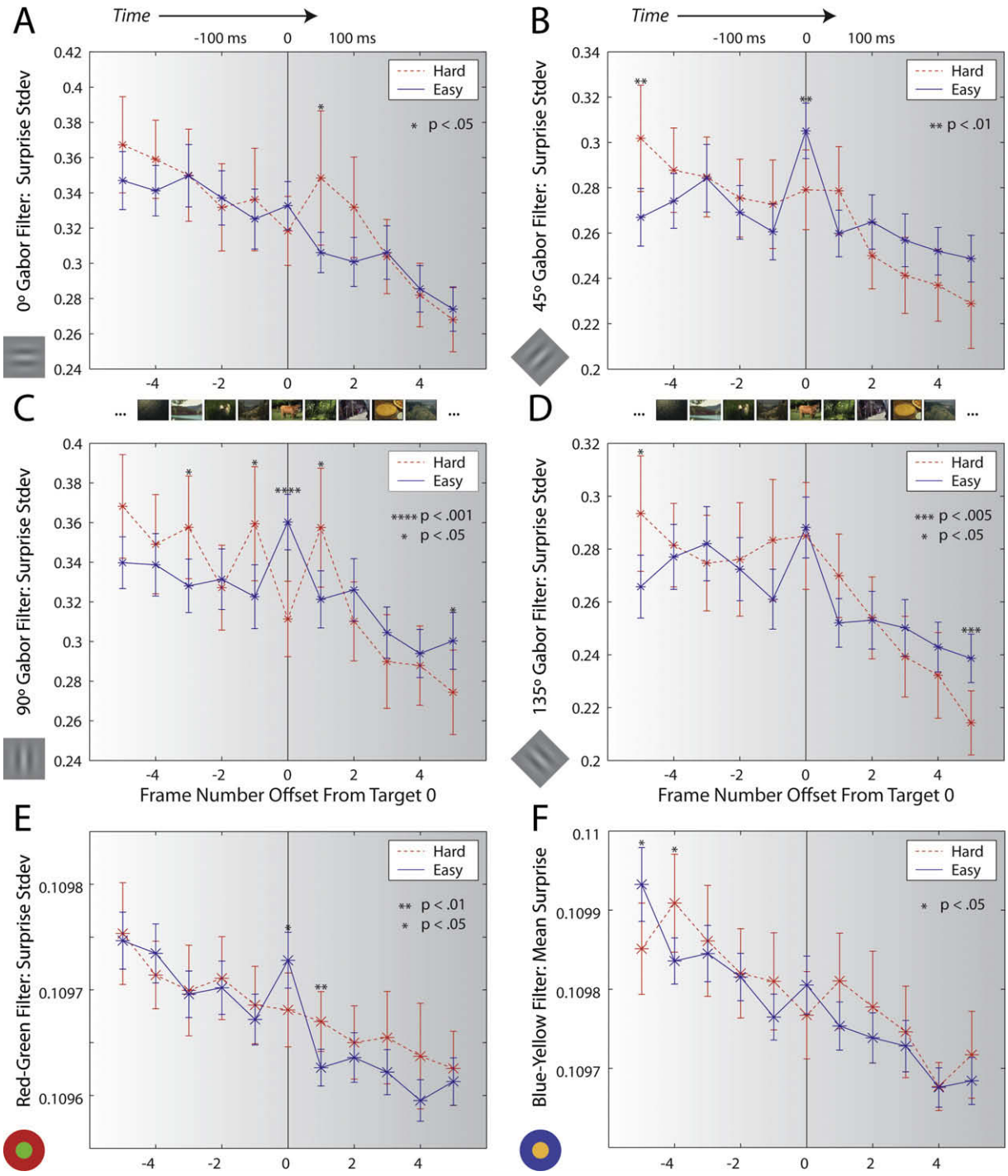


Fig. 5. The standard deviation of surprise per frame like the mean, also seems to be significant for hard vs. easy image sequences with the M–W pattern. However, significants seems more concentrated on the target image rather than the distractor images. Again, vertical lines are more prominent. Error bars have been Bonferroni corrected for multiple *t*-tests.

how well observers will detect and report them. Additionally, the system can be agnostic to whether or not a target image is in a sequence to determine if a sequence might have intrinsic memorability for *any* image for our given sequences. To do this, we added a test where we increased the difficulty for the system by testing to see if it can predict the difficulty of a sequence without knowing which image in the sequence contains the target. We added this as a comparison to our standard system which has knowledge of which frame contains the target. This also allows us to test the importance of knowing the target offset, which the evidence suggests is quite helpful in the prediction of observer performance.

Training our algorithm to predict performance on the RSVP task required three primary steps. Each step is outlined below. These include the initial gathering of human observer data on the RSVP task, processing of the image sequences to obtain surprise values, and training using back propagation on a feed-forward neural network, which attempts to give an answer as to how difficult a given sequence should be for human observers.

