Vision Research 49 (2009) 1295–1306

Contents lists available at ScienceDirect

Vision Research

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

Bayesian surprise attracts human attention

Laurent Itti [a,*], Pierre Baldi [b,1]

[a] Computer Science Department and Neuroscience Graduate Program, University of Southern California, Hedco Neuroscience Building, 3641 Watt Way, HNB-30A, Los Angeles, CA 90089, USA
[b] Computer Science Department and Institute for Genomics and Bioinformatics, University of California, Irvine, Irvine, CA 92697-3425, USA

ARTICLE INFO

Article history:
Received 3 October 2007
Received in revised form 2 September 2008

Keywords:
Attention
Surprise
Bayes theorem
Information theory
Eye movements
Natural vision
Free viewing
Saliency
Novelty

ABSTRACT

We propose a formal Bayesian definition of surprise to capture subjective aspects of sensory information. Surprise measures how data affects an observer, in terms of differences between posterior and prior beliefs about the world. Only data observations which substantially affect the observer's beliefs yield surprise, irrespectively of how rare or informative in Shannon's sense these observations are. We test the framework by quantifying the extent to which humans may orient attention and gaze towards surprising events or items while watching television. To this end, we implement a simple computational model where a low-level, sensory form of surprise is computed by simple simulated early visual neurons. Bayesian surprise is a strong attractor of human attention, with 72% of all gaze shifts directed towards locations more surprising than the average, a figure rising to 84% when focusing the analysis onto regions simultaneously selected by all observers. The proposed theory of surprise is applicable across different spatio-temporal scales, modalities, and levels of abstraction.

1. Introduction and background

In a world full of surprises, animals have developed an exquisite ability to quickly detect and orient towards unexpected events (Ranganath & Rainer, 2003). Yet, at present, our formal understanding of what makes an observation surprising is limited: Indeed, our everyday vocabulary lacks a quantitative notion of surprise, with qualities such as "wow factors" still ill-defined and thus far intractable to quantitative analysis. Here, within the Bayesian probabilistic framework, we develop a simple quantitative theory of surprise. Armed with this theory, we provide direct experimental evidence that Bayesian surprise best characterizes what attracts human gaze in large amounts of natural video stimuli.

Our effort to formally and mathematically define surprise is motivated by the fact that informal correlates of surprise have been described at nearly all stages of neural processing. Thus, surprise is an essential concept for the study of the neural basis of behavior. In sensory neuroscience, for example, it has been suggested that only the unexpected at one stage of processing is transmitted to the next stage (Rao & Ballard, 1999). Hence, sensory cortex may have evolved to adapt to, to predict, and to quiet down the expected statistical regularities of the world (Olshausen & Field, 1996; Müller, Metha, Krauskopf, & Lennie, 1999; Dragoi, Sharma, Miller, & Sur, 2002; David, Vinje, & Gallant, 2004), focusing instead on events that are unpredictable or surprising (Fairhall, Lewen, Bialek, & de Ruyter Van Steveninck, 2001). Electrophysiological evidence for this early sensory emphasis onto surprising stimuli exists from studies of adaptation in visual (Maffei, Fiorentini, & Bisti, 1973; Movshon & Lennie, 1979; Müller et al., 1999; Fecteau & Munoz, 2003), olfactory (Kurahashi & Menini, 1997; Bradley, Bonigk, Yau, & Frings, 2004), and auditory cortices (Ulanovsky, Las, & Nelken, 2003), subcortical structures like the LGN (Solomon, Peirce, Dhruv, & Lennie, 2004), and even retinal ganglion cells (Smirnakis, Berry, Warland, Bialek, & Meister, 1997; Brown & Masland, 2001) and cochlear hair cells (Kennedy, Evans, Crawford, & Fettiplace, 2003): neural responses greatly attenuate with repeated or prolonged exposure to an initially novel stimulus. At higher levels of abstraction, surprise and novelty are also central to learning and memory formation (Ranganath & Rainer, 2003), to the point that surprise is believed to be a necessary trigger for associative learning (Schultz & Dickinson, 2000; Fletcher et al., 2001), as supported by mounting evidence for a role of the hippocampus as a novelty detector (Knight, 1996; Stern et al., 1996; Li, Cullen, Anwyl, & Rowan, 2003). Finally, seeking novelty is a well-identified human character trait, possibly associated with the dopamine D4 receptor gene (Ebstein et al., 1996; Benjamin et al., 1996; Lusher, Chandler, & Ball, 2001).

Empirical and often *ad-hoc* formalizations of surprise, usually referred to as spatial "saliency" or temporal "novelty," are at the core of many laboratory studies of attention and visual search: The strongest attractors of attention are stimuli that pop-out from

* Corresponding author. Fax: +1 213 740 5687.
E-mail addresses: itti@usc.edu (L. Itti), pfbaldi@ics.uci.edu (P. Baldi).
[1] Tel.: +1 949 824 5809; fax: +1 949 824 4056.

0042-6989/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2008.09.007

their neighbors in space or time, like a salient vertical bar embedded within an array of horizontal bars (Treisman & Gelade, 1980; Wolfe & Horowitz, 2004), or the abrupt onset of a novel bright dot in an otherwise empty display (Theeuwes, 1995). Computationally, these notions may be summarized in terms of outliers (Markou & Singh, 2003) and Shannon information: stimuli which have low likelihood given a distribution of expected or learned stimuli, over space or over time, are outliers, are more informative in Shannon's sense, and capture attention (Duncan & Humphreys, 1989). We show that this line of thinking at best captures an approximation to surprise, but can be flawed in some extreme cases. To exacerbate the differences and to gauge their practical impact in ecologically relevant situations, we quantitatively compare Bayesian surprise to 10 existing measures of saliency and novelty, in their ability to predict human gaze recordings on large amounts of natural video data. We find that Bayesian surprise best characterizes where people look, even more so for stimuli that are consistently fixated by multiple observers. Our results suggest that surprise is an important formalization for understanding neural processing and behavior, and is the best known attractor of human attention.

This work extends Itti and Baldi (2006), through a more complete exposition of the theory and of the new proposed unit of surprise (the "wow"), simple examples of how surprise may be computed, and a broader set of experiments and comparisons with competing theories and models.

2. Theory

In this paper, we elaborate a definition of surprise that is general, information-theoretic, derived from first principles, and formalized analytically across spatio-temporal scales, sensory modalities, and, more generally, data types and data sources. Two elements are essential for a principled definition of surprise. First, surprise can exist only in the presence of uncertainty. Uncertainty can arise from intrinsic stochasticity, missing information, or limited computing resources. A world that is purely deterministic and predictable in real-time for a given observer contains no surprises. Second, surprise can only be defined in a relative, subjective, manner and is related to the expectations of the observer, be it a single synapse, neuronal circuit, organism, or computer device. The same data may carry different amounts of surprise for different observers, or even for the same observer taken at different times.

2.1. Defining surprise

In probability and decision theory it can be shown that, under a small set of axioms, the only consistent way for modeling and reasoning about uncertainty is provided by the Bayesian theory of probability (Cox, 1964; Savage, 1972; Jaynes, 2003). Furthermore, in the Bayesian framework, probabilities correspond to subjective degrees of beliefs in hypotheses (or so-called models). These beliefs are updated, as data is acquired, using Bayes' theorem as the fundamental tool for transforming prior belief distributions into posterior belief distributions. Therefore, within the same optimal framework, a consistent definition of surprise must involve: (1) probabilistic concepts to cope with uncertainty and (2) prior and posterior distributions to capture subjective expectations. These two simple components are at the basis of the proposed definition of surprise below.

The background information of an observer is captured by his/her/its prior probability distribution $\{P(M)\}_{M\in\mathscr{M}}$ over the hypotheses or models M in a model space $\mathscr{M}$. At a high level of abstraction and for, e.g., a human observer, the ensemble $\mathscr{M}$ may for instance consist of a number of cognitive hypotheses or models of the world, such as:

$$\mathscr{M} = \{\text{it will rain tomorrow; the cold war is over; the USC-Trojans football team is on a winning streak; my wallet is in my possession; my car is in good working order; my credit card information is secure; nobody at work knows that today is my birthday; etc}\} \quad (1)$$

At lower levels of abstraction and for less sophisticated observers, the model space may be much simpler, corresponding to straightforward hypotheses over well-defined quantities, such as, for example, the amount of light hitting a given photoreceptor:

$$\mathscr{M} = \{\text{light level is low; light level is medium; light level is high; etc}\} \quad (2)$$

With each of these hypotheses or models M is associated a likelihood function, $P(D|M)$, which quantifies how likely any data observation D is under the assumption that a particular model M is correct.

Given the prior distribution of beliefs before the next observation of data, the fundamental effect of a new data observation D on the observer is to change the prior distribution $\{P(M)\}_{M\in\mathscr{M}}$ into the posterior distribution $\{P(M|D)\}_{M\in\mathscr{M}}$ via Bayes' theorem, whereby

$$\forall M \in \mathscr{M}, \qquad P(M|D) = \frac{P(D|M)}{P(D)}P(M). \quad (3)$$

In this framework, the new data observation D carries no surprise if it leaves the observer's beliefs unaffected, that is, if the posterior distribution over the ensemble $\mathscr{M}$ is identical to the prior. Conversely, D is surprising if the posterior distribution after observing D significantly differs from the prior distribution. Therefore we formally measure surprise by quantifying the distance (or dissimilarity) between the posterior and prior distributions. Computing such distance between two probability distributions is best done using the relative entropy or Kullback-Leibler (KL) divergence (Kullback, 1959). Thus, surprise is defined by the average of the log-odd ratio:

$$S(D,\mathscr{M}) = KL(P(M|D),P(M)) = \int_{\mathscr{M}} P(M|D)\log\frac{P(M|D)}{P(M)}\,\mathrm{d}M \quad (4)$$

taken with respect to the posterior distribution over the model space $\mathscr{M}$. For example, using the premises of Eq. (1), if the data observation D consisted of patting your pocket and realizing that it feels unusually empty, that would create surprise as your posterior beliefs in the hypotheses "my wallet is in my possession" and "my credit card information is secure" would be dramatically lower than the prior beliefs in these hypotheses, resulting in a large KL distance between posterior and prior over all hypotheses, and in large surprise.

