Neural Networks 23 (2010) 649–666

Contents lists available at ScienceDirect

Neural Networks

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

Of bits and wows: A Bayesian theory of surprise with applications to attention

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

[a] Department of Computer Science, UCI, Irvine, CA 92697-3435, United States

[b] Department of Computer Science, USC, Los Angeles, CA 90089-2520, United States

ARTICLE INFO

Article history:
Received 29 April 2009
Received in revised form 3 December 2009
Accepted 18 December 2009

Keywords:
Information
Surprise
Relative entropy
Attention
Eye movements

ABSTRACT

The amount of information contained in a piece of data can be measured by the effect this data has on its observer. Fundamentally, this effect is to transform the observer's prior beliefs into posterior beliefs, according to Bayes theorem. Thus the amount of information can be measured in a natural way by the distance (relative entropy) between the prior and posterior distributions of the observer over the available space of hypotheses. This facet of information, termed "surprise", is important in dynamic situations where beliefs change, in particular during learning and adaptation. Surprise can often be computed analytically, for instance in the case of distributions from the exponential family, or it can be numerically approximated. During sequential Bayesian learning, surprise decreases as the inverse of the number of training examples. Theoretical properties of surprise are discussed, in particular how it differs and complements Shannon's definition of information. A computer vision neural network architecture is then presented capable of computing surprise over images and video stimuli. Hypothesizing that surprising data ought to attract natural or artificial attention systems, the output of this architecture is used in a psychophysical experiment to analyze human eye movements in the presence of natural video stimuli. Surprise is found to yield robust performance at predicting human gaze (ROC-like ordinal dominance score ∼0.7 compared to ∼0.8 for human inter-observer repeatability, ∼0.6 for simpler intensity contrast-based predictor, and 0.5 for chance). The resulting theory of surprise is applicable across different spatio-temporal scales, modalities, and levels of abstraction.

1. Introduction

The concept of information is central to science, technology, and many other human endeavors. While several approaches for quantifying information have been proposed, the most prominent one so far has been Claude Shannon's definition introduced over half a century ago (Aczel & Daroczy, 1975; Blahut, 1987; Cover & Thomas, 1991; McEliece, 1977; Shannon, 1948). According to this definition, the amount of information contained in a piece of data D is measured by $-\log_2 P(D)$ bits—a rare piece of data with small probability is considered more informative. Although eminently successful for the development of modern telecommunication and computer technologies, Shannon's definition does not capture all aspects of information. Here we look at information from a different angle. Starting from Bayes theorem, we notice that the fundamental effect that data has on a given observer is to change his/her/its prior beliefs into posterior beliefs. Thus we propose to measure the effect D has on the observer by the distance between his prior and posterior belief distributions. We call this facet of information surprise.

Surprise plays an important role in *dynamic* situations when the beliefs of the observer change significantly in time, as a result of consecutive applications of Bayes theorem. This can happen in at least two broad categories of situations: either when the beliefs keep changing all the time without converging to a stable value, or when the beliefs progressively converge to a stable value. The first case corresponds to tracking or adaption in a non-stationary environment. The second case corresponds to learning from a stationary data set, when beliefs evolve but finally converge to a stable value. In our framework, adaptation and learning are the results of the same fundamental operation: belief update using Bayes theorem. What distinguishes them is not the basic underlying mathematical operation, but rather the memory span and time scales involved.

In what follows, in Section 2 we first provide the mathematical definition of surprise. In Section 3, we show how surprise can be computed exactly or approximated efficiently in most common situations. In Section 4, we study how surprise changes during learning. In Section 5, we investigate the connections between

* Corresponding address: School of Information and Computer Sciences and Institute for Genomics and Bioinformatics, University of California Irvine, Irvine, CA 92697-3435, United States. Tel.: +1 949 824 5809; fax: +1 949 824 9813.
E-mail addresses: pfbaldi@ics.uci.edu (P. Baldi), itti@usc.edu (L. Itti).
[1] University of Southern California, Hedco Neuroscience Building, 3641 Watt Way Los Angeles, CA 90089-2520, United States. Tel.: +1 213 740 3527; fax: +1 213 740 5687.

0893-6080/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2009.12.007

surprise and other theories of information and novelty, including Shannon's theory of information. In Section 6, we describe a neural network architecture for computing surprise over image and video data in computer vision. Finally, in Section 7, we conduct psychophysical experiments and apply the surprise architecture to the problem of modeling attention and predicting rapid eye movements in humans watching natural stimuli.

The present paper builds on our previously published reports on the same theme (Baldi, 2002; Itti & Baldi, 2005, 2006, 2009). The two key new components here are: (1) a detailed and self-contained treatment of the theory with derivation of closed-form expressions for computing surprise in a number of important cases, analysis of the relationship between surprise and Bayesian learning, and comparison of surprise to other theories of information and novelty; and (2) a detailed presentation of the computational model of attention developed to investigate to which extent Bayesian surprise may predict what attracts human gaze while watching natural video clips. In addition, we also describe new experimental results comparing two variants of our computational model (using either Gaussian data and a Gaussian prior, or Poisson data and a Gamma prior; see below) on two open-access human gaze tracking datasets, and we develop a new temporal analysis that reveals a strong time contingency between the onset of a surprising event in a video clip and the execution of a human gaze shift towards such event. The reader is invited to explore our previous publications for more extensive discussions of the eye-tracking methodologies, general modeling of human saliency maps, and methodologies for comparing human gaze recordings to saliency maps generated by a number of different models available in the literature. All the data collected for the experiments have been made publicly available.

2. Mathematical definition of surprise

The definition we propose is best understood within the Bayesian or subjectivist framework of probability theory. In the subjectivist framework, degrees of belief or confidence are associated with hypotheses or models. It can be shown that under a small set of reasonable axioms, these degrees of belief can be represented by real numbers and that when rescaled to the [0, 1] interval they must obey the rules of probability and in particular Bayes theorem (Berger, 1985; Cox, 1964; Gelman, Carlin, Stern, & Rubin, 1995; Jaynes, 1986, 2003; Savage, 1972). The amount of surprise in the data for a given observer can be measured by looking at the changes that take place in going from the prior to the posterior distributions.