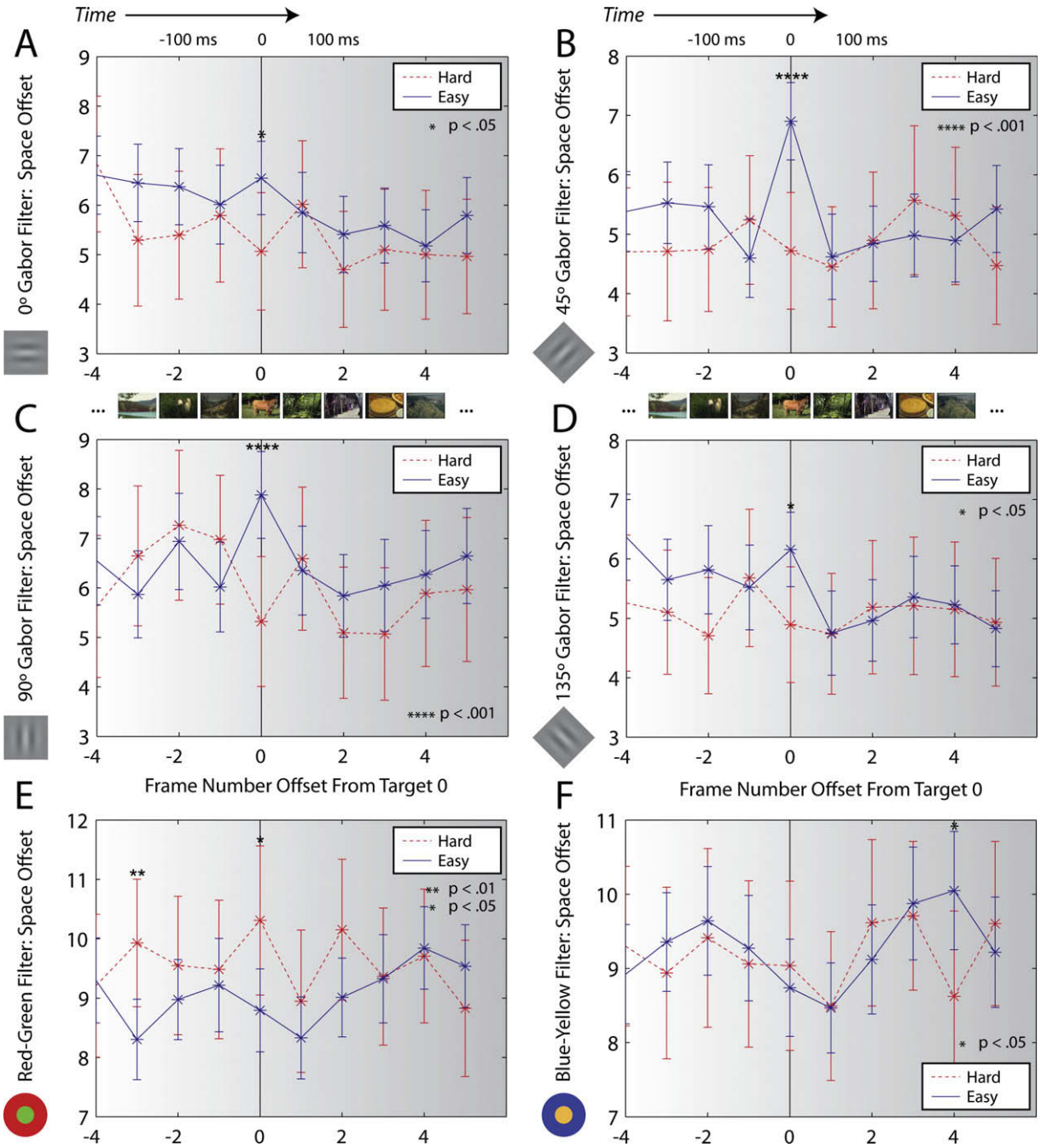


Fig. 6. Here the mean spatial offset of maximum surprise from one frame to the next (illustrated in Fig. 2) is shown for both easy and hard sequences. Frame 5 (−250 ms) is left out since it is always zero in a differential measurement. An M–W pattern is visible, but much less so than for Figs. 4 and 5. It is notable that in three of the feature types, easy sequence targets have maximum surprise peaks further in distance from the peaks in the flanking images than hard sequences do. Again, like with mean and standard deviation of surprise, the greatest power seems to be in the vertical orientation statistics. Red/green color results for space seem to be inverted suggesting that it has priming spatial effects to some extent even though strong mean red/green surprise blocks as seen in Fig. 4. Error bars have been Bonferroni corrected for multiple *t*-tests.

4.1. Data collection

RSVP data were collected from two different experiments where image sequences were shown to eight observers (Einhäuser et al., 2007b). From the two experiments, we obtained a total of 866 unique sequences of images which were randomly split into three balanced equal sized sets of 288 sequences for training, testing and final validation (two sequences were randomly dropped to maintain equal sizing for all three sets). As mentioned, each sequence could be given a rank based on how many observers correctly identified the target as being present in the sequence. The rank of difficulty was a number from zero to eight, rather than just easy, intermediate and hard. This matched how many observers managed to spot the animal target. This is used, as a training value, where the output of a network is its prediction as to how many observers in a new group of eight should correctly identify the target present in a given sequence. The control image sequences, the ones without animal targets, were not included for training since we are testing a computational model of bottom-up attention, which has no ability to actually identify an animal. The result of including non-target sequences in the system will only yield a prediction of how likely an observer is to spot a non-target which from

an attentional standpoint is the same as asking an observer to spot the target.

4.2. Surprise analysis

Each sequence was run through an algorithm, which extracted the statistics of surprise or for comparison, contrast. The contrast model (Parkhurst & Niebur, 2003), a common statistical metric which measures local variance in image intensity over 16 × 16 image patches, was used as a baseline measure. Thus, we created and trained two network models with the same image sets for contrast and surprise. The statistics were obtained by running the same set of 866 sequences through each of the systems (surprise or contrast), which returned statistics per frame.

4.3. Training using a neural network

Training was accomplished using a standard back-propagation trained feed-forward neural network (Fig. 7). The network was trained using gradient descent with momentum (Plaut, Nowlan, & Hinton, 1986). The input was first reduced in dimension to about 1/3 the original size from 528 features (listed in Fig. 7) to 170 using Principal Component Analysis (PCA) derived from the training set (Jollife, 1986). The input layer to the neural network had 170 nodes and the first hidden layer had 200 nodes. This was a number arrived at by trying different sizes in increments of 50 neurons (the validation set was not used in any way during fitting). A network with 150 hidden units has a very similar, but slightly larger error, while 250 neurons has an error which is a few percentage points larger. A second hidden layer had three nodes. For comparison, a two-layer network model was also tested. The output layer for all networks had one node. Tangential sigmoid transfer functions were used from the hidden layers. The output layer had a standard linear transfer function. Training was carried out for 20,000 epochs which is sufficient for the model to converge without over fitting. Training, testing and validation were done using three disjoint random sets with 288 samples of sequences each. The training set was used for training the network model. The testing set augmented the training set to test different meta parameters (e.g. number of neurons, PCA components). Only after all possible meta-parameter tuning was concluded, the validation set was used to evaluate performance. All results shown are from the final validation set, which was not used during training or for adjusting any of the meta-parameters (network size, PCA components, etc.).

As noted, the final output from the neural network was a number from zero to eight which represented how many observers it predicted should be able to correctly identify the target on the input sequence. That is, if of the eight observers, four registered a 'hit' on that sequence, then the target value was four. A single scalar output (a number 0–8) is used because it seems to exhibit resistance to noise in our system compared with several different network types with eight or nine outputs (e.g. a binary code using a normalized maximum). To keep the output from the neural network consistent, the output value was clamped (bounded) and rounded so that the final value given was forced to be an integer between zero and eight.

4.4. Validation and results of the neural network performance

Performance evaluation used a validation set with 288 sequences, which had not been used at any time during training or parameter setting. The idea was to try and predict how observers should perform on the RSVP task. This was done by running the validation set and obtaining general predicted performance of observers on the RSVP task. As mentioned, this was a number from zero to eight for each image sequence, depending on how many

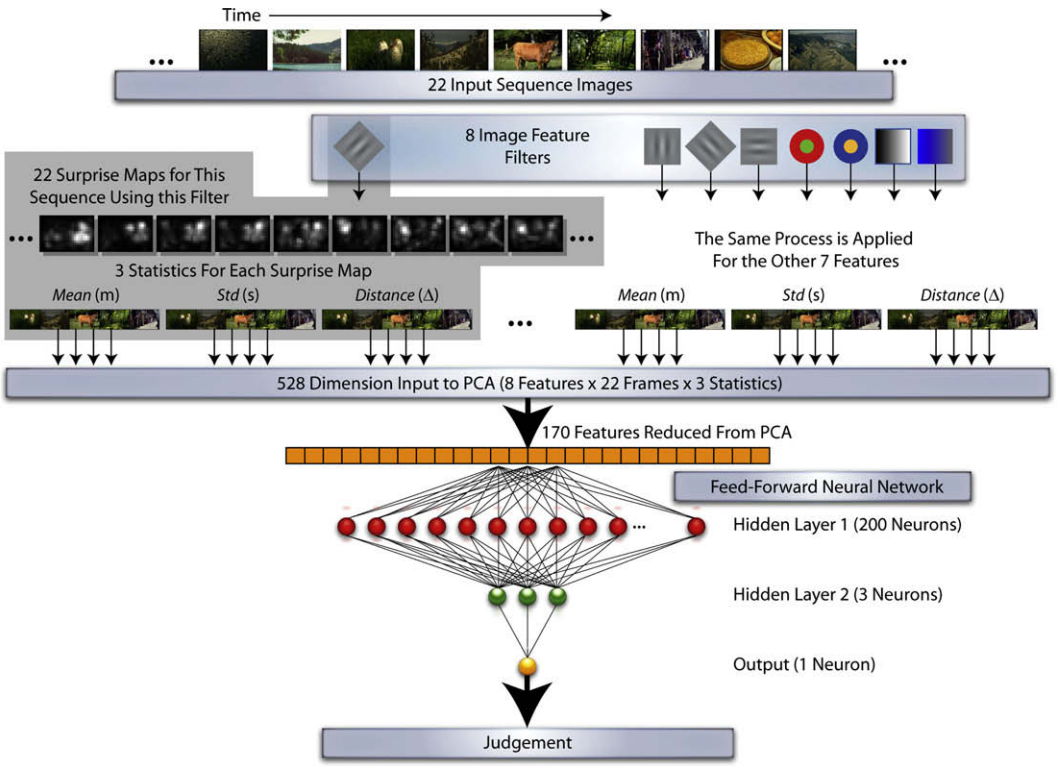


Fig. 7. The network uses the mean, standard deviation and spatial offset for maximum surprise per feature type (the same which were illustrated in Figs. 4–6) for each image in a given RSVP sequence. This is taken from 22 input frames along eight different feature types. The features are then reduced in dimension using PCA. The output is the judgment about how difficult an image sequence should be for human observers.

people the trained network expected to correctly recognize the presence of a target animal in the sequence. To make this a meaningful output, we needed to define a bound for testing the null hypothesis. In this case, we should see that the network and Surprise account for a large proportion of RSVP performance. The null hypothesis would predict that information prior to image sequence processing would be at least as sufficient as surprise for the task.