Note that KL is not symmetric but has well-known theoretical advantages, including invariance with respect to reparameterizations. A unit of surprise – a "*wow*" – may then be defined for a single model M as the amount of surprise corresponding to a two-fold variation between $P(M|D)$ and $P(M)$, i.e., as $\log P(M|D)/P(M)$ (with log taken in base 2). The total number of wows experienced when simultaneously considering all models is obtained through the integration in Eq. (4). In the following section, we provide a simple description of how surprise may be computed, and of how it fundamentally differs from Shannon's notion of information (notably, Shannon's entropy requires integration over the space $\mathscr{D}$ of all pos-

sible data observations, while surprise requires integration over the space $\mathcal{M}$ of all models of the observer). Surprise can always be computed numerically, but also analytically in many practical cases, in particular those involving probability distributions in the exponential family (Brown, 1986) with conjugate or other priors (see below).

The Kullback-Leibler divergence (*KL*) has been used extensively, at least since Shannon with the mutual information between two random variables *X* and *Y* defined as $KL(P(X,Y), P(X)P(Y))$. In particular, there is a rich history of using *KL* in machine learning, Boltzmann machines, and neural networks, especially in the context of computing the gradient of the *KL*, and using gradient descent on the *KL* for learning (Ackley, Hinton, & Sejnowski, 1985). It is important to note that here, however, we use it in a different way. In neural networks, for instance, training is often done to maximize the likelihood $P(D|M) = P(D|w)$, or, when there is a prior on the weight vector w, to maximize the posterior $P(w|D)$ (Note that the data vector D may include target values in the case of supervised learning). The *KL* often then appears in the expression of the error function, usually the negative log likelihood. In a typical multinomial classification problem, learning is done by gradient descent on the negative log likelihood associated with the *KL* between the data distribution $P(D)$ and the distribution produced by the network $P(D|w)$. That is, one tries to minimize the mismatch between $P(D)$ and $P(D|w)$ by adjusting w. Clearly this is different from the *KL* between the posterior $P(w|D)$ and the prior $P(w)$, which is surprise: surprise requires integration over the model space (or weights w) while previous methods integrate over the space of data D.

2.2. Surprise, shannon information, and the white snow paradox

To illustrate how surprise arises when data is observed, consider a human observer who just turned a television set on, not knowing which channel it is tuned to. The observer has a number of co-existing hypotheses or models about which channel may be on, for example, MTV, CNN, FOX, BBC, etc. (Fig. 1). Over the course of viewing the first few video frames of the unknown channel (here, CNN), the observer's beliefs in each hypothesis adjust, progressively favoring one channel over the others (leading to a higher prior probability for CNN in left panel). Consider next what happens if yet another video frame of the same program is observed (Fig. 1, top right), intuitively an unsurprising event. Through Bayesian update, the new frame only minimally alters the observer's beliefs, with the posterior distribution of beliefs over models showing a slightly reinforced belief into the correct channel at the expense of the others. In contrast, if a frame of snow was suddenly observed (Fig. 1, middle right), intuitively this should be a very surprising event, as it may signal storm, earthquake, toddler's curiosity, electronic malfunction, or a military putsch. Through Bayesian update, this observation would yield a large shift between the prior and posterior distributions of beliefs, with the posterior now strongly favoring a snow model (and possible associated earthquake, malfunction, etc. hypotheses), correspondingly reducing belief in all other television channels. In sum, unsurprising data yields little difference between posterior and prior distributions of beliefs over models, while surprising data yields a large shift: in mathematical terms, an event is surprising when the distance between posterior and prior distributions of beliefs over all models is large (see Eq. (4)).

While at onset snow is surprising (Fig. 1, middle right), after sustained viewing it quickly becomes boring to most humans. Indeed, no more surprise arises after the observer's beliefs have stabilized towards strongly favoring the snow model over all others (Fig. 1, bottom right). Thus surprise resolves the classical paradox that random snow, although in the long term the most boring of all television programs, carries the largest amount of Shannon information. This paradox arises from the fact that there are many more possible random images than there exists natural images. Thus, the entropy of snow is higher than that of natural scenes (Field, 2005). Even when the observer knows to expect snow, every individual frame of snow carries a large amount of Shannon information. Indeed, in a sample recording of 20,000 video frames from typical television programs, presumably of interest to millions of watchers, we measured approximately 20 times less Shannon information per second than in matched random snow clips, after compression to constant-quality MPEG4 to adaptively eliminate redundancy in both cases (Table 1). The situation was reversed when we measured that snow clips carried about 17 times less surprise per second than the television clips, evaluated using the average, over space and time, of the output of the surprise metric presented with our human experiments. Note that a clip where all frames are black would practically carry no Shannon information and yield no surprise. Thus, more informative data may not always be more important, interesting, worthy of attention, or surprising; in fact, the most interesting or surprising data may often carry intermediate amounts of Shannon information, between fully predictable data (black frames) and completely unpredictable data (snow frames).

2.3. Simple example of surprise computation

One class of examples where surprise can be computed exactly consists of contingency tables of any size. Consider for instance a parent who has two competing internal models or hypotheses about a new television channel, the first, M, according to which that new channel is appropriate for children, and the second, $\overline{M}$, according to which it is not. Assume that initially our observer is undecided and equally split across both models, that is, $P(M) = P(\overline{M}) = 1/2$. Next consider two possible data observations, D_1, a TV program that contains some nudity, and D_2, one that does not, with, for instance, $P(D_1) = (D_2) = 1/2$. Finally, assume that the observer initially believes that observing nudity is three times more likely on a channel that is inappropriate for children.

The initial beliefs of our observer may thus be tabulated as follows:

	D_1	D_2
M	$a = 1/8$	$c = 3/8$
$\overline{M}$	$b = 3/8$	$d = 1/8$

where the table verifies the above specifications, in that $P(D_1) = a + b = 1/2$, $P(D_2) = c + d = 1/2$, $P(M) = a + c = 1/2$, $P(\overline{M}) = b + d = 1/2$, and $P(D_1, \overline{M}) = b = 3 \times P(D_1, M) = 3 \times a$. Assume that D_1 is observed (a program with some nudity). Since $P(D_1) = 1/2$, this observation carries $-\log P(D_1) = 1$ bit of Shannon information (remember that the logarithm should be taken in base 2 for all numerical applications). The posterior probabilities of M and $\overline{M}$ are

$$P(M|D_1) = \frac{P(M, D_1)}{P(M, D_1) + P(\overline{M}, D_1)} = \frac{a}{a+b} = \frac{1}{4} \quad \text{and} \tag{5}$$

$$P(\overline{M}|D_1) = \frac{P(\overline{M}, D_1)}{P(M, D_1) + P(\overline{M}, D_1)} = \frac{b}{a+b} = \frac{3}{4} \tag{6}$$

That is, observing D_1 (a program with some nudity) shifted the observer's initial indecision between M and $\overline{M}$, now favoring $\overline{M}$ (the new TV channel is inappropriate for children) over M (it is appropriate) by a factor 3. The amount of surprise resulting from this shift,

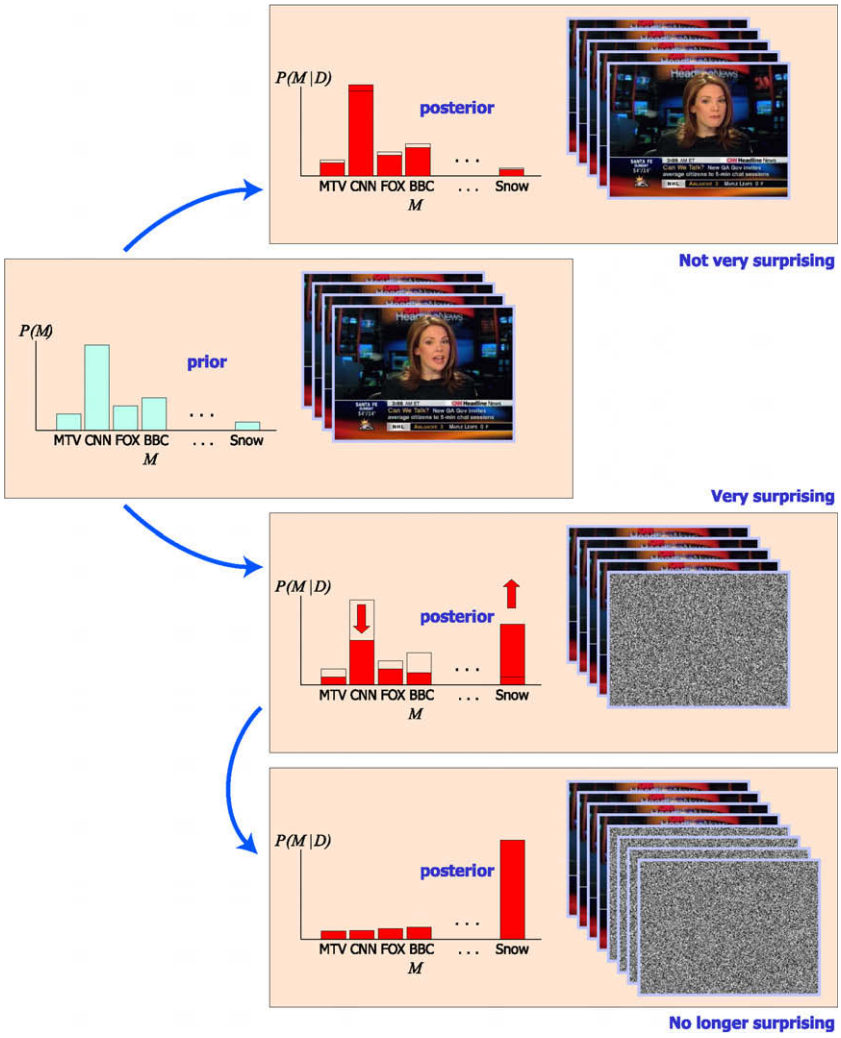

Fig. 1. Simple description of how surprise may be computed at a high level of abstraction, for an observer who has beliefs about possible television channels that she or he may be watching. Section 3 for further details.