Specifically, consider an observer with a prior distribution $P(M)$ over a set $\mathcal{M}$ of possible models or hypotheses. The collection of a piece of data D leads to a reevaluation of beliefs and the transformation of the prior probability into a posterior distribution according to Bayes theorem,

$$P(M|D) = \frac{P(D|M)}{P(D)} P(M). \tag{1}$$

From this equation, the effect of D is clearly to change $P(M)$ to $P(M|D)$. *In other words, we view the data D as an operator acting on the space of distributions over the space of models. Thus, one basic way of measuring information carried by D is to measure the distance between the prior and the posterior distributions.* To distinguish it from Shannon's communication information, we call this notion of information the surprise information, or just *surprise* (Baldi, 2002):

$$S(D, \mathcal{M}) = d[P(M), P(M|D)] \tag{2}$$

where d is a distance or dissimilarity measure. There are different ways of measuring distance or dissimilarity between probability distributions. In what follows, for standard well-known theoretical reasons such as invariance with respect to reparameterizations, we use the relative entropy or Kullback–Liebler (Kullback, 1968) divergence K

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

where H denotes the entropy.

The alternative version $K(P(M|D), P(M))$ of the relative entropy may also be used (and may even be slightly preferable in settings where the "true" or "best" distribution is used as the first argument). While the basic principles in the following derivations apply to both forms, here we use the version in Eq. (3) because in general it leads to slightly simpler analytical expressions. This is simply because the prior distribution, which occurs twice in Eq. (3), in general has a simpler expression than the posterior distribution which contains additional data-dependent terms. Alternatively, the relative entropy might also be symmetrized by taking $[K(P(M), P(M|D)) + K(P(M|D), P(M))]/2$. Although the symmetric version is rarely used, the analytical formula to be derived could be applied to the symmetric version with the proper and obvious adjustments. The same applies to other variations, such as the Jensen–Shannon divergence (Lin, 1991; Wong & You, 1985). In considering the symmetric version, note, however, that there is no reason why the intuitive notion of surprise ought to be symmetric with respect to the distributions involved. In fact, introspection dictates that the contrary ought to be true. A broad prior distribution followed by a narrow posterior distribution corresponds to a reduction in uncertainty, while a narrow prior distribution followed by a broad posterior distribution corresponds to an increase in uncertainty, and both lead to different subjective experiences.

Equivalently, we can define the single-model surprise by the log-odd ratio

$$S(D, M) = \log \frac{P(M)}{P(M|D)} \tag{4}$$

and the surprise by its average

$$S(D, \mathcal{M}) = \int_{\mathcal{M}} S(D, M) P(M) \mathrm{d}M \tag{5}$$

taken with respect to the prior distribution over the model class. In statistical mechanics terminology, surprise can also be viewed as the free energy of the negative log-posterior at a temperature $t = 1$, with respect to the prior distribution over the space of models (Baldi & Brunak, 2001).

A unit of surprise – the "wow" – can be defined for a single model M as the amount of surprise corresponding to a two-fold variation between the prior and the posterior, i.e., as $-\log_2 P(M)/P(M|D)$. Note that unless we use absolute values, this ratio can be positive or negative depending on whether the observer's belief in model M increases or decreases. The total number of wows experienced when simultaneously considering all models is obtained by integrating over $\mathcal{M}$ and, as a relative entropy, is always positive.

3. Analytical computation or approximation of surprise

As a relative entropy, surprise can always be estimated, at least numerically. But for the concept to be really useful, one ought to be able to compute surprise analytically, at least in the most standard statistical cases.

More precisely, consider a data set $D = \{d_1, \ldots, d_N\}$ containing N points. For simplicity, although this does not correspond to any restriction in the general theory, we consider the case of conjugate priors, where the prior and the posterior have the same functional form. In addition to their theoretical interest, conjugate priors are also important for efficient implementations of iterative Bayesian learning where the posterior at iteration t becomes the prior for iteration $t+1$. In order to compute surprise in Eq. (3) with conjugate priors, we need only to compute general terms of the form

$$F(P_1, P_2) = \int P_1 \log P_2 \mathrm{d}x \tag{6}$$

where $P_1(x)$ and $P_2(x)$ are two distributions with the same functional form. The surprise is then given by

$$S = F(P_1, P_1) - F(P_1, P_2) \tag{7}$$

where P_1 is the prior and P_2 is the posterior. Note also that in this case the symmetric divergence can easily be computed using $F(P_1, P_1) - F(P_1, P_2) + F(P_2, P_2) - F(P_2, P_1)$. Details for the calculation of $F(P_1, P_2)$ in the following specific examples are given in the Appendix.

3.1. Discrete data: Multinomial model

Consider the case where d_i is binary. The simplest class of models for D is then $M(p)$, the first order Markov model with a single parameter p representing the probability of emitting a 1. The conjugate prior on p is the Beta distribution (or Dirichlet distribution in the general multinomial case)

$$\begin{aligned} D_1(a_1, b_1) &= \frac{\Gamma(a_1 + b_1)}{\Gamma(a_1)\Gamma(b_1)} x^{a_1 - 1}(1 - x)^{b_1 - 1} \\ &= C_1 x^{a_1 - 1}(1 - x)^{b_1 - 1} \end{aligned} \tag{8}$$

with parameters $a_1 \geq 0, b_1 \geq 0$, and $a_1 + b_1 > 0$. The expectation is $a_1/(a_1 + b_1)$. With n successes in the sequence D of N samples, the posterior is a Dirichlet distribution $D_2(a_2, b_2)$ with Baldi and Brunak (2001)

$$a_2 = a_1 + n \quad \text{and} \quad b_2 = b_1 + (N - n). \tag{9}$$

The surprise can be computed exactly

$$\begin{aligned} S(D, \mathcal{M}) &= K(D_1, D_2) \\ &= \log \frac{C_1}{C_2} + n[\Psi(a_1 + b_1) - \Psi(a_1)] \\ &\quad + (N - n)[\Psi(a_1 + b_1) - \Psi(b_1)] \end{aligned} \tag{10}$$

where Ψ is the derivative of the logarithm of the Gamma function (see Appendix). When $N \to \infty$, and $n = pN$ with $0 < p < 1$ we have

$$\begin{aligned} S(D, \mathcal{M}) &\approx NK(p, a_1) \\ &= NK\left((p, 1 - p), \left(\frac{a_1}{a_1 + b_1}, \frac{b_1}{a_1 + b_1}\right)\right) \end{aligned} \tag{11}$$