To do this, we transformed the sequence prediction into a probability. The idea was to ascertain how well the network would perform if it was asked to tell us for each image whether it expected an observer to correctly identify a target. Basically, we will compare the results of different systems, such as Surprise to that of an ideal predictor. So for instance, if the network output the number six for a sequence, then 75% of the time we would predict that an observer should notice the target (six out of eight times). Error is specified from this as how often the network will predict a hit compared with how often it should be expected to if it is an ideal predictor. This gives us an error with an applied utility since it can be used directly to measure how often our network should correctly predict subject performance. An RMSE (Root Mean Square Error) gives us a good metric for network performance, but it does not provide a direct prediction of expected error against subjects in this case.

To compute error (see Fig. 8), we find the empirical error given the network output difficulty rating and how many hits should be expected given how the network classified the sequence's difficulty. This gives us an error of expected hits by the network compared with expected hits given the actual target value for all trials in the validation set. Here, the neural network performance is compared with that of the ideal predictor. As an example, for all trials in which 4 out of 8 observers correctly spotted the target animal (a difficulty of 4), our most reasonable expectation is that observers should have a 50% chance of spotting the target. Ideally, if the neural network predictor is accurate, then it should also predict that subjects will hit on those targets 50% of the time. If however it rates some of the "50%" targets as "100%" (gives them a difficulty of 8), then the model will be too optimistic and predict a greater than 50% chance of hits for the "50%" targets.

To do this (Fig. 8), the output from the network is sorted into an $n \times m$ confusion matrix of *hits* where the rows represent the expected prior target output t and the columns represented the actual network output y. Thus, if a target in a sequence had six out of eight hits from the original human observers, this would give it a difficulty, and a target value of six. If the neural network then ranked it as a five, then it would incrementally go into cell bin $(i, j) = (5, 6)$. Where j is the prior observed for how easy this target was for subjects:

$$0 \leqslant i \leqslant 8;\ 0 \leqslant j \leqslant 8 \tag{1}$$

Expected hits by the network per cell is given by the product of the number of hits binned in a cell N^Y_{ij} by the probability that the hits will be observed given the prior observed difficulty j as $P(j)$:

$$P(j) = j/8; \quad 0 \leqslant P(j) \leqslant 1 \tag{2}$$

Symbolically we will also assume that:

$$i = j \leftrightarrow P(i) = P(j) \tag{3}$$

N^Y_{ij} is how many sequences were binned by the network into each cell bin in the confusion matrix N at cell i, j. Summing the columns of the expected hits gives the number of expected hits from the neural network given the prior determined difficulty class j and the network determined difficulty i. Note that this must be less than the total number of trials which is 288.

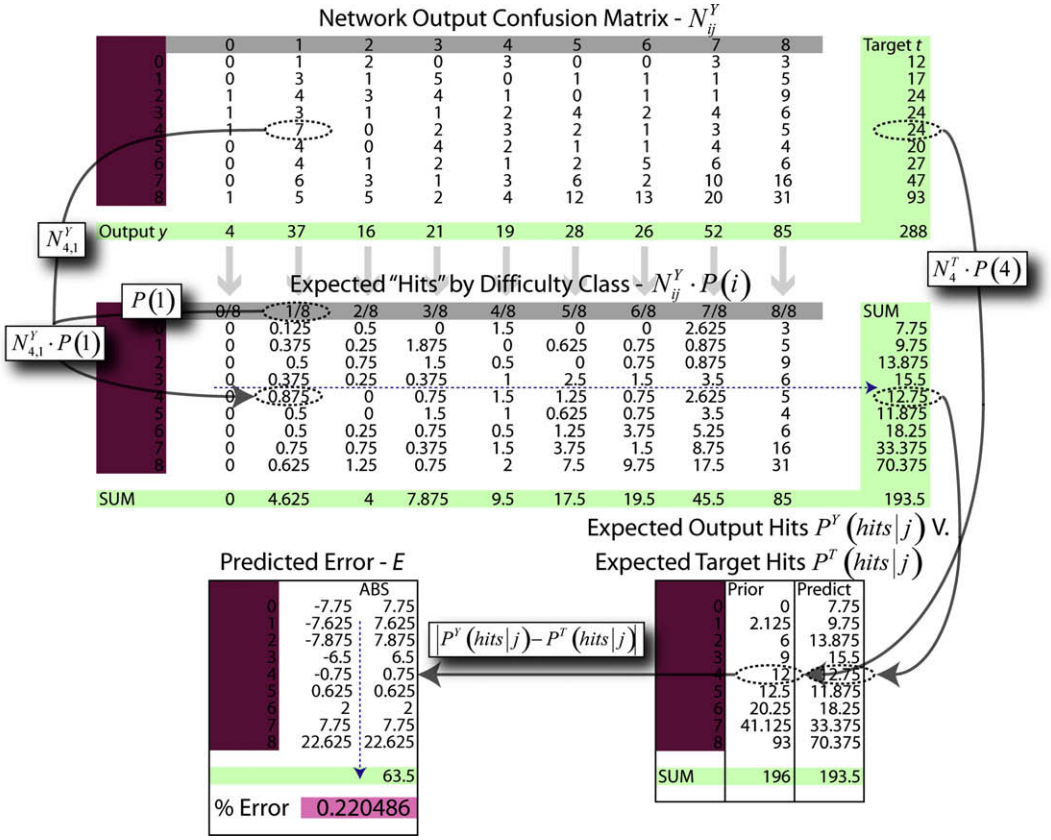


Fig. 8. An error metric is derived as how well we should be able to predict a ninth new observer's RSVP performance for the image sequences we already have data for. We compute the predicted hits for each image from the output of the neural network based on the difficulty rank it gave to each sequence. The final error is the difference between the network prediction and that of an ideal predictor.

$$P^Y(hits|j) = \sum_{i=0}^{8} N_{ij}^Y \cdot P(i); \quad 0 \leqslant P^Y(hits|j) \leqslant 288 \tag{4}$$

Also note that the *sum* of the expected hits must be less than (or equal to) the total number of trials:

$$0 \leqslant \sum_j P^Y(hits|j) \leqslant 288 \tag{5}$$

For comparison, we take the prior target data and compute its expected values given the difficulty as well:

$$P^T(hits|j) = N_j^T \cdot P(j); \quad 0 \leqslant P^T(hits|j) \leqslant 288 \tag{6}$$

A final sum error is derived by subtracting the number of expected hits as determined by the neural network y (4) with the expected number of hits from the real targets t (6). That is divided by the total number of possible hits which yields the final empirical error E_{which} is the number of expected hits in error divided by the total number of trials.

$$E = \frac{\sum_j \left| P^Y(hits|j) - P^T(hits|j) \right|}{\sum_j N_j^T} \tag{7}$$

If E is 0, then the system being tested performs the same as an ideal predictor. It should be noted, that the ideal predictor in this case is highly ideal since it is assumed to have perfect knowledge of sequence difficulty.

Baseline conditions for testing the null hypothesis were created by generating different sets of N_φ. The condition N_{naive} was created by putting the same number of hits into each bin. Thus, it is the output one would expect given a random uniform guess by the neural network. N_{bayes} was created by binning hits at the mean (mean number of hits for all sequences which is 5.44 out of 8). That is, in the absence of additional information, we will tend to guess each sequence difficulty as the *mean* difficulty. Metaphorically speaking, this is particularly important since neural networks can learn to guess the mean value when they cannot understand how to produce the target value from the input features.