Table 1
Could you guess how interesting a video clip is before you watch it, by just looking at how many megabytes of Shannon information it contains? Here we find that a clip of snow would be about 20 times a larger file than a clip of typical television. Surprise resolves this so-called "white snow paradox".

	TV	Snow	TV: Snow ratio
Shannon information (Mbyte/s)	0.25 ± 0.16	4.90 ± 0.01	1:20
Surprise (wows/s)	50.83 ± 0.43	2.99 ± 0.02	17:1

first considering only model M, is $S(D_1, M) = \log \frac{P(M|D_1)}{P(M)} = -1.00$ wow. Similarly, with respect to $\overline{M}$, the surprise is $S(D_1, \overline{M}) = \log \frac{P(\overline{M}|D_1)}{P(\overline{M})} = 0.58$ wows. After averaging over the model family $\mathscr{M} = \{M, \overline{M}\}$ weighted by the posterior (Eq. 2), the total surprise experienced by the observer is

$$S(D_1, \mathscr{M}) = P(M|D_1)S(D_1, M) + P(\overline{M}|D_1)S(D_1, \overline{M}) \quad (7)$$

$$= \frac{a}{a+b} \log \frac{a}{(a+b)(a+c)} + \frac{b}{a+b} \log \frac{b}{(a+b)(b+d)} \quad (8)$$

$$\approx 0.19 \text{wows}. \quad (9)$$

The new beliefs of the observer may hence be tabulated as follows, using the posterior resulting from our above observation as new prior:

	D_1	D_2
M	$a' = 1/16$	$c' = 3/16$
$\overline{M}$	$b' = 7/16$	$d' = 5/16$

Consider next what happens if D_1 is observed once more. We intuitively expect this second observation to carry less surprise than the previous one, since our observer now already fairly strongly believes that the new TV channel is inappropriate, and observing nudity once again should only incrementally consolidate that belief. Indeed, proceeding as above, the total surprise now experienced by the observer is $S(D_1, \mathscr{M}) = 0.07$ wows, nearly three times less than on the previous observation.

2.4. Analytical computations of surprise with N data points

Exponential Family. Consider a family of models $\mathscr{M}$ parameterized by w with likelihood $P(D|M) = P(D|w)$. By definition, the conjugate prior $P(M) = P(w)$ has the same functional form as the likelihood. In this case, by Bayes' theorem, the posterior also has the same functional form. While surprise can be computed with any prior, conjugate priors are useful for their mathematical simplicity and ease of implementation during Bayesian learning,

where the posterior at one iteration becomes the prior of the following iteration.

A likelihood is in the exponential family with parameter vector w if it can be expressed in the form, for a single datum d

$$P(d|w) = h(d)c(w)\exp\left(\sum_{i=1}^{k}\theta_i(w)t_i(d)\right) \quad (10)$$

With N independent data points ($D = d_1, \ldots, d_N$),

$$P(D|w) = [c(w)]^N\left[\prod_{j=1}^{N}h(d_j)\right]\exp\left(\sum_{i=1}^{k}\theta_i(w)T_i(D)\right) \quad (11)$$

letting $T_i(D) = \sum_{j=1}^{N} t_i(d_j)$ be the sufficient statistics. Most common distributions belong to the exponential family. The conjugate prior has a similar exponential form

$$P(w;\alpha_i) = C\exp\left(\sum_{i=1}^{k}\alpha_i\theta_i(w)\right) \quad (12)$$

parameterized by the α_i's. Using Bayes' theorem, the posterior has the same exponential form with normalizing constant C' and $\alpha_i' = \alpha_i + T_i(D)$. Calculation of surprise yields

$$S(D,\mathscr{M}) = \log\frac{C'}{C} - \sum_{i=1}^{k}T_i(D)E[\theta_i(w)] \quad (13)$$

where $E[\theta_i(w)]$ is the expectation of $\theta_i(w)$ with respect to the posterior. Surprise can be rewritten as:

$$S(D,\mathscr{M}) = N\left(\log c(w) + <\log h(d)> - <\log P(d)> - \sum_{i=1}^{k}<t_i(d)>E[\theta_i(w)]\right) \quad (14)$$

where $<>$ denotes averages over the data points. Thus in general, for large N, surprise grows linearly with the number of data points. Below is an application.

Binary Data Modeled as a Series of Independent and Identical Coin Tosses (Binomial Model). The family $\mathscr{M}$ of models is parameterized by the probability $0 \leqslant w \leqslant 1$ of observing "heads" on a coin toss, thus encompasses models of biased coins (small and large w values) and of fair coins ($w = 0.5$). The conjugate prior is the Beta prior $P(w;\alpha,\beta) = Cw^{\alpha-1}(1-w)^{\beta-1}$ with $C = \Gamma(\alpha+\beta)/[\Gamma(\alpha)\Gamma(\beta)]$ and parameters α, β. With a number n of heads observed after tossing a coin N times, the posterior is also a Beta distribution with $\alpha' = \alpha + n$ and $\beta' = \beta + (N - n)$. Integrating over models, surprise is

$$S(D,\mathscr{M}) = \log\frac{C'}{C} - n[\Psi(\alpha+\beta+N) - \Psi(\alpha+n)] - (N-n)[\Psi(\alpha+\beta+N) - \Psi(\beta+N-n)] \quad (15)$$

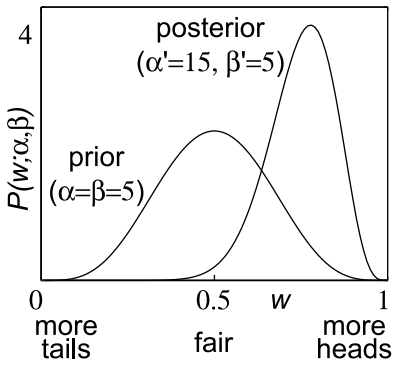

Fig. 2. Simple example of surprise computation for series of coin tosses. Here the prior and posterior distributions of beliefs about how fair the coin may be are formalized as Beta distributions.

where Ψ is the digamma function. For example, assume an observer who initially believes most coins are fair, i.e., whose prior is concentrated around $w = 0.5$ (e.g., $\alpha = \beta = 5$; Fig. 2). Assume that $N = 10$ tosses of a coin are observed and happen to yield exactly $n = 10$ heads. This observation is surprising and shifts the observer's beliefs towards favoring the models of coins that yield more heads ($\alpha' = 15$, $\beta' = 5$; Fig. 2), resulting in 2.26 wows of surprise. An outcome of five heads and five tails would elicit only 0.15 wows from slight sharpening of the prior around $w = 0.5$ ($\alpha' = 10$, $\beta' = 10$).

3. Methods

To test the surprise hypothesis – that Bayesian surprise attracts human attention in dynamic natural scenes – we recorded eye movements from eight naïve observers. Each watched a subset (about half) from 50 videoclips totaling over 25 min of playtime. Clips comprised outdoors daytime and nighttime scenes of crowded environments, video games, and television broadcast including news, sports, and commercials.

To characterize image regions selected by participants, we process videoclips through computational metrics that output a topographic dynamic master response map, assigning in real-time a response value to every input location. A good master map would highlight, more than expected by chance, locations gazed to by observers. To score each metric we hence sample, at onset of every human saccade, master map activity around the saccade's future endpoint, and around a uniformly random endpoint (random sampling was repeated 100 times to evaluate variability). We quantify differences between histograms of master map samples collected from human and random saccades using again the Kullback–Leibler (KL) distance: metrics which better predict human scanpaths exhibit higher distances from random. This scoring presents several advantages over simpler scoring schemes (Reinagel & Zador, 1999, Parkhurst, Law, & Niebur, 2002), including agnosticity to putative mechanisms for generating saccades and the fact that applying any continuous nonlinearity to master map values would not affect scoring.

3.1. Subjects and stimuli

Subjects were USC students and staff, three females and five males, ages 23–32, normal or corrected-to-normal vision. Informed consent was obtained from all subjects prior to the experiments. Each subject watched a subset of the collection of videoclips, so that eye movement traces from four distinct subjects were obtained for each clip. Videoclips were presented on a 22" CRT monitor (LaCie, Inc.; 640×480, 60.27 Hz double-scan, mean screen luminance 30 cd/m^2, room 4 cd/m^2, viewing distance 80 cm, field of view $28^\circ \times 21^\circ$). The clips comprised between 164 to 2814 frames or 5.5–93.9 s, totaling 46,489 frames or 25:42.7 playback time. Frames were presented on a Linux computer under SCHED_FIFO scheduling which ensured microsecond-accurate timing (Finney, 2001).