where $K(p, a_1)$ is a concise notation to represent the Kullback–Liebler divergence between the empirical distribution $(p, 1-p)$ and the expectation of the prior $(a_1/(a_1 + b_1), b_1/(a_1 + b_1))$. Thus asymptotically surprise grows linearly with the number of data points with a proportionality coefficient that depends on the discrepancy between the expectation of the prior and the observed distribution. The same relationship is true in the case of a multinomial model. When the prior is symmetric $(a_1 = b_1)$, a slightly more precise approximation is provided by

$$S(D, \mathcal{M}) = K(D_1, D_2) \approx N\left[\sum_{k=a_1}^{2a_1 - 1} \frac{1}{k} - H(p)\right] \tag{12}$$

where $H(p)$ denotes the entropy $H(p) = -p \log p - (1-p)\log(1-p)$. For instance, when $a_1 = 1$ then $K(D_1, D_2) \approx N(1 - H(p))$, and when $a_1 = 5$ then $K(D_1, D_2) \approx N[0.746 - H(p)]$.

3.2. Discrete data: Poisson model

As a second discrete example, consider the case where d_i is an integer. A simple class of models for D is the class of Poisson models parameterized by λ. The conjugate prior on λ is the Gamma prior

$$\Gamma_1(a_1, b_1) = \frac{b_1^{a_1}}{\Gamma(a_1)} x^{a_1 - 1} \mathrm{e}^{-b_1 x} = C_1 x^{a_1 - 1} \mathrm{e}^{-b_1 x} \tag{13}$$

with $x \geq 0$, shape $a_1 > 0$, inverse scale $b_1 > 0$. The expectation is a_1/b_1. With N observations, the posterior is also a Gamma distribution $\Gamma_2(a_2, b_2)$ with

$$a_2 = a_1 + N\bar{m} \quad \text{and} \quad b_2 = b_1 + N \tag{14}$$

where $\bar{m}$ is the sample mean. The surprise can be computed exactly

$$\begin{aligned} S(D, \mathcal{M}) &= K(\Gamma_1, \Gamma_2) \\ &= a_1 \log \frac{b_1}{b_1 + N} - N\bar{m} \log(b_1 + N) + \log \frac{\Gamma(a_1 + N\bar{m})}{\Gamma(a_1)} \\ &\quad + \frac{Na_1}{b_1} + N\bar{m}[\log b_1 - \Psi(a_1)]. \end{aligned} \tag{15}$$

When $N \to \infty$, Stirling's formula yields the approximation

$$S(D, \mathcal{M}) \approx N\left(\frac{a_1}{b_1} - \bar{m}[1 - \log \bar{m} + \Psi(a_1) - \log b_1]\right). \tag{16}$$

Thus asymptotically surprise information grows linearly with the number of data points with a proportionality coefficient that depends on the difference between the mean a_1/b_1 of the prior and mean $\bar{m}$ of the sample plus an offset.

3.3. Continuous data: Unknown mean/known variance

When the d_i are real, we can consider first the case of unknown mean with a known variance. We have a family $M(\mu)$ of models, with a Gaussian prior $G_1(\mu_1, \sigma_1^2)$. If the data has known variance σ^2, then the posterior distribution is Gaussian $G_2(\mu_2, \sigma_2^2)$ with parameters given by (Gelman et al., 1995)

$$\mu_2 = \frac{\frac{\mu_1}{\sigma_1^2} + \frac{N\bar{m}}{\sigma^2}}{\frac{1}{\sigma_1^2} + \frac{N}{\sigma^2}} \quad \text{and} \quad \frac{1}{\sigma_2^2} = \frac{1}{\sigma_1^2} + \frac{N}{\sigma^2} \tag{17}$$

where $\bar{m}$ is the observed mean. In this case

$$\begin{aligned} S(D, \mathcal{M}) &= K(G_1, G_2) \\ &= \log \frac{\sigma}{\sqrt{\sigma^2 + N\sigma_1^2}} + N\frac{\sigma_1^2}{2\sigma^2} + \frac{N^2\sigma_1^2(\mu_1 - \bar{m})^2}{2\sigma^2(\sigma^2 + N\sigma_1^2)} \\ &\approx \frac{N}{2\sigma^2}[\sigma_1^2 + (\mu_1 - \bar{m})^2] \end{aligned} \tag{18}$$

the approximation being valid for large N. In the special case where the prior has the same variance as the data $\sigma_1 = \sigma$ then the formula simplifies a little and yields

$$\begin{aligned} S &= K(G_1, G_2) \\ &= \frac{N}{2} - \frac{1}{2}\log(N + 1) + \frac{N^2(\mu_1 - \bar{m})^2}{2(N + 1)\sigma^2} \\ &\approx \frac{N}{2\sigma^2}[\sigma^2 + (\mu_1 - \bar{m})^2] \end{aligned} \tag{19}$$

the last approximation being valid when N is large. In any case, surprise grows linearly with N with a coefficient that is the sum of the prior variance and the square difference between the expected mean and the empirical mean scaled by the variance of the data.

3.4. Continuous data: Unknown variance/known mean

When the d_is are real, we can also consider the case of unknown variance with a known mean. We then have a family $M(\sigma^2)$ of models, with a conjugate scaled inverse gamma prior (Gelman et al., 1995)

$$\Gamma_1(\nu_1, s_1) = \frac{(\frac{\nu_1}{2})^{\nu_1/2} s_1^{\nu_1}}{\Gamma(\frac{\nu_1}{2})} (\sigma^2)^{-(\frac{\nu_1}{2}+1)} e^{-\frac{\nu_1 s_1^2}{2\sigma^2}} = C_1(\sigma^2)^{-(\frac{\nu_1}{2}+1)} e^{-\frac{\nu_1 s_1^2}{2\sigma^2}}. \quad (20)$$

The posterior is then a scaled inverse gamma distribution (Gelman et al., 1995) with

$$\nu_2 = \nu_1 + N \quad \text{and} \quad s_2^2 = \frac{\nu_1 s_1^2 + N\bar{\sigma}^2}{\nu_1 + N}. \quad (21)$$