It should be noted that N_{bayes} is not naïve. It still uses prior information to determine the most likely performance for sequences in the absence of better evidence. It is used as a null hypothesis since it is the best guess for observer performance prior to image processing but knowing the summary distribution of observer performance. Additionally, it presents a strong challenge to any models performance since as a maximum likelihood estimate; it has the potential to do very well depending on the nature of observer performance variance.

In addition to the surprise system, a contrast system was also trained in the same way. A range of error was computed by training both the contrast and surprise systems 25 times and computing the final validation error each time. This gives a representation of the error range produced by the initial random weight sets in the neural network training. The error is the standard 5% error over the mean and variance of the validation set error for 25 runs. It should be noted that we attempted to maintain as close a parity as possible between the training of the contrast system and the surprise system. However, there is enough difference between the two systems such that there is room for broader interpretation as to performance comparison. Since, we introduce the contrast system as a general metric and this paper is not a model comparison paper the results should not be interpreted as a critique of it.

Fig. 9 shows the expected error from both the contrast system and surprise system compared with the worst possible predictions. It can be seen that the surprise system (77.1% correct, error bars shown in figure) performs better than the contrast system (74.1% correct) and that the differences are significant within the bounds of the neural network variance. The surprise and contrast data can be combined by feeding both data sets in parallel into the network. If performance of the two combined is better than the surprise system by itself, then it suggests that the contrast system contains extra useful information which the surprise system alone does not have. However, the combined model (76.6% correct) does not perform better than the original surprise system. Its performance appears slightly worse, but is not significantly different than the surprise system alone.

Both the contrast and surprise systems perform much better than the worst possible Naïve model (67.1% correct) and noticeably better than the null hypothesis Bayes model (71.7% correct). It is important to note that while neither the surprise nor contrast systems perform perfectly, they should not be expected to do so since they have no notion of the difficulty of actually recognizing the particular animal in a target image. They are only judging difficulty based on simple image statistics for attention in each sequence. Additionally, as mentioned, the task was made more difficult by withholding knowledge of the position of the target in the sequence, from the training system. It can be seen that when the target frame is known, the model performs its best at 79.5% correct.

We can put this performance in perspective by using the Naïve model as a baseline. We compute a standard performance increase f'_{bayes} as the difference between the performance of the Naïve model f_{naive} and the Bayes model f_{bayes}.

$$f'_{bayes} = f_{bayes} - f_{naive} \tag{8}$$

The same metric for the surprise system is:

$$f'_{surprise} = f_{surprise} - f_{naive} \tag{9}$$

The gain of the surprise system over the Bayes model is computed as:

$$g_{surprise} = \frac{f'_{surprise}}{f'_{bayes}} \tag{10}$$

This tells us that the performance increase of the surprise system is 2.7 times that of the Bayes model.

5. Discussion

The results shown in Figs. 3–6 are not unexpected given that the M–W curves closely resemble the timing and shape of both visual masking (Breitmeyer & Öğmen, 2006) and RSVP attentional blink study results (Chun & Potter, 1995; Einhäuser et al., 2007a; Keysers & Perrett, 2002; Raymond et al., 1992). What is interesting is that surprise elegantly reveals the masking behavior in natural images. This allows efficient use of natural scenes in vision research since low level bottom-up masking behavior can be accounted for and controlled. As an added note on computational efficiency, a single threaded surprise system will process several images a second with current hardware. This is important since it has allowed us to extract surprise from thousands of sequences. Since the surprise system is open source and downloadable, other researchers should be able to take advantage of it to create RSVP sequences with a variety of masking scenarios. Ease of use has also been factored into the surprise system since it is built upon the widely used saliency code of Itti and Koch (2001).

In addition to the basic metric of surprise, when we combine it with neural network learning, it predicts which RSVP sequences of natural images are difficult for observers from those that are easy. This provides two major contributions. The first is that surprise has been shown to deliver a good metric for visual detection and recall experiments. Thus, a variety of vision experiments can be con-

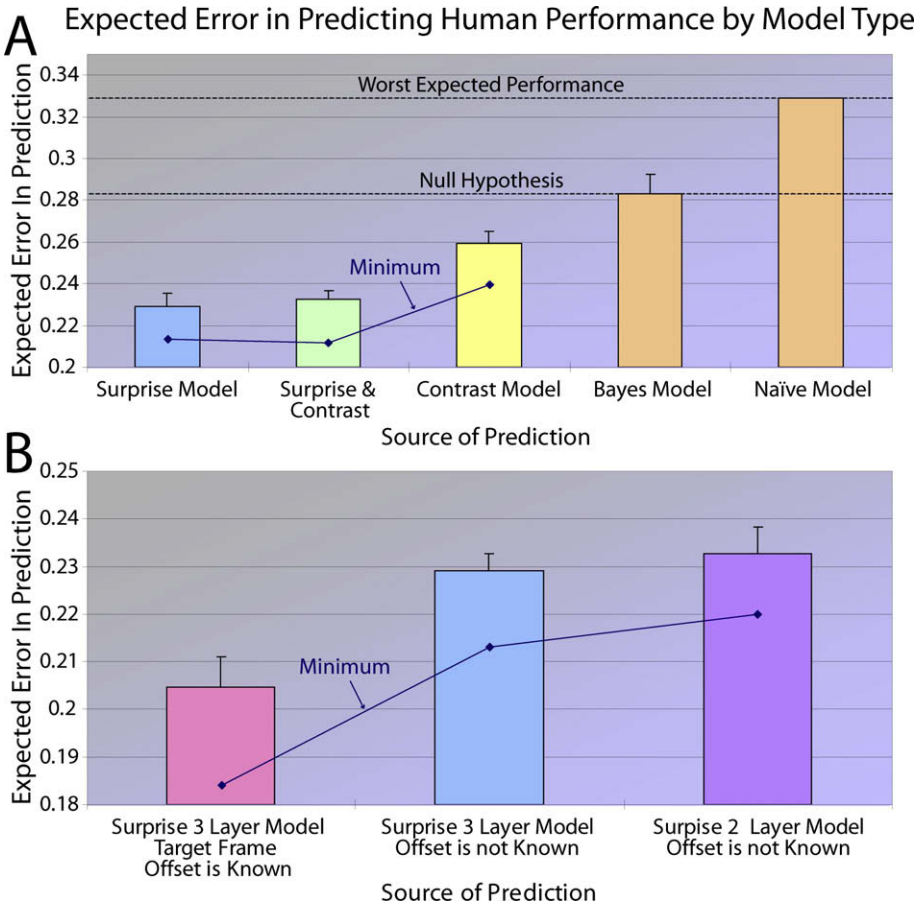


Fig. 9. (A) is the expected error for prediction in performance for the RSVP task. Error bars represent the expected range of performance for the neural networks trained on the task. Both the surprise system and the contrast model perform much better than the null hypothesis would predict. Combining the image surprise data and contrast data does not yield better results. (B) shows that a three-layer network performs slightly better than a two-layer network. Additionally, as expected, when the frame offset for the target is provided to the network, performance increases further since it can now look for the characteristic M–W patterns in the flanking images as seen in Figs. 4–6. Note that (A) and (B) show the same type of information and can be directly compared.

ducted, which can use surprise to gain insight into the intrinsic attentional statistics, especially for natural images. For instance, as we have done, surprise was successfully manipulated by re-ordering images in sequences, to modulate recall (Einhäuser et al., 2007b). Here we have extended on this work by creating a non-relative metric that can compare difficulty for 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 gives us a parametric and absolute measure based on how many observers should be able to spot a target.