Right-eye position was tracked at 240 Hz using a video-based device (ISCAN RK-464), which robustly estimates gaze from comparative real-time measurements of both the center of the pupil and the reflection of an infrared light source onto the cornea. Saccades were defined by a velocity threshold of 20°/s and amplitude threshold of 2°.

Observers were instructed to follow the stimuli's main actors and actions, so that their gaze shifts reflected an active search for nonspecific information of subjective interest. Two hundred calibrated eye movement traces (10,192 saccades) were analyzed, corresponding to four distinct observers for each of the 50 clips. Fig. 4a shows sample scanpaths for one videoclip.

Sampling of master map values around human or random saccade targets used a circular aperture of diameter 5.6°, approximating the size of the fovea and parafovea. Saccade initiation latency was accounted for by subjecting the master maps to a temporal low-pass filter with time constant $\tau = 500$ ms. This provided an upper bound, allowing the analysis to compensate for delays between the physical appearance of stimuli on the screen and the start of human saccades. The random sampling process was repeated 100 times, yielding the (very small) error bars of the random histograms of Fig. 5.

Note that instead of using a uniform random sampling, the random saccade distribution could also have been derived by randomly shuffling the human saccades (Tatler, Baddeley, & Gilchrist, 2005). However, it is important to understand that using such biased random distribution would confer particular knowledge to the random sampling process: general knowledge aggregated over the human dataset under study would be exploited to decide how random samples are to be selected. Our computational gaze prediction metrics do not have any such knowledge about the human dataset, and do not have any built-in spatial bias, central or otherwise (except, for some metrics, a slight bias against the extreme image borders due to boundary conditions on the filters applied to the images). Hence, we here chose to retain knowledge-free computational and random metrics, rather than to contaminate them with knowledge about the dataset under study.

3.2. Simulations

Dynamic master maps generated by the computational gaze prediction metrics were 40×30 lattices of metric responses computed over 16×16 image patches, given 640×480 stimuli. The master maps were internally updated at a rate of 10,000 frames/s (simulated time, actual CPU time was much longer), receiving new input video frames every 33.185 ms. Simulations were parallelized across Beowulf clusters of computers, totaling in excess of one CPU-year to evaluate all computational metrics against our 41.5 GB of raw video data and 1,542,752 eye movement samples.

3.3. Human-derived metric

With our clips and instructions, observers agreed with each other quite often on where to look (Fig. 4b). Hence our data presents ample opportunity for characterizing in the image space what most strongly attracted human observers. An upper-bound *KL* score can be computed from a human-derived metric, whose master map is built, at every video frame, from eye movement positions of the three observers other than that under test (Fig. 4c). Computational metrics are expected to yield *KL* scores between zero (chance level) and this upper bound of $KL = 0.679 \pm 0.011$ reflecting inter-observer consistency. To build the human-derived maps, a Gaussian blob with $\sigma = 3$ master map pixels (4.5°) was continuously painted at each of the eye positions of the three observers other than that under test, with some forgetting provided by the master map's temporal low-pass filter. High metric responses were hence sampled if and only if a saccade of the observer under test was aimed to approximately a location where other observer(s) were currently looking. Because this metric is not predictive like the others, sampling occurred when a saccade ended (and other humans were expected to also be reaching the endpoint) rather than when it started (and other humans possibly also started).

3.4. Static metrics

The simplest computational metrics tested only exploit local and static image properties. The variance metric computes local variance of pixel luminance within 16×16 image patches (Reinagel & Zador, 1999). The Shannon entropy metric computes the entropy of the local histogram of grey-levels in 16×16 image patches (Privitera & Stark, 2000). The DCT-based (Discrete Cosine Transform) information metric similarly computes in image patches the number of DCT coefficients above detection threshold, for the luminance and two chrominance channels (Itti, Koch, & Niebur, 1998). The color, intensity and orientation contrast metrics are derived from reduced versions of our previously proposed bottom-up saliency metric (Itti & Koch, 2001). They compute local contrast in each feature dimension using difference-of-Gaussian center-surround contrast detectors operating at six different spatial scales.

3.5. Dynamic and saliency metrics

The flicker and motion metrics rely on the same center-surround architecture as for color, intensity, and orientation. The saliency metric combines intensity contrast (six feature maps), red/green and blue/yellow color opponencies (12 maps), four orientation contrasts (24 maps), flicker (six maps) and motion energy in four directions (24 maps), totaling 72 feature maps. Central to the saliency metric and each of its center-surround feature channels is neurally-inspired non-classical spatial competition for saliency (Sillito, Grieve, Jones, Cudeiro, & Davis, 1995; Itti & Koch, 2001), by which distant active locations in each feature map inhibit each other, giving rise to pop-out and attentional capture (Wolfe & Horowitz, 2004). Thus, these metrics are not necessarily attracted to locally information-rich image regions, as many highly informative regions will be discarded if they resemble their neighbors. For this reason, these metrics typically yielded sparser maps than the contrast, entropy, and DCT-based information metrics, which are purely local. These metrics represent biologically-plausible heuristics to an outlier detection metric described below.

3.6. Surprise and outlier metrics

The surprise metric retains the 72 raw feature detection mechanisms of the saliency metric (but without the non-classical competition for saliency), and attaches local surprise detectors to each location in each of the 72 feature maps. Surprise detectors compute both local temporal surprise (generalizing outliers-based temporal novelty) and spatial surprise (generalizing outliers-based spatial saliency).

In our implementation, image patches are described by a 72D feature vector representing the responses from the 72 low-level feature channels (color, motion, etc. at six spatial scales). A model of an image patch, then, is a 72D vector of 1D Poisson random variables, under the assumption that each low-level feature detector outputs 1D trains of Poisson-distributed spikes in response to visual stimulation (Softky & Koch, 1993). We consider the model family that comprises all possible such 72D vectors of Poisson models, parameterized by a single 72D Poisson rate vector. For instance, patches of motionless vs. trembling foliage correspond to two different models, described by two vectors of 72 Poisson firing rates (with, among other differences, lower rates for motion features in the motionless foliage model). Note that with these simple models, there is no single explicit model M_{snow} that can capture random snow. Rather, when snow is observed, the prior quickly becomes uniform, indicating that every model is believed to be equally bad and that the observer does not have any strong belief in favor of any one model. More complex models (Doretto, Chiuso, Wu, & Soatto, 2003) could be used as well, without affecting the theory.

Consider a neuron at a given location in one of the 72 feature maps, receiving Poisson spikes as inputs from low-level feature detectors. We compute surprise independently in that

feature map and at that location using a family of models which are all the 1D Poisson distributions for all possible firing rates $\lambda > 0$.

Using the theory of surprise outlined above, we consider conjugate priors, whereby the posterior belongs to the same functional family as the prior. In such case, the posterior at one video frame can directly serve as prior for the next frame, as is customary in Bayesian learning. Thus, we use for $P(M)$ a functional form such that $P(M|D)$ has the same functional form when D is Poisson-distributed. It is easy to show that $P(M)$ satisfying this property is the Gamma probability density:

$$P(M(\lambda)) = \gamma(\lambda;\alpha,\beta) = \frac{\beta^{\alpha}\lambda^{\alpha-1}e^{-\beta\lambda}}{\Gamma(\alpha)} \tag{16}$$

with shape $\alpha > 0$, inverse scale $\beta > 0$, and $\Gamma(.)$ the Euler Gamma function. Given an observation $D = \bar{\lambda}$ at one of our surprise detectors and prior density $\gamma(\lambda;\alpha,\beta)$, the posterior $\gamma(\lambda;\alpha',\beta')$ obtained by Bayes' theorem is also a Gamma density, with:

$$\alpha' = \alpha + \bar{\lambda} \quad \text{and} \quad \beta' = \beta + 1 \tag{17}$$

To prevent these from increasing unbounded over time, we add a forgetting factor $0 < \zeta < 1$, yielding:

$$\alpha' = \zeta\alpha + \bar{\lambda} \quad \text{and} \quad \beta' = \zeta\beta + 1 \tag{18}$$

ζ preserves the prior's mean α/β but increases its variance α/β^2, embodying relaxation of belief in the prior's precision; our simulations use $\zeta = 0.7$, based on a reproduction of neural recordings from Müller et al. (1999). Local temporal surprise S_T resulting from the update is computed exactly using the *KL* divergence to quantify the differences between posterior and prior distributions over models:

$$\begin{aligned} S_T(D,\mathscr{M}) &= KL(\gamma(\lambda;\alpha',\beta'),\gamma(\lambda;\alpha,\beta)) \\ &= -\alpha + \alpha\log\frac{\beta'}{\beta} + \log\frac{\Gamma(\alpha)}{\Gamma(\alpha')} + \beta\frac{\alpha'}{\beta'} + (\alpha'-\alpha)\Psi(\alpha') \end{aligned} \tag{19}$$

with $\Psi(.)$ the digamma function. Spatial surprise S_S is computed similarly. At every visual location, a Gamma neighborhood prior is computed as the weighted combination of priors from local models, over a large neighborhood with 2D Difference-of-Gaussians profile ($\sigma_+ = 20$ and $\sigma_- = 3$ feature map pixels, i.e., 29° and 4.5° resp.). As new data arrives, spatial surprise is the *KL* between the posterior neighborhood distribution after update by local samples from the neighborhood's center, and the prior. Temporal and spatial surprise