Here $\bar{\sigma}^2 = \sum (x_i - m)^2/N$ is the observed variance, based on the known mean m. The surprise

$$S(D, \mathcal{M}) = K(\Gamma_1, \Gamma_2) = \log \frac{C_1}{C_2} - \frac{N}{2}\left[\Psi\left(\frac{\nu}{2}\right) + \log \frac{2}{\nu_1 s_1^2}\right] + \frac{N\bar{\sigma}^2}{2s_1^2}. \quad (22)$$

For large values of N,

$$S = K(\Gamma_1, \Gamma_2) \approx \frac{N}{2}\left(\frac{\bar{\sigma}^2}{s_1^2} + \log \frac{\nu_1 s_1^2}{2\bar{\sigma}^2} - \Psi\left(\frac{\nu_1}{2}\right)\right). \quad (23)$$

Thus surprise information scales linearly with N, with a coefficient of proportionality that typically depends mostly on the ratio of the empirical variance to the scale parameter s_1^2, which is roughly the expectation of the prior [the expectation of the prior is $\nu_1 s_1^2/(\nu_1 - 2)$ provided $\nu_1 > 2$]. The effects of very large or very small values of $\bar{\sigma}$ or ν_1 can also be seen in the formula above. In particular, surprise is largest when the empirical variance $\bar{\sigma}^2$ goes to 0 or infinity, i.e., is very different from the prior expectation.

3.5. Continuous data: Unknown mean/unknown variance

Finally, we can consider the case of unknown mean with unknown variance. We have a family $M(\mu, \sigma^2)$ of models, with a conjugate prior $G_1\Gamma_1 = P(\mu|\sigma^2)P(\sigma^2) = G_1(\mu_1, \sigma^2/\kappa_1)\Gamma_1(\nu_1, s_1)$, product of a normal with a scaled inverse Gamma distribution. Thus the prior has four parameters $(\mu_1, \kappa_1, \nu_1, s_1)$, with $\kappa_1 > 0$, $\nu_1 > 0$, and $s_1 > 0$. The conjugate posterior has the same form, with similar parameters $(\mu_2, \kappa_2, \nu_2, s_2)$ satisfying (see for instance Gelman et al. (1995))

$$\mu_2 = \frac{\kappa_1}{\kappa_1 + N}\mu_1 + \frac{N}{\kappa_1 + N}\bar{m}$$
$$\kappa_2 = \kappa_1 + N$$
$$\nu_2 = \nu_1 + N$$
$$\nu_2 s_2^2 = \nu_1 s_1^2 + (N-1)\bar{\sigma}^2 + \frac{\kappa_1 N}{\kappa_1 + N}(\bar{m} - \mu_1)^2$$

with $\bar{m} = \sum x_i/N$ and $\bar{\sigma}^2 = \sum (x_i - \bar{m})^2/(N-1)$. The surprise is

$$S(D, \mathcal{M}) = K(G_1\Gamma_1, G_2\Gamma_2) = \frac{1}{2}\log \frac{\kappa_1}{\kappa_1 + N} + \frac{N}{2\kappa_1}\frac{\kappa_1 + N}{2}\left[\frac{N(\bar{m} - \mu_1)}{(\kappa_1 + N)s_1}\right]^2 + \log \frac{C_1}{C_2} - \frac{N}{2}\left[\Psi\left(\frac{\nu_1}{2}\right) + \log \frac{2}{\nu_1 s_1^2}\right] + \frac{(N-1)\bar{\sigma}^2 + \frac{\kappa_1 N}{\kappa_1 + N}(\bar{m} - \mu_1)^2}{2s_1^2}. \quad (24)$$

For large values of N,

$$K(G_1\Gamma_1, G_2\Gamma_2) \approx \frac{N}{2}\left(\frac{1}{\kappa_1} + \frac{\bar{\sigma}^2}{s_1^2} + \log \frac{\nu_1 s_1^2}{2\bar{\sigma}^2} - \Psi\left(\frac{\nu_1}{2}\right) + \frac{(\bar{m} - \mu_1)^2}{s_1^2}\right). \quad (25)$$

Surprise information is linear in N with a coefficient that is essentially the sum of the coefficients derived in the unknown mean and unknown variance partial cases.

3.6. Generalization: Exponential families with conjugate priors

The previous examples can be generalized by considering a family $M(\theta)$ of models parameterized by the parameter vector θ with a likelihood function associated with the exponential family of distributions (Brown, 1986)

$$P(d|M) = h(d)c(\theta)\exp\left(\sum_{i=1}^{k} w_i(\theta)t_i(d)\right) \quad (26)$$

where $h(d)$, $c(\theta)$, $w_i(\theta)$, and $t_i(d)$ are known functions of the respective variables. With N independent data points $(D = d_1, \ldots, d_N)$

$$P(D|M) = [c(\theta)]^N h(D)\exp\left(\sum_{i=1}^{k} w_i(\theta)\sum_{j=1}^{N} t_i(d_j)\right) = [c(\theta)]^N h(D)\exp\left(\sum_{i=1}^{k} w_i(\theta)T_i(D)\right) \quad (27)$$

where $h(D) = \prod_{j=1}^{N} h(d_j)$, and $T_i(D) = \sum_{j=1}^{N} t_i(d_j)$ are the sufficient statistics. The conjugate prior has a similar exponential form

$$P(\theta) = C_1(\theta)\exp\left(\sum_{i=1}^{k} A_i^1 w_i(\theta)\right) \quad (28)$$

parameterized by the A_i^1's. Using Bayes theorem, the posterior has the same exponential form

$$P(\theta) = C_2(\theta)\exp\left(\sum_{i=1}^{k} A_i^2 w_i(\theta)\right) \quad (29)$$

parameterized by the A_i^2's satisfying

$$A_i^2 = A_i^1 + T_i(D). \quad (30)$$

Calculation of surprise yields

$$S(D, \mathcal{M}) = \log \frac{C_1}{C_2} - \sum_{i=1}^{k} T_i(D)E_{A^1}[w_i(\theta)] \quad (31)$$

where $E_{A^1}[w_i(\theta)]$ is the expectation of $w_i(\theta)$ with respect to the prior. Note that if surprise is defined by $K(P(M|D), P(M))$, the same calculation yields

$$S(D, \mathcal{M}) = \log \frac{C_2}{C_1} + \sum_{i=1}^{k} T_i(D)E_{A^2}[w_i(\theta)]. \quad (32)$$

Thus for members of the exponential family (Brown, 1986) of distributions, the posterior depends entirely on the sufficient statistics and therefore the surprise also depends crucially on them. The $T_i(D)$ terms typically grow linearly with the data, and so does surprise.