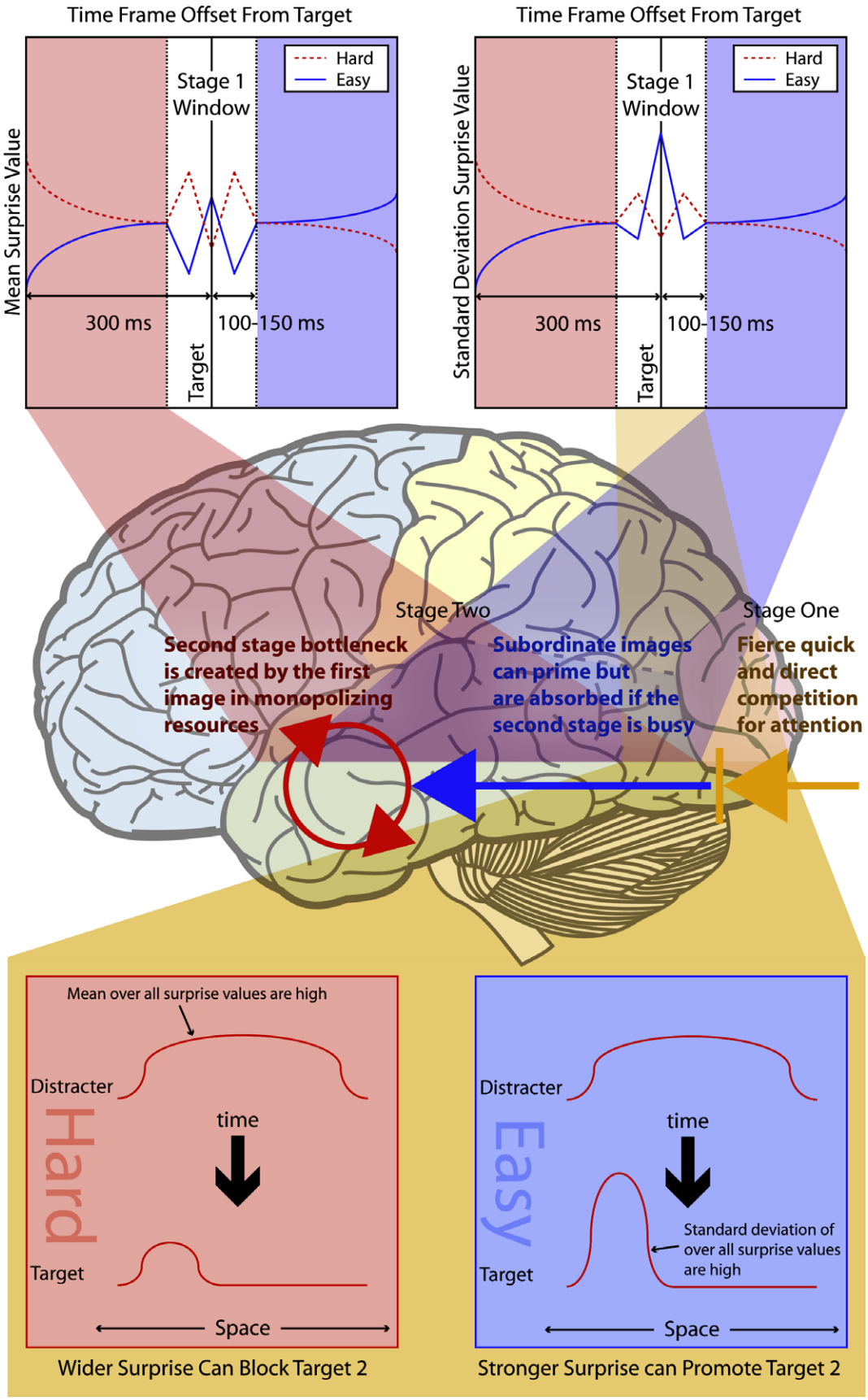
The second contribution is that a system can be created to predict performance on a variety of detection tasks. In this article, we focused on the specific RSVP task of animal detection. However, the stimulus which we used is general enough as to suggest that the same procedure could be used for many other types of images and sequences.

5.1. Patterns of the two-stage model

While the interaction between space, time and features is not completely understood, the interactions within these domains are well illustrated by our experimental results. Surprise in flanking images aligned spatially and along feature dimensions block the perception of a target particularly in a window of ±50 ms around the target image. Additionally, the time course is consistent at least with respect to Gabor statistics, which show the M–W pattern of attentional capture and gating. However, we must note that this study does not explicitly establish a constant time course for any image sequence, but there is some expectation of the time course to be bounded based on the reviewed masking, RSVP and attentional blink experiments.

With significant peaks at widely divergent times, our results are compatible with the two-stage model described by (Chun & Potter, 1995) (for a conceptual diagram, see Fig. 10). In the ≈ 50 ms sur-

Fig. 10. The top row is a hypothetical summary of Figs. 4 and 5. Flanking images block the target image based on increase in mean surprise within 100 ms and show the characteristic M–W pattern. A strong tail is observed as much as 250 ms prior. Notably, the flat portion at 100 ms is at the lag sparing interval for attentional blink. This suggests that a two-stage model similar to (Chun & Potter, 1995) is at work. In the first stage fierce competition creates a mask with the power of an images surprise. Wider masks are more likely to block information, while stronger surprise in an image is more likely to break through a prior blocking mask even if it is wide. In the later stages, parts of images which have passed the first stage are assembled into a coherent object or set of objects. Once the assembly is coherent enough, subordinate information from new images can help prime it and make it more distinguishable. However, information from the subordinate images will be absorbed if the second stage has reached a critical coherence. This can be summarized by stating that the effects from surprise in the first stage are a result of direct competition and blocking while the effects seen in the latter stages are a side effect of the result of priming, absorption or integration.


Time Frame Offset From Target
Mean Surprise Value
Hard
Easy
Stage 1 Window
300 ms
100-150 ms
Target
Time Frame Offset From Target
Standard Deviation Surprise Value
Hard
Easy
Stage 1 Window
300 ms
100-150 ms
Target
Stage Two
Second stage bottleneck is created by the first image in monopolizing resources
Subordinate images can prime but are absorbed if the second stage is busy
Stage One
Fierce quick and direct competition for attention
Mean over all surprise values are high
Distracter
time
Hard
Target
Space
Wider Surprise Can Block Target 2
Distracter
time
Easy
Standard deviation of over all surprise values are high
Target
Space
Stronger Surprise can Promote Target 2

rounding the target, it is easier for images with targets of higher surprise to directly interfere with each other. Since targets can both forward and backward mask at this interval, competition to pass the first stage is more direct. Already present targets can be suppressed by new more surprising ones or block less surprising targets from entering. Then we see a period of about 100 ms where the target image is in a sense safer, this is at the same point in time that lag sparing is observed in attentional blink studies. After that ($\geqslant$200 ms), the data shows that forward masking can block the perception of the target. Conversely, the data indicates that backwards masking does not occur at this interval, but instead we see enhancement.

The results seen here might be a result of the manner in which the visual cortex gates and pipelines information and may be thought of as a two or three stage model seen in Fig. 10 (stages two and three may be the same stage, but we parse them here for illustrative purposes). (1) Very surprising stimuli capture attention from other less surprising stimuli at ±50 ms. Additionally, competition is spatially local allowing for some types of parallel attention to occur. That is, some image features can get past the first stage in parallel if they are far enough apart from each other. Parallelism, while not illustrated by the data presented here is expected based on the reviewed studies on split attention (Li et al., 2002; McMains & Somers, 2004; Rousselet et al., 2002, 2004). (2) Stimuli is pipelined at 100 ms to the next stage of visual processing. This leads to a lag sparing effect or the safety period observed in Figs. 4 and 5. This is also observed in the literature we have reviewed which illustrates the lag-1 sparing in RSVP (Raymond et al., 1992) or a relaxed period of masking with a Type-B function (Breitmeyer, 1984; Breitmeyer & Öğmen, 2006; Cox & Dember, 1972; Hogben & Di Lollo, 1972; Reeves, 1982; Weisstein & Haber, 1965). The pipeline may do further processing, but is highly efficient and most importantly avoids conflicts. (3) After 150 ms, a new processing bottleneck is encountered. Unlike (1) which is a strict attention gate meant to limit input into the visual system, we suggest as others have (Chun & Potter, 1995; Sperling et al., 2001) that (3) is a bottleneck by the fact of its intrinsic limited processing resources. Conflict can occur if an image that arrived first is not finished processing when a second image arrives. The situation is made more difficult for the second image if the first image has sequestered enough resources to prevent the second image from processing. The second image in may be integrated or blocked while at the same time enhancing the first image into the third stage. That is, an image or set of targets may dominate visual resources after $\approx$250 ms, causing new input stimuli into a subordinate role. This later stage may also resemble strongly several current theories of visual processing (Dehaene, Changeux, Naccache, Sackur, & Sergent, 2006; Lamme, 2004) if it is thought of in terms of a global workspace or as visual short term memory (Shih & Sperling, 2002).