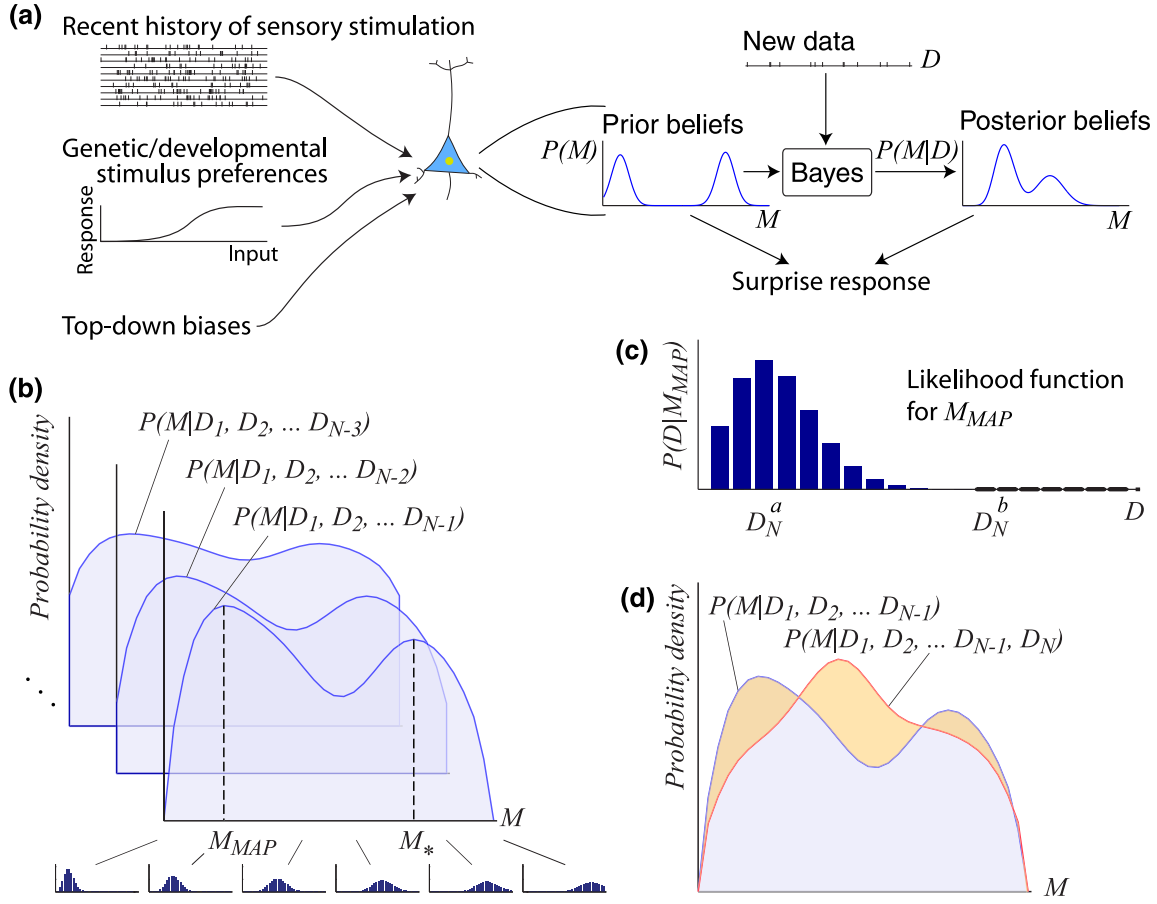

Fig. 3. Hypothetical implementation of surprise computation in a single neuron. (a) Prior data observations, tuning preferences, and top-down influences contribute to shaping a set of "prior beliefs" a neuron may have over a class of internal models or hypotheses about the world. For instance, $\mathscr{M}$ may be a set of Poisson processes parameterized by the rate λ, with $\{P(M)\}_{M\in\mathscr{M}} = \{P(\lambda)\}_{\lambda\in\mathbb{R}^{+*}}$ the prior distribution of beliefs about which Poisson models well describe the world as sensed by the neuron. New data D updates the prior into the posterior using Bayes' theorem. Surprise quantifies the difference between the posterior and prior distributions over the model class $\mathscr{M}$. The remaining panels detail how surprise differs from conventional model fitting and outlier-based novelty. (b) In standard iterative Bayesian model fitting, at every iteration N, incoming data D_N is used to update the prior $\{P(M|D_1,D_2,\ldots,D_{N-1})\}_{M\in\mathscr{M}}$ into the posterior $\{P(M|D_1,D_2,\ldots,D_N)\}_{M\in\mathscr{M}}$. Freezing this learning at a given iteration, one then picks the currently best model, usually using either a maximum likelihood criterion, or a maximum a posteriori one (yielding M_{MAP} shown). (c) This best model is used for a number of tasks at the current iteration, including outlier-based novelty detection. New data is then considered novel at that instant if it has low likelihood for the best model (e.g., D_N^b is more novel than D_N^a). This focus onto the single best model presents obvious limitations, especially in situations where other models are nearly as good (e.g., M_* in panel (b) is entirely ignored during standard novelty computation). One palliative solution is to consider mixture models, but this just amounts to shifting the problem into a different model class. (d) Surprise directly addresses this problem by simultaneously considering all models and by measuring how data changes the observer's distribution of beliefs from $\{P(M|D_1,D_2,\ldots,D_{N-1})\}_{M\in\mathscr{M}}$ to $\{P(M|D_1,D_2,\ldots,D_N)\}_{M\in\mathscr{M}}$ over the entire model class $\mathscr{M}$ (orange shaded area).

are combined additively to yield the final surprise metric. Additional implementation details have been described previously (Itti & Baldi, 2005, 2006).

The outlier detection metric uses exactly the same Poisson models and low-level visual features as the surprise metric, but fundamentally differs from surprise in that it focuses onto the single best model at a given moment, instead of simultaneously considering all models like the surprise metric. Thus, at every location in every feature map, the best Poisson model $M(\lambda_{\text{best}})$ given the observed data to date is considered, and is used to compute the likelihood of the new data sample, yielding:

$$O(D, M(\lambda_{\text{best}})) = \frac{1}{P(D|M(\lambda_{\text{best}}))} - 1. \quad (20)$$

Thus, a data observation D which is an outlier, that is, has low likelihood $P(D|M(\lambda_{\text{best}})) \approx 0$ given the currently best model, yields a large response, while an inlier data observation with high $P(D|M(\lambda_{\text{best}})$ yields a lower response.

Fig. 3 illustrates how surprise differs from the notions of saliency and novelty based on outliers and Shannon information, by examining a hypothetical implementation of surprise computation in a single neuron.

4. Results

We compare the ten computational metrics described above, which encompass and extend the state-of-the-art found in previous studies, to Bayesian surprise (Table 2). The first six metrics quantify static image properties while the remaining four, and Bayesian surprise, also respond to dynamic events. The first three metrics compute local variance, Shannon entropy, and DCT-based (discrete cosine transform) information within 16×16 image patches, as previously proposed to characterize attractors of human gaze over static images (Reinagel & Zador, 1999; Privitera & Stark, 2000; Itti et al., 1998). We find that humans are significantly attracted by image regions with higher metric responses (Fig. 5, Table 2). However, these purely local metrics typically respond vigorously at numerous visual locations. Hence they are poorly specific and yield relatively low *KL* scores between humans and random: while humans preferentially gaze towards locations with high metric responses, such locations are often so numerous that high metric responses are also collected at random saccade targets.

The next three metrics increase specificity by focusing on spatial image outliers in the dimensions of color, intensity, and orientation, using heuristic biologically-inspired center-surround detectors operating at six spatial scales. These metrics yield sparser maps, as local responses are inhibited unless they contrast with neighboring regions. While we find that humans also significantly gaze towards regions with outlier color, intensity, and orientation, these metrics do not score substantially better than the previous three.

The next two metrics consider the dimensions of flicker (onsets/offsets of light intensity) and directional motion energy, again employing six center-surround scales, and hence measure spatiotemporal novelty. Both score equivalently well and significantly higher than the static metrics (nearly 20% better than the best static metric, entropy), providing a quantitative evaluation of the stronger impact of dynamic features onto human attentional selection over natural video scenes.

The last three metrics – biologically-inspired saliency, outlier detection, and surprise – employ a common front-end which consists of a set of linear filters tuned to the five features of color, intensity, orientation, flicker, and motion at six spatial scales (Itti & Koch, 2001). They differ in the computations applied to the outputs of these linear detectors to yield a master map. The biologically-inspired saliency metric implements a heuristic detection of spatial and temporal outliers in each of the low-level feature channels and spatial scales: non-linear competitive interactions between distant visual locations, mimicking the non-classical surround suppression effects observed in primary visual cortex (Sillito et al., 1995; Itti et al., 1998), enhance isolated or outlier stimuli while suppressing more extended regions with high linear filter outputs. Because it considers both static and dynamic features, our saliency metric combines both notions of spatial saliency and temporal novelty into a generalized biologically-inspired measure of saliency. It scores better than any of the single visual features taken in isolation, suggesting that all features do contribute to human gaze allocation.

The explicit outlier detection metric retains the same front-end as the saliency metric, but instead of biologically-inspired long-range interactions it explicitly computes likelihood of the incoming pixel data given an adaptive model for that data at every location in each of the feature maps. This metric yields a strong output when incoming data is an outlier (low likelihood) given the adaptive model (Methods). Hence this metric exactly computes outlier probabilities instead of relying on biologically-plausible heuristics as used in the saliency metric. We find that explicit outlier detection and biologically-inspired saliency score equally well, suggesting that the neural competitive interactions in the saliency metric well approximate a true detection of outliers.