4. Learning and surprise

There is an immediate connection between surprise and computational learning theory. If we imagine that data points from a training set are presented sequentially, we can consider that the posterior distribution after the N-th point becomes the prior for the next iteration (sequential Bayesian learning). As a system learns from examples with a static distribution, new data points ought to become less and less surprising. Thus in this case we can expect on average surprise to decrease after each iteration. We shall compute the exact rate of decrease using examples taken from the distributions studied in the previous section.

4.1. Learning curves: Discrete data

Consider first a sequence of 0–1 examples $D = (d_N)$. The learner starts with a Dirichlet prior $D_0(a_0, b_0)$. With each example d_N, the learner updates its Dirichlet prior $D_N(a_N, b_N)$ into a Dirichlet posterior $D_{N+1}(a_{N+1}, b_{N+1})$ with $(a_{N+1}, b_{N+1}) = (a_N + 1, b_N)$ if $d_{N+1} = 1$, and $(a_{N+1}, b_{N+1}) = (a_N, b_N + 1)$ otherwise. When $d_{N+1} = 1$, the corresponding surprise is easily computed using Eqs. (56) and (59) (detailed in Appendix). For simplicity, and without much loss of generality, let us assume that a_0 and b_0 are integers, so that a_N and b_N are also integers for any N. Then if $d_{N+1} = 1$ the relative surprise is

$$S(D_N, D_{N+1}) = \log \frac{a_N}{a_N + b_N} + \sum_{k=0}^{b_N - 1} \frac{1}{a_N + k} \tag{33}$$

and similarly in the case $d_{N+1} = 0$ by interchanging the role of a_N and b_N. By using the standard integral bound for series based on monotonically decreasing functions, we have

$$\sum_{k=0}^{b_N - 1} \frac{1}{a_N + k} \leq \frac{1}{a_N} + \log \frac{a_N + b_N - 1}{a_N}. \tag{34}$$

By combining the last two equations we get

$$0 \leq S(D_N, D_{N+1}) \leq \frac{1}{a_N} + \log\left(1 - \frac{1}{a_N + b_N}\right). \tag{35}$$

Asymptotically we have $a_N \approx a_0 + pN$ and $b_N \approx b_0 + (1 - p)N$ (this is exactly true in expectation). Therefore, by taking the first order expansion of $\log(1 - u)$ and substituting these approximate values, we see that asymptotically the bound gives

$$\frac{1}{a_N} + \log\left(1 - \frac{1}{a_N + b_N}\right) \approx \frac{1 - p}{pN}. \tag{36}$$

Thus surprise decreases in time with the number of examples as $1/N$ (Fig. 1). A similar calculation can be done for the Poisson model.

4.2. Learning curves: Continuous data

In the case of continuous Gaussian data with, for instance, known variance σ^2, the learner starts with a Gaussian prior $G_0(\mu_0, \sigma_0^2)$ on the mean. With each example d_N, the learner updates its Gaussian prior $G_N(\mu_N, \sigma_N^2)$ into a Gaussian posterior $G_{N+1}(\mu_{N+1}, \sigma_{N+1}^2)$ with

$$\mu_{N+1} = \frac{\frac{\mu_N}{\sigma_N^2} + \frac{d_{N+1}}{\sigma^2}}{\frac{1}{\sigma_N^2} + \frac{1}{\sigma^2}} \quad \text{and} \quad \frac{1}{\sigma_{N+1}^2} = \frac{1}{\sigma_N^2} + \frac{1}{\sigma^2}. \tag{37}$$

From Eq. (18), the relative surprise is

$$S(G_N, G_{N+1}) = \log \frac{\sigma}{\sqrt{\sigma^2 + \sigma_n^2}} + \frac{\sigma_N^2}{2\sigma^2}\left(1 + \frac{(\mu_N - d_{N+1})^2}{\sigma^2 + \sigma_N^2}\right). \tag{38}$$

Asymptotically

$$E[S(G_N, G_{N+1})] \leq \frac{\sigma_N^2}{2\sigma^2}. \tag{39}$$

From Eq. (17), we have $\frac{1}{\sigma_{N+1}^2} = \frac{1}{\sigma_0^2} + \frac{(N+1)}{\sigma^2}$, or $\sigma_{N+1}^2 = \frac{\sigma_0^2 \sigma^2}{\sigma^2 + (N+1)\sigma_0^2}$, which asymptotically behaves like σ^2/N.

Combining this asymptotic form with Eq. (39), we see that in this case surprise can be expected to decrease as $1/2N$, again proportionally to the inverse of the number of data points.

Similar calculations can be done for the general exponential case (Eqs. (30) and (31)) by noticing that, as $N \to \infty$, in a stationary environment $T_i(D) \approx N\bar{t}_i(d)$, where $\bar{t}_i(d)$ is the average value of t_i.

5. Relations of Bayesian surprise to other theories of information and surprise

5.1. Theories of information

Several theories have been proposed over the years to try to capture the concept of information and entropy (Aczel & Daroczy, 1975; Blahut, 1987; Cover & Thomas, 1991; Jumarie, 1990; McEliece, 1977; Renyi, 1961; Shannon, 1948) and not all of them can be reviewed here. Shannon's theory has been by far the most successful one, and many of the other theories that have been proposed (Aczel & Daroczy, 1975; Renyi, 1961) can be viewed as variations on Shannon's definition. Thus for conciseness here we focus on the relationship of Bayesian surprise to Shannon's definition and then separately on the relationship of Bayesian surprise to other specific definitions of surprise found in the literature.

Shannon's theory of communication defines the information contained in D at the level of an individual model M by

$$I(D, M) = -\log P(D|M) \tag{40}$$

with the corresponding entropy

$$I(\mathcal{D}, M) = H(P(D|M)) = -\int_{\mathcal{D}} P(D|M) \log P(D|M) \mathrm{d}D. \tag{41}$$

This entropy corresponds to an integral over data, whereas Bayesian surprise corresponds to an integral over models hence at this level surprise and information are dual facets of the data.

Shannon's theory can also be applied at the level of the model class $\mathcal{M}$. In this case the information carried by D is

$$I(D) = I(D, \mathcal{M}) = -\log P(D|\mathcal{M}) = -\log P(D) \tag{42}$$

where $P(D) = P(D|\mathcal{M}) = \int_{\mathcal{M}} P(D|M)P(M)\mathrm{d}M$ is also called the *evidence* and plays a key role in Bayesian analysis and model class comparison.