To further refine what we have said, the first stage is selective for images based more strongly on surprise and is not strictly order-dependant, which is why Figs. 4 and 5 show the strong M–W pattern. The utility of this is perhaps to triage image information so that the most important nuggets are transported to later stages which have limited processing capability. The third stage is selective based on more complex criteria such as order which is why we observe the asymmetry at ±250 ms in Figs. 4 and 5. However, evidence also suggests that if a second target is salient enough, it may be able to co-occupy later stages of visual processing with the first target allowing for detection of both targets in RSVP (Shih & Reeves, 2007). Importantly, the relationship between the first target and the second target appears to be asymmetric in the later stages, which was illustrated in Fig. 4 from the surprise system at the ±250 ms extremes.

Surprise and attention in the first stage does not necessitate that later stages do not also contain attention mechanisms. An attentional searchlight (Crick, 1984; Crick & Koch, 2003) theoretical model could allow for attention at many levels. That is, as information is processed, refined and sent to higher centers of processing, additional attention processing can allow a continuation of the refinement. Selection can be pushed from the bottom-up or pulled from top-down processes. Two-stage models which place attention in a critical role in later stages have been proposed. For instance, (Ghorashi, Smilek, & Di Lollo, 2007) suggest that lag sparing in RSVP attentional blink are due to a pause while the visual system is configured for search in latter stages of processing. While it is unclear if a pause causes lag sparing from the research presented, significants of specific features we have illustrated from surprise at ±250 ms might lend credence to feature specific attention in later stages,

5.2. Information necessity, attention gating and biological relevance

The results show that the gating of information in the visual cortex during RSVP may be based on a sound statistical mechanism. That is, target information which carries the most novel evidence is of higher value and is what is selected to be processed in a window of 100 ms. Target information separated by longer time intervals may also be indirectly affected by surprise because of interference in later stages of processing. This happens if more than one overlapping target is able to enter later visual processing at the same time. Since surprise effects are indirect at this stage, surprise may never be able to account for all RSVP performance effects.

The mechanism for surprise in the human brain we believe is based on triage. Target information is held briefly during early processing allowing a second set of target features in the following stages to also be processed. Masking at the shortest interval happens when the two sets of image data are competed against each other. If the new target information is more surprising, it usurps the older target information (attention capture), otherwise it is blocked. Masking in this paradigm depends on competition within a visual pipeline. Its purpose is to create a short competition which may even introduce a delay for the sake of keeping the higher visual processes from becoming overloaded. After 50 ms, if target information is still intact, it moves beyond this staging area and out of the reach of competition from very new visual information. Masking and the loss of target detection after that point, might be based on how the visual system constructs an image from the data it has. After 150 ms, image data is lost in assembly by virtue of an integrative mechanism in a larger working memory or global workspace.

Interestingly, if surprise does account well for the Type-B masking effects many have observed, it suggests that masking in the short interval is not a result of a flaw in the visual system, but rather it is the result of an internally evolved mechanism to directly filter irrelevant stimuli. That is, in the ±50 ms bounding a target, masking is an effect of an optimal statistical process. The absence of masking, if it was observed, would indicate that too much information is being processed by the visual system.

5.3. Generalization of results

Returning to the topic of feed-forward neural networks, another facet to mention is on the limitations of generalization for networks such as the one we used here. Validation was carried out using images of the same type. As such, we do not necessarily know that it will perform the same for a completely different type of image set. As an example, one might imagine target pictures of

rocks with automobile distractors. However, evidence suggests that at less than 150 ms only coarse coding is carried out by the visual system (Rousselet et al., 2002). Thus, we would expect generalization to other types of images so long as coarse information statistics are similar. Additionally, the input to the neural network was extremely general and contained a relative lack of strong identifying information. That is, we only used whole image statistics of mean, standard deviation and spatial offset for surprise. It is quite conceivable that images of other types of targets, and distractors, may very well exhibit the same basic statistics and, as a result, the neural network should generalize.

5.4. Comparison with previous RSVP model prediction work

An important item to note here is the difference between the system we have presented and the one presented by Serre et al. (Serre, Oliva, & Poggio, 2007). The primary difference is that our system is based on the temporal differentiation across images in a sequence whereas the model by Serre et al. is an instantaneous model. That is, it has no explicit notion of the procession of images in a sequence, which as we have shown has effect on the performance of RSVP. This is particularly true since the same images in different order around the target can affect performance on the RSVP task (Einhäuser et al., 2007b).

However, the RSVP task performed by the Serre model is sufficiently different, that any real cross-comparison between the two models is difficult on anything other than a superficial level. Indeed, the Serre model has much better target feature granularity, which perhaps gives it better insight into feature interactions within an image. Thus, one may be able to produce an even better RSVP predictor by combining the Serre model with the one we presented here, thereby exploiting both sequence and fine granularity aspects.

5.5. Network performance

It is also interesting to note that with the neural network, the contrast system performed almost as well as the surprise system. However, when the surprise and contrast data are combined, performance is not increased suggesting that the contrast system does not contain any useful new information that the surprise system does not already have. That is, given that the combined surprise/contrast system performs almost exactly the same as the surprise system by itself, and that the contrast system still performs much better than chance, suggests that the surprise system intrinsically contains the contrast system's information (indeed, intensity contrast, computed with center-surround receptive fields, is one of the features of the surprise system which is similar but not the exact same thing as the contrast system). Otherwise, we should expect to see an improvement in the performance of the combined surprise contrast system.

Another network performance issue that should be mentioned is that in one condition, the neural network performance was hindered by forcing it to not know the target offset in each sequence. This was done to see if a sequence of images could be labeled with a general metric so that we could conceivably gauge the recall of any image in the sequence, not just the target. However, if knowing the target offset frame is a luxury one can afford, then we illustrated that prediction performance can be improved even further. That is, knowing which frame is the target frame in advance, aids in prediction, but the network can make a prediction, though less accurate, otherwise. This is to be expected given the M–W shape for surprise as seen in Fig. 10 where surprise from the flanking images yields the most information to RSVP performance on the target.

5.6. Applications of the surprise system

In addition to the theoretical implications of our results there are also applications which can be derived. For instance, the surprise analysis could be used to help predict recall in movies, commercials or – by adjustment of timescale – even magazine ads. Since it gives us a good metric of where covert attention is expected, it may also have utility in optimizing video compression to alter quality based on where humans are expected to notice things (Itti, 2004). As long as images are viewed in a sequence, the paradigm should hold. Even large static images, whose scanning requires multiple eye-movements, effectively provide a stimulus sequence in retinal coordinates and will be an interesting issue for further research.