Finally we evaluate the surprise metric, which retains the raw linear visual features of the saliency and outlier detection metrics, but attaches surprise detectors to every location in each of the feature maps. This metric quantifies low-level surprise in image patches over space and time, and at this point does not account for cognitive beliefs of our human observers, nor does it attempt to consider high-level, possibly semantically-rich, models for the video frames (such as the models of television channels discussed in Methods). Rather, the surprise metric assumes a family of simple models for image patches and computes surprise from shifts in the distribution of beliefs about which models better describe the

Table 2
KL scores for the eleven computational metrics studied. (a) Metrics based on static image properties overall scored lowest. (b) Metrics also sensitive to dynamic image properties scored higher, with surprise significantly the highest. For all metrics shown, humans saccaded towards image locations of higher metric response more often than expected by chance (nonparametric sign tests, $p < 10^{-100}$ for every metric). Consequently, *KL* distances between human and random (mean±S.D. from 100-times repeated random sampling) were all significantly higher than zero, which would indicate a metric not predicting human saccades better than chance (*t*-tests, $p < 10^{-100}$ or better). *KL* distances differed from one another with $p < 10^{-100}$ or better on *t*-tests for equality of the *KL* scores, except for orientation vs. intensity contrasts ($p < 10^{-12}$), variance vs. color ($p < 10^{-9}$), motion vs. flicker ($p \geqslant 0.15$), outliers vs. saliency ($p \geqslant 0.10$), suggesting a strict ordering of all eleven metrics except for equivalent performance of flicker and motion, and of saliency and outliers. While significantly above chance level, obviously, image-based, or purely bottom-up, computational metrics only explain a fraction of the correlation among humans, which encompasses both bottom-up and top-down factors.

Model	Human to random *KL* distance
(a)	
Intensity variance	0.074 ± 0.003
DCT-based information	0.101 ± 0.004
Entropy	0.151 ± 0.005
Color center-surround	0.077 ± 0.004
Intensity center-surround	0.089 ± 0.004
Orientation center-surround	0.084 ± 0.004
(b)	
Flicker center-surround	0.179 ± 0.005
Motion center-surround	0.180 ± 0.005
Outliers	0.204 ± 0.006
Saliency	0.205 ± 0.006
Surprise	0.241 ± 0.006

patches (Methods). Notably, the models used in the surprise metric are the same as in the outlier detection metric. The only difference between these two metrics is that one detects outliers based on computing likelihood while the other computes surprise. Consequently, any difference in performance at predicting human gaze patterns cannot be due to the low-level front-end or class of models used, but must reflect a difference between computing outliers and computing surprise.

We find that the surprise metric significantly outperforms the outlier and all other computational metrics ($p < 10^{-100}$ or better on *t*-tests for equality of *KL* scores), scoring nearly 20% better than the second-best metric (saliency) and 60% better than the best static metric (entropy). Surprising stimuli often substantially differ from simple feature outliers; for example, a shower of randomly-colored pixels continually excites all low-level feature detectors and outlier detection mechanisms, but rapidly becomes unsurprising.

Clearly, in our and previous eye-tracking experiments, in some situations potentially interesting targets were more numerous than in others. With many possible targets, different observers may orient towards different locations, making it more difficult for a single metric to accurately predict all observers. To investigate this, we consider (Fig. 6) subsets of human saccades where at least two, three, or all four observers simultaneously agreed on a general location of interest. Observers could have agreed based on bottom-up factors (e.g., only one visual location had strikingly interesting image appearance at that time), top-down factors (e.g., only one object qualified as the main actor), or both (e.g., a single actor was present who also had distinctive appearance). Irrespectively of the cause for agreement, it indicates consolidated belief that a location was attractive. While overall the *KL* scores of all metrics improved when progressively focusing the analysis onto only those consensus locations, dynamic metrics improved more steeply, indicating that stimuli which more reliably attracted all observers carried more flicker, motion, saliency, and surprise (Fig. 6). Surprise remained significantly the best metric to characterize these agreed-upon attractors of human gaze ($p < 10^{-100}$ or better on *t*-tests for equality of *KL* scores).

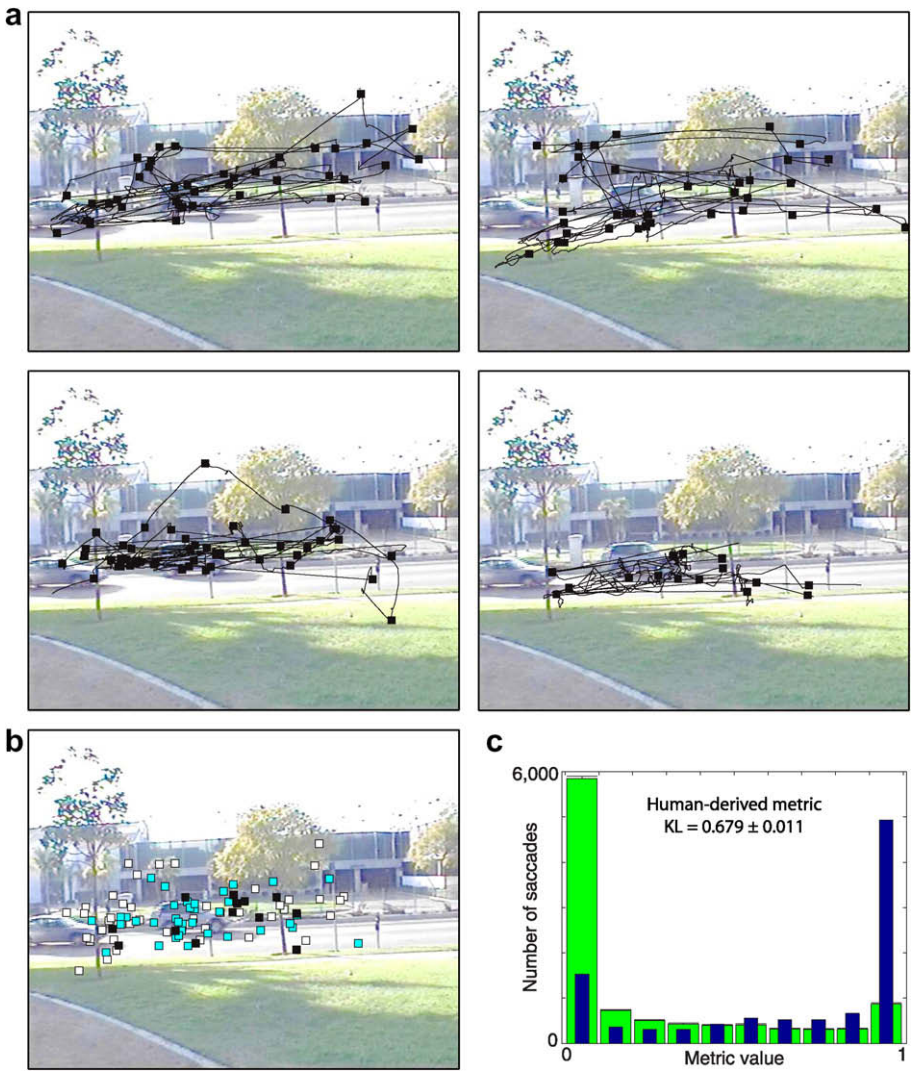

Fig. 4. (a) Sample eye movement traces from four observers (CZ, NM, RC, VN) watching one video clip (545 frames, 18.1 s) that showed cars passing by on a fairly static background. Squares denote saccade endpoints (42, 36, 48, and 16 saccades for CZ, NM, RC, and VN). (b) Our data shows high inter-individual overlap of saccade targets, as shown here with the locations where one human saccade endpoint was nearby (within 5.6°) the instantaneous eye position of one (white squares, 47 saccades), two (cyan squares, 36 saccades) or all three (black squares, 13 saccades) other humans. (c) Given this high overlap, a metric where the master map was created from the three eye movement traces other than that being tested yielded an upper-bound *KL* score, computed by comparing the histograms of metric values at human (blue) and random (green) saccade targets. Indeed, this metric's map was very sparse, as demonstrated by the high number of random saccades landing on locations with near-zero metric response. Yet humans preferentially saccaded towards the three active hotspots corresponding to the instantaneous eye positions of three other humans, as demonstrated by the high number of human saccades landing on locations with near-unity metric responses.

Overall, surprise explained the greatest fraction of human saccades, indicating that humans are significantly attracted towards surprising locations in video displays. Over 72% of all human saccades were targeted to locations predicted to be more surprising than on average. When only considering saccades where two, three, or four observers agreed on a common gaze target, this figure rose to 76, 80, and 84%, respectively.