The corresponding entropy is given by

$$I(\mathcal{D}, \mathcal{M}) = -\int_{\mathcal{D}} P(D|\mathcal{M}) \log P(D|\mathcal{M}) \mathrm{d}D. \tag{43}$$

For a fixed data set D, the surprise is

$$S(D, \mathcal{M}) = -I(D, \mathcal{M}) + \int_{\mathcal{M}} P(M) I(D, M) \mathrm{d}M \tag{44}$$

and therefore it can also be viewed as the difference between the average Shannon's information per model, taken with respect to the prior, and the Shannon's information based on the evidence.

If we integrate the surprise with respect to the evidence

$$\int_{\mathcal{D}} P(D) S(D, \mathcal{M}) \mathrm{d}D = \int_{\mathcal{D}, \mathcal{M}} P(D)P(M) \log \frac{P(D)P(M)}{P(D, M)} \mathrm{d}D\mathrm{d}M \tag{45}$$

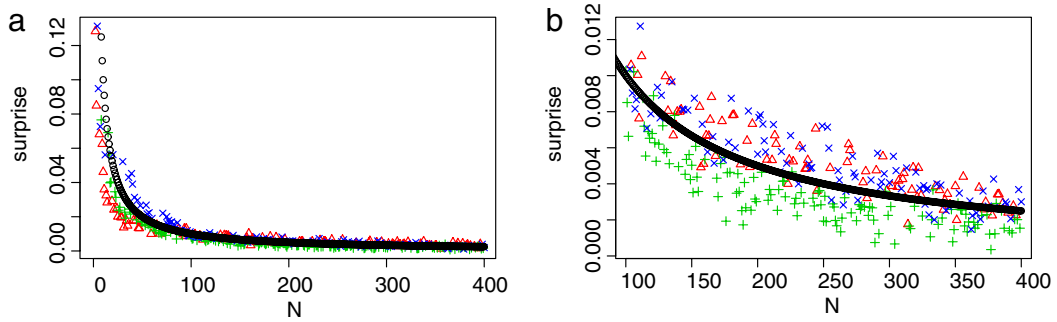

Fig. 1. Simulation results corresponding to flips of a 3-sided (red, blue, green) die with a corresponding multinomial learning model. Curves are derived using 400 random samples, drawn from the distribution red = 0.3, blue = 0.3, and green = 0.4. The x axis correspond to learning iterations. The y axis corresponds to surprise. As predicted by the theory, during Bayesian learning, surprise decreases on average as $1/N$ (black curve) as learning progresses. (a) curve corresponding to epochs 1–400; (b) magnified view corresponding to epochs 100–400. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

we get the Kullback–Liebler divergence $K(P(D)P(M), P(D, M))$, which is the permuted version of the mutual information MI between $\mathcal{D}$ and $\mathcal{M}$: $MI(\mathcal{D}, \mathcal{M}) = K(P(D, M), P(D)P(M))$. If surprise is defined by $K(P(M|D), P(M))$ then the integral of the surprise is equal to the mutual information between data and models. Note in contrast that the integral over models of the per-model entropy (Eq. (41)) is *not* equal to the mutual information, but related to it by a convex inequality, as shown by Eq. (44).

In short, Bayesian surprise measures a facet of information that is different and complementary to Shannon's definition.

5.2. Theories of surprise

The concepts of "surprise" and "surprising event" have also been raised multiple times in the statistical literature (Bartlett, 1952; Evans, 1997; Good, 1956; Kvalseth, 1987; Redheffer, 1951; Weaver, 1948).

One simple approach corresponds to outlier detection theory, whereby surprising events are defined as rare events, i.e. events having a low probability. Such a definition of course is closely related to Shannon's theory of information since, by definition, rare events ($P(D)$ small) have high Shannon's information ($I(D) = -\log P(D)$ large). While it is easy to see that in many cases, Shannon's $I(D)$ and the Bayesian surprise $S(D)$ are closely related, there exist also specific situations where these two approaches provide clearly distinct answers, and Bayesian surprise matches intuition better. We illustrate here two somewhat extreme classes of examples corresponding to high Shannon's information and low surprise, and vice versa.

Many bits with few wows: The most simple example is obtained when $\mathcal{M}$ contains a single model M. The prior is necessarily given by $P(M) = 1$, and the posterior distribution is always equal to the prior. Thus if there is data D satisfying $P(D) = \epsilon \ll 1$, the Shannon's information $\log \epsilon$ can be arbitrarily large, whereas the surprise is always zero. For a more complex and instructive example, let $\mathcal{M} = \{M_1, \ldots, M_N\}$ with a uniform prior $P(M_i) = 1/N$ for every i. Assume that for each model M_i, $P(D|M_i) = \epsilon$, and hence $P(D) = \epsilon$; that is, a datum is observed which is unlikely for any of the models. By Bayes theorem, we have immediately $P(M_i|D) = P(M_i) = 1/N$. Thus in this case Shannon's information $-\log \epsilon$ grows to infinity as we decrease the value of ϵ. On the other hand, the prior and the posterior distributions being identical, the surprise is zero. Thus while the number of wows is zero, the number of bits grows to infinity as $-\log \epsilon$. In this case, although D is a strong outlier, D is a false positive, in the sense that it carries no useful information for discriminating between the alternative hypotheses M_i. Therefore D carries no surprise as its observation leaves the observer's expectations unaffected.

Few bits with many wows: Conversely, consider $\mathcal{M} = \{M_1, \ldots, M_N\}$ with a non-uniform prior given by $P(M_1) = a$ and $P(M_i) = (1 - a)/(N - 1)$ for $i = 2, \ldots, N$. Consider data D with the likelihood $P(D|M_1) = (1 - a)/(N - 1)$ and $P(D|M_i) = a$ for $i = 2, \ldots, N$. A simple calculation shows that $P(D) = a(1 - a)N/(N - 1)$ while the posterior distribution is uniform and given by $P(M_i) = 1/N$ for any i. Thus for large N, the Shannon's information converges to the constant value $I(D) = -\log[a(a - 1)]$ bits, determined by the parameter a. For instance, if $a = 0.5$ the Shannon's information converges to 2 bits. The surprise, however, is given by $S(D) = a \log Na + (1 - a) \log[(1 - a)N/(N - 1)]$ which, for large values of N, converges to $a \log N + H(a)$. Thus while the number of bits is finite, the number of wows grows to infinity as $a \log N$. Data with these properties would go undetected by standard outlier theory, but would be picked up by surprise since it is associated with a significant change from prior to posterior distribution.