6. Conclusion

Surprise is a useful metric for predicting the performance of target recall on an RSVP task if natural images are used. Additionally, in past experiments, we have shown that this can be used to augment image sequences to change the performance of observers. An improved system of prediction can be made using a feed-forward neural network, which integrates spatial, temporal, and feature concerns in prediction, and can even be done so without knowing the offset of the target in the sequence. Type-B and spatial masking effects are revealed by surprise which provides further evidence that RSVP performance is related to masking and attentional capture. This also suggests that the masking and attentional capture are accounted for in a statistical process, perhaps as a triage mechanism for visual information in the first stage of a two-stage model of visual processing.

Appendix A

Given a set M of possible models (or hypotheses), an observer with prior distribution $P(M)$ over the models, and data D such that:[3]

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)} \tag{11}$$

Notably, the data can in this case be many things such as a pixel value or the value of the output of a feature detector. We are also not constrained to just vision; a variety of data is useful. To quantitatively measure how much effect the data observation (D) had on the observer beliefs (M), one can simply use some distance measure d between the two distributions $P(M|D)$ and $P(M)$. This yields the definition of surprise as (Itti & Baldi, 2005, 2006):

$$S(D,M) = d[P(M|D), P(M)] \tag{12}$$

One such measure is the Kullback–Leibler (KL) distance (Kullback & Leibler, 1951) generically defined as:

$$L = -\int p(\mathbf{x}) \ln \frac{\tilde{p}(\mathbf{x})}{p(\mathbf{x})} dx \tag{13}$$

By combining (13) and (11) we get the surprise given the data and the model as:

$$S(D,M) = \log P(D) - \int_M P(M) \log P(D|M) dM \tag{14}$$

[3] A general surprise framework is downloadable for Matlab from http://sourceforge.net/projects/surprise-mltk. The full vision code in gcc is downloadable from http://ilab.usc.edu.

Surprise can have as its base a variety of statistical processes, so long as the KL distance can be derived. In this work, we used a Gamma/Poisson distribution since it models neural firing patterns (Ricciardi, 1995) and since it gives a natural probability over the occurrence of events. We can define the Gamma process as:

$$P(M(\lambda)) = \gamma(\lambda;\alpha,\beta) = \lambda^{\alpha-1}\frac{\beta^{\alpha}e^{-\beta\lambda}}{\Gamma(\alpha)} \text{ for } \lambda > 0 \tag{15}$$

and its corresponding KL distance as:

$$\mathrm{KL}(\gamma(\lambda;\alpha',\beta'),\gamma(\lambda;\alpha,\beta)) = -\alpha' + \alpha\log\frac{\beta'}{\beta} + \log\frac{\Gamma(\alpha)}{\Gamma(\alpha')} + \beta\frac{\alpha'}{\beta'} + (\alpha'-\alpha)\Psi(\alpha') \tag{16}$$

Here we note that Γ is the gamma function (not to be confused with the gamma probability distribution given in (15) and Ψ is the digamma function which is the log derivative of the gamma function Γ. Surprise itself in this case is the KL distance.

To make this model work, we need to know what α and β are and how to update them to get α' and β'. Importantly, in this notation, α' is just the update of α at the next time step. Given some input data d such as a feature filter response, we can derive α', which is abstractly analogous to the *mean* in the normal distribution. It is given as:

$$\alpha' = \alpha\cdot\zeta + \frac{d}{\beta} \tag{17}$$

Here ζ is a decay *forgetting* term which has been set as 0.7 based on earlier work that suggests this is a good value (Itti & Baldi, 2006). To reiterate, this assumes that the underlying process is Poisson. Otherwise one must use a more complex method to compute α' such as the *Newton–Raphson* method (Choi & Wette, 1969) or use a completely different process for computing surprise such as a Gaussian process.

The value of β', which is abstractly analogous to the *standard deviation* in the normal distribution can be computed as:

$$\beta' = \beta\cdot\zeta + 1 \tag{18}$$

In this formulation, β is independent of new data for temporal surprise, but not for spatial surprise which computes the β term from the image surround. (18) is sufficient for analysis of RSVP sequences since model evidence collection can start in a naïve state. Otherwise, for long sequences such as movies, β should be allowed to float based on longer term changes in the state of evidence.

To deal with different time scales which have different gains, the results from a model are fed forward into successive models from 1 to i where i is the ith model out of n time scales ($1 \leqslant i \leqslant n$). In this case, six time scales are used. For each scale we will compute a different α' and thus a different surprise value. For each time scale, this takes the form of:

$$\alpha_1' = \alpha_1\cdot\zeta + \frac{\cdots^{\alpha_2\cdot\zeta+\frac{\alpha_1\cdot\zeta+\frac{d}{\beta_1}}{\beta_2}}}{\beta_1} \tag{19}$$

Over a series of images, surprise can be computed in temporal and spatial terms. Temporal surprise is a straight forward computation with updates (17) and (18) where d is the value of a single location in an image and α is the hyper parameter computed on the previous frame as α'.

Spatial surprise is computed with d also being the value of a single location, but α and β are treated like the Gamma mean and variance given values of the surrounding locations in an image. This makes spatial surprise the surprising difference between a location and its surround. Given the data from m locations in the surround D and locations j ($j \in D$) and a Difference of Gaussian (DoG) weighting kernel w, we compute a weight sum W of the kernel as:

$$W = \sum_{j=1}^{m} w_j \tag{20}$$

The spatial variance $\bar{\beta}$ is computed as:

$$\bar{\beta} = \frac{1}{W}\cdot\sum_{j=1}^{m}\beta_j\cdot w_j \tag{21}$$

Since this operates at multiple time scales, all β_j are initialized to 1 for the first time scale, but are set to the previous time scales β' value thereafter.

The expected value at an image location $\bar{\alpha}$ computed from $\bar{\beta}$ is:

$$\bar{\alpha} = \frac{\bar{\beta}}{W^2}\cdot\sum_{j=1}^{m}\alpha_j\cdot w_j \tag{22}$$

Again, α_j is initialized as the data value for the surrounding image locations in the first time scale, but is set to the previous value as in (19) after that. The values $\bar{\alpha}$ and $\bar{\beta}$ can then be plugged into (17) and (18) for α and β respectively which allows surprise to be computed as in (16) for space at each location in an image.

Total surprise for a feature channel S at each location is then computed as the product of surprise for the sum of temporal t and space p across all time scales n:

$$S = (S_{p1} + S_{t1}\cdot\ldots\cdot S_{pn} + S_{tn})^{1/(3\cdot n)} \tag{23}$$

Each surprise map S for each feature type can be combined into a single surprise map by taking a weighted average over all the surprise maps. The combined surprise map is created from individual surprise maps in the same way we have done in the past with saliency maps (Itti & Koch, 2001; Itti, Koch, & Braun, 2000; Itti, Koch, & Niebur, 1998) to create a combined saliency map.

References

Breitmeyer, B. G. (1984). *Visual masking*. New York: Oxford University Press.

Breitmeyer, B. G., & Öğmen, H. (2006). *Visual Masking: Time slices through conscious and unconscious vision*. New York: Oxford University Press.

Choi, S. C., & Wette, R. (1969). Maximum likelihood estimation of the parameters of the gamma distribution and their bias. *Technometrics, 11*(4), 683–690.

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

Cox, S., & Dember, W. N. (1972). U-shaped metacontrast functions with a detection task. *Journal of Experimental Psychology, 95*, 327–333.

Crick, F. (1984). Function of the thalamic reticular complex: the searchlight hypothesis. *PNAS, 81*, 4586–4590.

Crick, F., & Koch, C. (2003). A framework for consciousness. *Nature Neuroscience, 6*(2), 119–126.

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences, 10*(5), 204–211.

Di Lollo, V., Enns, J. T., & Rensink, R. A. (2000). Competition for consciousness among visual events: the psychophysics of reentrant visual processes. *Journal of Experimental Psychology: General, 129*(4), 481–507.

Didday, R. L. (1976). A model of visuomotor mechanisms in the frog optic tectum. *Mathematical Biosciences, 30*, 169–180.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General, 113*(4), 501–517.