5. Discussion

While previous research has shown with either static scenes or dynamic synthetic stimuli that humans preferentially fixate regions of high entropy (Privitera & Stark, 2000), contrast (Reinagel & Zador, 1999), saliency (Parkhurst et al., 2002), novelty (Theeuwes, 1995), or motion (Abrams & Christ, 2003), our data provides direct

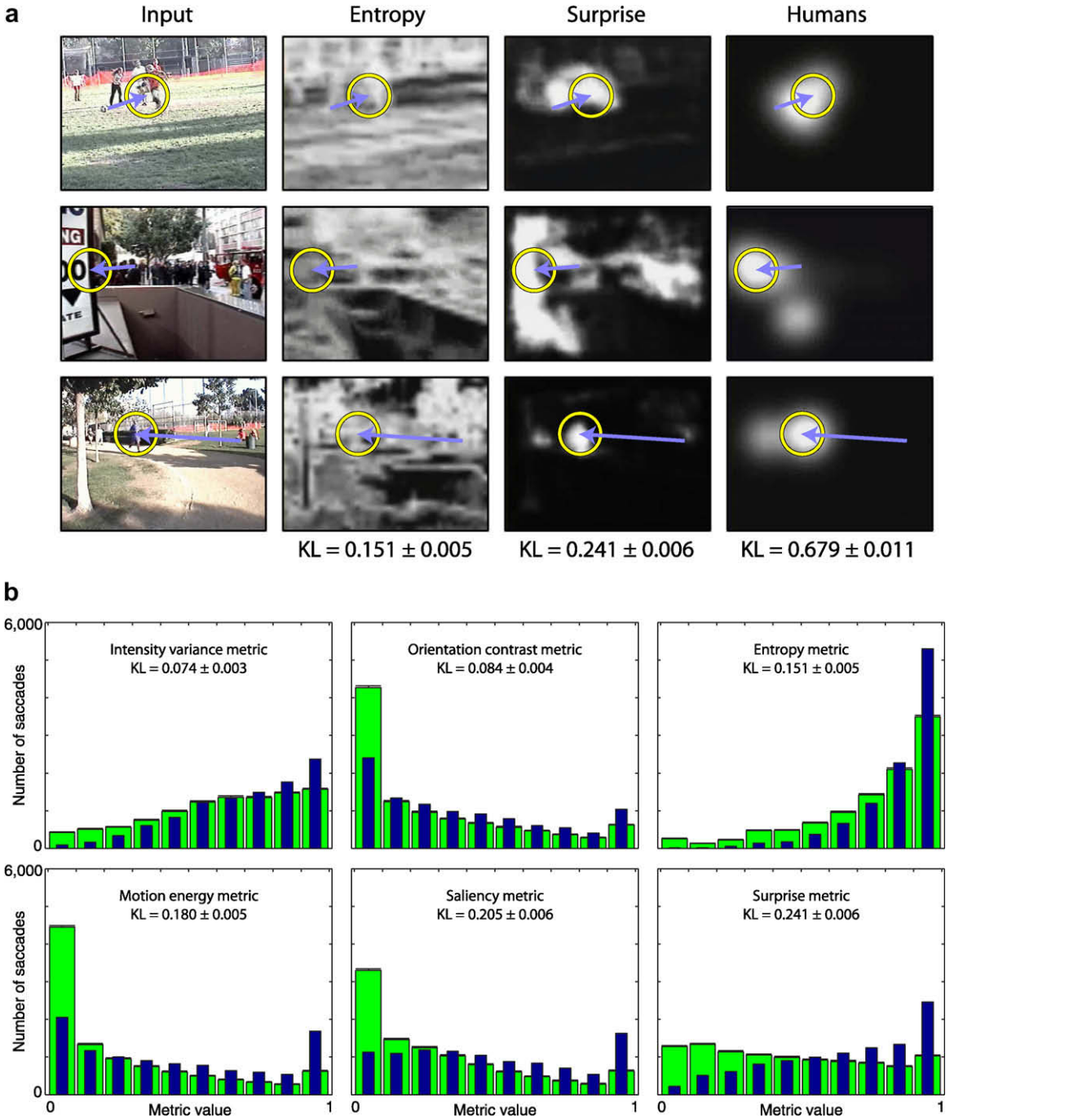

Fig. 5. (a) Sample frames from our video clips, with corresponding human saccades and predictions from the entropy, surprise, and human-derived metrics. Entropy maps, like variance and DCT-based information maps, exhibited many locations with high responses, hence had low specificity and were poorly discriminative. In contrast, surprise and human-derived maps were much sparser and more specific. For three example frames (first column), saccades from one subject are shown (arrows) with corresponding apertures over which master map activity was sampled (circles). Associated master maps exemplify the varying degrees of sparseness and specificity of the metrics tested. (b) *KL* scores quantify the tendency of human saccades (narrow blue bars) to pick hotspots with high values in the master maps, compared to chance (wide green bars, which reflect the intrinsic distributions of hotspots for each metric). A *KL* score of zero would indicate that humans did not look at hotspots in a master map more often than expected solely by chance. For all metrics studied, *KL* scores were significantly above zero, and reflected significantly different performance levels, with a strict ranking of variance < orientation < entropy < motion < saliency < surprise < human-derived (also see Table 1). Among eleven computational metrics tested in total, surprise performed best, in that surprising locations were relatively few yet reliably gazed to by humans.

experimental evidence that humans fixate surprising locations even more reliably. This conclusion was made possible by developing new analysis methods to quantify what attracts human gaze in dynamic natural scenes, and by applying these methods to large-scale data analysis totaling over one CPU-year of numerical simulations. Using these new methods and the proposed Bayesian definition of surprise, we find that surprise explains best where humans look when considering all saccades, and even more so when restricting the analysis to only those saccades for which human observers tended to agree. Surprise hence represents an easily computable shortcut towards events which deserve attention.

Beyond the early aforementioned studies comparing predictions of simple image processing metrics to human gaze patterns, a number of recent studies have further exploited such metrics under a wider variety of stimuli and behavioral conditions (Henderson, 2003; Tatler et al., 2005; Einhauser, Kruse, Hoffmann, & Konig, 2006; Foulsham & Underwood, 2008; Einhauser, Rutishauser, & Koch, 2008). However, except for a few exceptions (Navalpakkam & Itti, 2005; Torralba, Oliva, Castelhano, & Henderson, 2006; Peters & Itti, 2007), this has typically not yielded new computational metrics which would perform better than the previously proposed ones. Hence an important contribution of the present study is to develop a new theory and computational metric which significantly improves upon the state of the art.

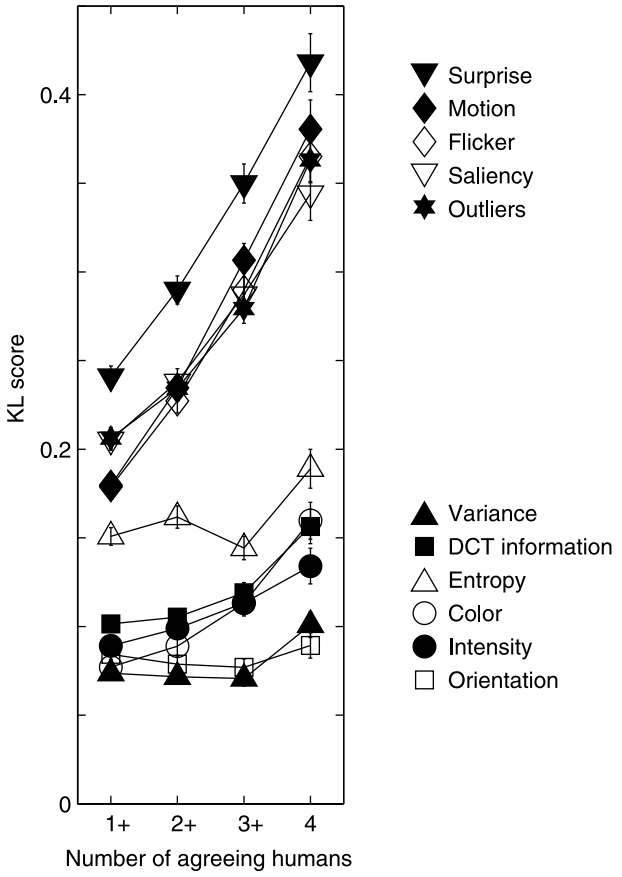

Fig. 6. *KL* scores when considering only saccades where at least one (all 10,192 saccades), two (7948 saccades), three (5565 saccades), or all four (2951 saccades) humans agreed on a general area of interest in the video clips (their gazes were within 5.6° of each other), for all eleven computational metrics. Scores of static metrics (bottom) improved substantially when progressively focusing onto only saccades with stronger inter-observer agreement (average slope 0.56 ± 0.37 percent *KL* score units per 1000 pruned saccade). Hence, when humans agreed on an important location, they also tended to be more reliably predicted by the computational metrics. Furthermore, all dynamic metrics (top) improved nearly 4.25 times more steeply (slope 2.37 ± 0.39), suggesting a stronger role of dynamic events in attracting human attention. Among those, surprising events were significantly the strongest (Bonferroni-corrected *t*-tests for equality of *KL*scores between surprise and other metrics, $p < 10^{-100}$).

In the absence of formal and quantitative tools to measure surprise, most experimental and modeling work to date has adopted the approximation that novel events are surprising, and has focused on experimental scenarios which are simple enough to ensure an overlap between the informal notions of novelty and surprise: for example, a stimulus is novel during testing if it has not been seen during training (Fecteau & Munoz, 2003). Bayesian surprise should enable the design more of sophisticated experiments, where surprise theory can directly be used to compare and calibrate surprise elicited by different stimuli, and to make predictions at the single-unit as well as behavioral levels.

The definition of surprise – as the distance between the posterior and prior distributions of beliefs over models – is entirely general and readily applicable to the analysis of auditory, olfactory, gustatory, or somatosensory data. While here we have focused on behavior rather than detailed biophysical implementation, it is worth noting that detecting surprise in neural spike trains does not require semantic understanding of the data carried by the spike trains, and thus could provide guiding signals during self-organization and development of sensory areas. In fact, the time constant of our surprise metric (Methods) was derived from fitting neural data from complex cells in areas V1 of monkeys (Müller et al., 1999).