Note that examples where D carries few bits but many wows do not require the number N of models to go to infinity. Consider the case $N = 2$ with $P(M_1) = a$ and $P(M_2) = 1 - a$. Assume that $P(D|M_1) = b$ and $P(D|M_2) = c$. We then have $P(D) = ab + (1-a)c$, $P(M_1|D) = ba/P(D)$, and $P(M_2|D) = (1 - a)c/P(D)$. The surprise is equal to $\log[ab + (1 - a)c] - a \log b - (1 - a) \log c$. By letting either $b \to 0$ or $c \to 0$, but not both, we can easily achieve a diverging amount of surprise in combination with a finite amount of Shannon's information.

In any case, in the outlier detection approach, if all possible events or datasets have very low probability, then they are all very "surprising", which is not very useful in practice. Thus it is clear that whether an event or data set is surprising or not cannot be decided on its probability alone. As a minimum, the probabilities of the other events must also be taken into consideration. Thus another approach, introduced in Weaver (1948) and further developed in, for instance, (Good, 1956; Redheffer, 1951), tries to compare the probability of an event to the probabilities of all the other possible events by using a "surprise index". Weaver (1948) considers an experiment with n possible outcomes, with probabilities $p_1, \ldots, p_n$ and defines the Surprise Index by

$$SI = \frac{E(p)}{p_i} = \frac{\sum_{i=1}^{n} p_i^2}{p_i}. \tag{46}$$

The SI measures "whether the probability realized, namely, p_i is small as compared with the probability that one can expect on the average to realize, namely, $E(p)$. If this ratio is small and SI correspondingly large, then one has a right to be surprised" (Weaver, 1948). Unlike simple outlier detection, the surprise index does consider an event in the context of other events. However, the surprise index does not consider explicitly prior and posterior

distributions and is obviously quite different from the concept of Bayesian surprise described in this paper.

Perhaps closest in spirit to our work, but still different and derived completely independently, is the work of Evans (Evans, 1997; Evans, Guttman, & Swartz, 2006) which takes a Bayesian perspective and considers prior and posterior distributions and their ratios. More precisely, consider a family of models parameterized by θ, and a function $T(\theta)$ with a set of possible values t_i. Evans proposes to introduce a total ordering on the t_i and a surprise inference principle by considering that t_1 is strictly preferred to t_2 if "the relative increase in belief for t_1, from a priori to a posteriori, is greater than the corresponding increase for t_2". In turn, this preference ordering is used to determine inferences and applied to estimation, hypothesis testing, and model checking procedures (Evans, 1997). This statistical work, however, does not take an explicit information theoretic perspective, and does not define surprise as the relative entropy between the prior and posterior distributions.

6. A neural network implementation of surprise for computer vision

Here we describe a neural network architecture for processing and computing surprise over image and video data in a bottom-up fashion. The architecture is inspired by the neurobiology of early visual processing in the primate brain (Itti & Baldi, 2005, 2006, 2009; Itti & Koch, 2001). In this section we first describe the low-level, Hubel and Wiesel-like, visual feature detection front-end of the proposed system followed by two alternative surprise computation back-ends operating on the feature responses.

6.1. Low-level visual feature extraction

The proposed system employs a relatively mature and standard low-level feature extraction front-end (Itti & Koch, 2000; Itti, Koch, & Niebur, 1998). This front-end analyzes the incoming input images at several spatial resolutions and along several low-level feature dimensions, including color contrast, luminance contrast, oriented edges, and motion energy. In a manner inspired by how early visual processing is organized in the primate brain, the front-end processing thus decomposes the image into a number of sub-bands. Surprise is then computed at the level where the responses from the low-level feature detectors are integrated, as opposed to directly at the pixel level.

A schematic diagram of the system is given in Fig. 2. Input video frames are analyzed in dyadic image pyramids with 9 scales (from scale 0 corresponding to the original image, to scale 8 corresponding to the image reduced by a factor of 256 horizontally and vertically). The pyramids are constructed by iteratively filtering and decimating the input image. In the implementation used here, pyramids are computed for the following low-level visual features thought to guide human attention (Wolfe & Horowitz, 2004): (1) luminance; (2) red–green color opponency; (3) blue–yellow color opponency; (4) four oriented edge filters (using Gabor kernels) spanning 180°; (5) luminance flicker (as computed from the difference between the previous image and the current one); and (6) four directions of motion spanning 360°. Additional details about the implementation of these image pyramids have been published previously (Itti, Dhavale, & Pighin, 2003; Itti et al., 1998). The final feature output is in the form of "feature maps" which are obtained by taking across-scale differences between pairs of levels within each feature pyramid. These differences coarsely approximate center-surround contrast enhancement mechanisms found in the early stages of biological visual processing (Grossberg & Raizada, 2000; Hubel & Wiesel, 1962; Suder & Worgotter, 2000). Center-surround differences are computed for the following scale pairs: 2–5, 2–6, 3–6, 3–7, 4–7, 4–8. All feature maps are then resampled to scale 4 (where the final combined surprise map is later computed); for 640 × 480 input videos, these maps thus have 40×30 pixels. In total, 72 such feature maps are computed (6 for luminance, 12 for color opponencies, 24 for oriented Gabor edges, 6 for flicker, and 24 for motion). Fig. 2c shows all the feature maps computed for the example input image shown.

6.2. Surprise computation in feature space

Surprise is computed for every pixel in each of the 72 center-surround feature maps. The underlying motivation is that simulated neurons in the feature maps may establish some very simple beliefs about the world as seen through their spatially- and feature-selective center-surround receptive fields. For instance, a neuron sensitive to red/green opponent contrast at a given location and scale may accumulate over time beliefs about the amount of red/green contrast present in the small portion of the world that is captured by the neuron's receptive field. When new data is observed with each new incoming video frame, the beliefs established thus far are used as prior, and Bayes' rule is applied to compute the posterior. The posterior at one video frame then becomes the prior for the next video frame. Using conjugate priors facilitates this process by ensuring that the posterior has the same functional form as the prior. The current implementation derives prior distributions at time t entirely from past inputs combined through Bayesian learning; however, the theory does not limit what may influence the prior distributions. Other sources such as top-down knowledge, behavioral states, or individual preferences could also influence the prior within the same general framework.