Duncan, J., Ward, R., & Shapiro, K. (1994). Direct measurement of attentional dwell time in human vision. *Nature, 369*(26), 313–315.

Einhäuser, W., Koch, C., & Makeig, S. (2007a). The duration of the attentional blink in natural scenes depends on stimulus category. *Vision Research, 47*, 597–607.

Einhäuser, W., Mundhenk, T. N., Baldi, P., Koch, C., & Itti, L. (2007b). A bottom-up model of spatial attention predicts human error patterns in rapid scene recognition. *Journal of Vision, 7*(10), 1–13.

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*(6), 1476–1492.

Ghorashi, S. M. S., Smilek, D., & Di Lollo, V. (2007). Visual search is postponed during the attentional blink until the system is suitably reconfigured. *Journal of Experimental Psychology: Human Perception and Performance, 33*(1), 124–136.

Hoffman, J. E., Nelson, B., & Houck, M. R. (1983). The role of attentional resources in automatic detection. *Cognitive Psychology, 15*(3), 379–410.

Hogben, J., & Di Lollo, V. (1972). Effects of duration of masking stimulus and dark interval on the detection of a test disk. *Journal of Experimental Psychology, 95*(2), 245–250.

Itti, L. (2004). Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing, 13*(10).

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

Itti, L., & Baldi, P. (2006). Bayesian surprise attracts human attention. In *Advances in 1215 Neural Information Processing Systems (NIPS)* (Vol. 19, pp. 547–554): MIT Press.

Itti, L., & Koch, C. (2001). Computational modeling of visual attention. *Nature Reviews Neuroscience, 2*(3), 194–203.

Itti, L., Koch, C., & Braun, J. (2000). Revisiting spatial vision. Towards a unifying model. *JOSA-A, 17*(11), 1899–1917.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.

Jollife, I. T. (1986). *Principle component analysis*. New York: Springer-Verlag.

Keysers, C., & Perrett, D. I. (2002). Visual masking and RSVP reveal neural competition. *Trends in Cognitive Sciences, 6*(3), 120–125.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*, 79–86.

Lamme, V. A. F. (2004). Separate neural definitions of visual consciousness and visual attention; a case for phenomenal awareness. *Neural Networks, 17*, 861–872.

Lee, S.-H., & Blake, R. (2001). Neural synergy in visual grouping: when good continuation meets common fate. *Vision Research, 41*, 2057–2064.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *PNAS, 99*(14), 9596–9601.

Li, W., & Gilbert, C. D. (2002). Global contour saliency and local colinear interactions. *Journal of Neurophysiology, 88*, 2846–2856.

Li, Z. (1998). A neural model of contour integration in the primary visual cortex. *Neural Computation, 10*, 903–940.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences, 6*(1), 9–16.

Mack, A., & Rock, I. (1998). *Inattentional blindness*. Cambridge, MA: MIT Press.

Maki, W. S., & Mebane, M. W. (2006). Attentional capture triggers an attentional blink. *Psychonomic Bulletin and Review, 13*(1), 125–131.

Marois, R., Yi, D.-J., & Chun, M. M. (2004). The neural fate of consciously perceived and missed events in the attentional blink. *Neuron, 41*, 465–472.

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

Mounts, J. R. W., & Gavett, B. E. (2004). The role of salience in localized attentional interference. *Vision Research, 44*, 1575–1588.

Mundhenk, T. N., Everist, J., Landauer, C., Itti, L., & Bellman, K. (2005). Distributed biologically based real time tracking in the absence of prior target information. In *SPIE* (Vol. 6006, pp. 330–341): Boston, MA.

Mundhenk, T. N., & Itti, L. (2005). Computational modeling and exploration of contour integration for visual saliency. *Biological Cybernetics, 93*(3), 188–212.

Mundhenk, T. N., Landauer, C., Bellman, K., Arbib, M.A., & Itti, L. (2004). Teaching the computer subjective notions of feature connectedness in a visual scene for real time vision. In: *SPIE* (Vol. 5608, pp. 136–147): Philadelphia, PA.

Navalpakkam, V., & Itti, L. (2006). Optimal cue selection strategy. *Advances in Neural Information Processing Systems (NIPS)*, 987–994.

Navalpakkam, V., & Itti, L. (2007). Search goal tunes visual features optimally. *Neuron, 53*(4), 605–617.

Neisser, U., & Becklen, R. (1975). Selective looking: attending to visually specified event. *Cognitive Psychology, 7*(4), 480–494.

Parkhurst, D. J., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision, 16*(2), 125–154.

Plaut, D., Nowlan, S., & Hinton, G.E. (1986). *Experiments on learning by back propagation. Technical Report CMU-CS-86-126.* Pittsburgh, PA: Department of Computer Science, Carnegie Mellon University.

Potter, M. C., Staub, A., & O'Conner, D. H. (2002). The time course of competition for attention: attention is initially labile. *Journal of Experimental Psychology, 28*(5), 1149–1162.

Raab, D. H. (1963). Backward masking. *Psychological Bulletin, 60*(2), 118–129.

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.

Reeves, A. (1980). Visual imagery in backward masking. *Perception and Psychophysics, 28*, 118–124.

Reeves, A. (1982). Metacontrast U-shaped functions derive from two monotonic functions. *Perception, 11*, 415–426.

Ricciardi, L. M. (1995). Diffusion models of neuron activity. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 299–304). Cambridge, MA: The MIT Press.

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience, 5*(7), 629–630.

Rousselet, G. A., Thorpe, S. J., & Fabre-Thorpe, M. (2004). Processing of one, two or four natural scenes in humans: the limit of parallelism. *Vision Research, 44*, 877–894.

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

Sergent, C., Baillet, S., & Dehaene, S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. *Nature Neuroscience, 8*(10), 1391–1400.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *PNAS, 104*(15), 6424–6429.

Shih, S.-I. (2000). Recall of two visual targets embedded in RSVP streams of distractors depends on their temporal and spatial relationship. *Perception and Psychophysics, 62*(7), 1348–1355.

Shih, S.-I., & Reeves, A. (2007). Attention capture in rapid serial visual presentation. *Spatial Vision, 20*(4), 301–315.

Shih, S.-I., & Sperling, G. (2002). Measuring and modeling the trajectory of visual spatial attention. *Psychological Review, 109*(2), 260–305.

Sperling, G., Reeves, A., Blaser, E., Lu, Z.-L., & Weichselgartner, E. (2001). Two computational models of attention. In J. Braun, C. Koch, & J. L. Davis (Eds.), *Visual attention and cortical circuits* (pp. 177–214). Cambridge, MA: MIT Press.

Tanaka, Y., & Sagi, D. (2000). Attention and short-term memory in contrast detection. *Vision Research, 40*, 1089–1100.

VanRullen, R., & Koch, C. (2003a). Competition and selection during visual processing of natural scenes and objects. *Journal of Vision, 3*, 75–85.

VanRullen, R., & Koch, C. (2003b). Visual selective behavior can be triggered by a feed-forward process. *Journal of Cognitive Neuroscience, 15*(2), 209–217.

VanRullen, R., Reddy, L., & Koch, C. (2004). Visual search and dual tasks reveal two distinct attentional resources. *Journal of Cognitive Neuroscience, 16*(1), 4–14.

Visser, T. A. W., Zuvic, S. M., Bischof, W. F., & Di Lollo, V. (1999). The attentional blink with targets in different spatial locations. *Psychonomic Bulletin and Review, 6*(3), 432–436.

Weisstein, N., & Haber, R. N. (1965). A U-shaped backward masking function in vision. *Psychonomic Science, 2*, 75–76.

Wolfe, J. M., Horowitz, T. S., & Michod, K. O. (2007). Is visual attention required for robust picture memory? *Vision Research, 47*, 955–964.

Yen, S., & Fenkel, L. H. (1998). Extraction of perceptually salient contours by striate cortical networks. *Vision Research, 38*(5), 719–741.