Our implementation of surprise theory into a simple computational video processing model is certainly limited, in particular by the fact that the prior beliefs are learned here from previous video frames, at a relatively short time scale. Hence, our implementation does not yet fully exploit one of the more powerful facets of surprise theory, by which prior beliefs can in principle be of a more subjective and top-down nature.

At higher processing levels, top–down cues and task demands are known to combine with stimulus novelty in capturing attention and triggering learning (Ranganath & Rainer, 2003; Wolfe & Horowitz, 2004), ideas which may now be formalized and quantified in terms of priors, posteriors, and surprise. For instance, surprise theory can further be tested and utilized in experiments where the prior is biased by top-down instructions or prior exposures to stimuli (Wolfe & Horowitz, 2004). Indeed, within the prior distribution, surprise can in principle incorporate information coming from past experience and/or top-down. Thus the formalism is in place and this is not a theoretical limitation of the notion of surprise itself.

For instance, one possible direction for future work is to use text-based stimuli where the amount of semantic information can be finely tuned, progressively transitioning from random assemblages of letters, to Markov models of order 1, 2, 3, or higher at the letter level, to Markov models of order 1, 2, 3, or higher at the word level, to full sentences (Shannon, 1948). One question which can then be addressed is that of how the increasing level of semantic information might influence eye movements, and whether a computational implementation of surprise theory can be embodied which captures such influences.

In addition, surprise-based behavioral measures, such as the eye-tracking one used here, may prove useful for early diagnostic of human conditions including autism and attention-deficit hyperactive disorder, as well as for quantitative comparison between humans and animals which may have lower or different priors, including monkeys, frogs, and flies. Beyond sensory neurobiology and human psychology, computable surprise could guide the development of data mining and compression systems (allocating more resources to surprising regions of interest), to find surprising agents in crowds, surprising sentences in books or speeches, surprising medical symptoms, surprising odors in airport luggage racks, surprising documents on the world-wide-web, or to design surprising advertisements.

Acknowledgments

Supported by NSF, DARPA, HFSP, and NGA (L.I.), and NIH and NSF (P.B.). We thank UCI's Institute for Genomics and Bioinformatics and USC's HPCC center for access to their computing clusters, both used to carry out the model predictions reported here. The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.

References

Abrams, R. A., & Christ, S. E. (2003). Motion onset captures attention. *Psychological Science, 14*(5), 427–432.

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science, 9*, 147–169.

Benjamin, J., Li, L., Patterson, C., Greenberg, B. D., Murphy, D. L., & Hamer, D. H. (1996). Population and familial association between the D4 dopamine receptor gene and measures of Novelty seeking. *Nature Genetics, 12*(1), 81–84.

Bradley, J., Bonigk, W., Yau, K. W., & Frings, S. (2004). Calmodulin permanently associates with rat olfactory CNG channels under native conditions. *Nature Neuroscience, 7*(7), 705–710.

Brown, L. D. (1986). *Fundamentals of statistical exponential families*. Hayward, CA: Institute of Mathematical Statistics.

Brown, S. P., & Masland, R. H. (2001). Spatial scale and cellular substrate of contrast adaptation by retinal ganglion cells. *Nature Neuroscience, 4*(1), 44–51.

Cox, R. T. (1964). Probability, frequency and reasonable expectation. *American Journal of Physics, 14*, 1–13.

David, S. V., Vinje, W. E., & Gallant, J. L. (2004). Natural stimulus statistics alter the receptive field structure of v1 neurons. *Journal of Neuroscience, 24*(31), 6991–7006.

Doretto, G., Chiuso, A., Wu, Y., & Soatto, S. (2003). Dynamic textures. *International Journal of Computer Vision, 51*(2), 91–109.

Dragoi, V., Sharma, J., Miller, E. K., & Sur, M. (2002). Dynamics of neuronal sensitivity in visual cortex and local feature discrimination. *Nature Neuroscience, 5*(9), 883–891.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Reveiw, 96*(3), 433–458.

Ebstein, R. P., Novick, O., Umansky, R., Priel, B., Osher, Y., Blaine, D., et al. (1996). Dopamine D4 receptor (D4DR) exon III polymorphism associated with the human personality trait of Novelty seeking. *Nature Genetics, 12*(1), 78–80.

Einhauser, W., Kruse, W., Hoffmann, K. P., & Konig, P. (2006). Differences of monkey and human overt attention under natural conditions. *Vision Research, 46*(8-9), 1194–1209.

Einhauser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision, 8*(2), 201–219.

Fairhall, A. L., Lewen, G. D., Bialek, W., & de Ruyter Van Steveninck, R. R. (2001). Efficiency and ambiguity in an adaptive neural code. *Nature, 412*(6849), 787–792.

Fecteau, J. H., & Munoz, D. P. (2003). Exploring the consequences of the previous trial. *Nature Reviews. Neuroscience, 4*(6), 435–443.

Field, D.J. 2005. Entropy, visual non-linearities and the higher-order statistics of natural scenes. In: Proceedings of CVR Vision Conference.

Finney, S. A. (2001). Real-time data collection in Linux: A case study. *Behavior Research Methods Instruments and Computers, 33*, 167–173.

Fletcher, P. C., Anderson, J. M., Shanks, D. R., Honey, R., Carpenter, T. A., Donovan, T., et al. (2001). Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. *Nature Neuroscience, 4*(10), 1043–1048.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision, 8*(2), 601–617.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences, 7*(11), 498–504.

Itti, L., & Baldi, P. 2005. A Principled Approach to Detecting Surprising Events in Video. Proceedings in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 631–637.

Itti, L., & Baldi, P. (2006). *Bayesian surprise attracts human attention. Advances in neural information processing systems* (19). Cambridge, MA.: MIT Press. pp. 547–554 NIPS*2005.

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

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.

Jaynes, E. T. (2003). *Probability theory. The logic of science*. Cambridge University Press.

Kennedy, H. J., Evans, M. G., Crawford, A. C., & Fettiplace, R. (2003). Fast adaptation of mechanoelectrical transducer channels in mammalian cochlear hair cells. *Nature Neuroscience, 6*(8), 832–836.

Knight, R. (1996). Contribution of human hippocampal region to novelty detection. *Nature, 383*(6597), 256–259.

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

Kurahashi, T., & Menini, A. (1997). Mechanism of odorant adaptation in the olfactory receptor cell. *Nature, 385*(6618), 725–729.

Li, S., Cullen, W. K., Anwyl, R., & Rowan, M. J. (2003). Dopamine-dependent facilitation of LTP induction in hippocampal CA1 by exposure to spatial novelty. *Nature Neuroscience, 6*(5), 526–531.

Lusher, J. M., Chandler, C., & Ball, D. (2001). Dopamine D4 receptor gene (DRD4) is associated with Novelty seeking (NS) and substance abuse: The saga continues. *Molecular Psychiatry, 6*(5), 497–499.

Maffei, L., Fiorentini, A., & Bisti, S. (1973). Neural correlate of perceptual adaptation to gratings. *Science, 182*(116), 1036–1038.

Markou, M., & Singh, S. (2003). Novelty detection: A review – Part 1: Statistical approaches. *Signal Processing, 83*(12), 2481–2497.

Movshon, J. A., & Lennie, P. (1979). Pattern-selective adaptation in visual cortical neurones. *Nature, 278*(5707), 850–852.

Müller, J. R., Metha, A. B., Krauskopf, J., & Lennie, P. (1999). Rapid adaptation in visual cortex to the structure of images. *Science, 285*(5432), 1405–1408.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research, 45*(2), 205–231.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*(6583), 607–609.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*(1), 107–123.

Peters, R.J., & Itti, L. 2007 (Jun). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Privitera, C. M., & Stark, L. W. (2000). Algorithms for defining visual regions-of-interest: comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(9), 970–982.

Ranganath, C., & Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. *Nature Reviews. Neuroscience, 4*(3), 193–202.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79–87.

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

Savage, L. J. (1972). *The foundations of statistics*. New York: Dover. First Edition in 1954.

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience, 23*, 473–500.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423. 623–656.

Sillito, A. M., Grieve, K. L., Jones, H. E., Cudeiro, J., & Davis, J. (1995). Visual cortical mechanisms detecting focal orientation discontinuities. *Nature, 378*(6556), 492–496.

Smirnakis, S. M., Berry, M. J., Warland, D. K., Bialek, W., & Meister, M. (1997). Adaptation of retinal processing to image contrast and spatial scale. *Nature, 386*(6620), 69–73.

Softky, W. R., & Koch, C. (1993). The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *Journal of Neuroscience, 13*(1), 334–350.

Solomon, S. G., Peirce, J. W., Dhruv, N. T., & Lennie, P. (2004). Profound contrast adaptation early in the visual pathway. *Neuron, 42*(1), 155–162.

Stern, C. E., Corkin, S., Gonzalez, R. G., Guimaraes, A. R., Baker, J. R., Jennings, P. J., et al. (1996). The hippocampal formation participates in novel picture encoding: Evidence from functional magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America, 93*(16), 8660–8665.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45*(5), 643–659.

Theeuwes, J. (1995). Abrupt luminance change pops out; abrupt color change does not. *Percept Psychophys, 57*(5), 637–644.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Reveiw, 113*(4), 766–786.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136.

Ulanovsky, N., Las, L., & Nelken, I. (2003). Processing of low-probability sounds by cortical neurons. *Nature Neuroscience, 6*(4), 391–398.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews. Neuroscience, 5*(6), 495–501.