Here we explore two model classes to implement surprise: Gaussian (with a Gaussian conjugate prior) which is formally simple and parallels background adaptation techniques used in computer vision (Grimson, Stauffer, Romano, & Lee, 1998), and Poisson (with a Gamma conjugate prior) which may more accurately model incoming neural spike trains from the low-level feature extraction stages (Softky & Koch, 1993).

To accommodate for changing data and events at multiple temporal scales, we employ a chained cascade of surprise detectors at every pixel in every feature map, where the output of one surprise detector serves as input to the next detector in the cascade. Our implementation uses 5 such cascaded feature detectors at every pixel and for every feature. The first (fastest) is updated with feature map data from the low-level feature computations, and detector $i + 1$ samples from i, so that time constants increase exponentially with i. In total, the system thus comprises $72[\text{maps}] \times (40 \times 30)[\text{pixels}] \times 5[\text{timescales}] = 432{,}000$ surprise detectors.

Finally, to account not only for temporally surprising events (e.g., sudden appearance of an object) but also for spatially surprising items (e.g., a red object among many green objects), we compute surprise both locally over time, and spatially in an instantaneous manner. For local temporal surprise, a single neuron in one of the feature maps is considered, and the prior is established over time from the observations received to date in the receptive field of that one neuron. For spatial surprise, a single neuron is also considered, but its prior is now derived from the compound instantaneous activity of the surrounding neurons in the map, at the next faster time scale. For every video frame at time t, location (x, y), feature f, and time scale i, a neighborhood distribution of models is computed as the weighted combination of distributions from the next-faster local models, over a large neighborhood with two-dimensional Difference-of-Gaussians profile ($\sigma_+ = 20$ and $\sigma_- = 3$ feature map pixels). As new data arrives, spatial surprise is the KL divergence between the prior neighborhood distribution and the posterior after update by local samples from

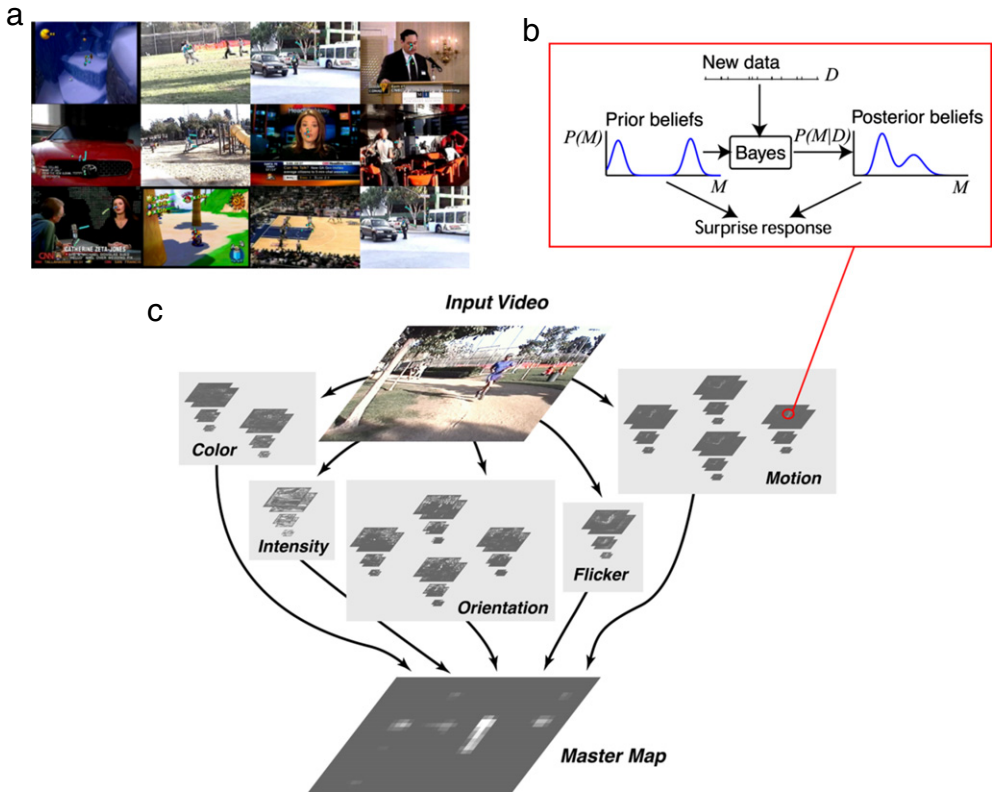

Fig. 2. (a) Sample frames of video clips to be processed by the model and for which human eye movement recordings are available (see next section). (b) Computation of surprise at the single-neuron level, each time new data is received from a new video frame. (c) Architecture of the full computational system which analyzes video frames along the dimensions of color, intensity, orientation, flicker, and motion, computing surprise at multiple spatial and temporal scales within each of these feature channels. The surprise output from all feature channels finally gives rise to the master surprise maps. Higher (brighter) values in this map represent the system's prediction of where the strongest attractors of attention currently are in the video inputs (in the example shown, the man running towards the camera is most surprising).

the neighborhood's center. For example, consider a neuron that observes a locally red image patch surrounded by a large green background. That neuron's neighbors would contribute a spatial prior that strongly suggests that the world is green. However, when that prior is combined with the local data received by the neuron of interest, which indicates that the world is red, a large difference between prior and posterior arises, and consequently a large spatial surprise (see Fig. 3 for pseudo-code).

Our theory does not constrain how temporal and spatial surprises may combine. In previous work (Itti & Baldi, 2005), we addressed this issue by turning to empirical single-unit recordings of complex cells in striate cortex of anesthetized monkey (Müller, Metha, Krauskopf, & Lennie, 1999). From fitting the neural data, total surprise S is given by:

$$S = \left[S_T + \frac{S_S}{20} \right]^{\frac{1}{3}} \tag{47}$$

where S_T is the temporal and S_S the spatial surprise. This formulation resulted from a least-squares fit of a function of the form $S = [a_1 S_T + a_2 S_S + a_3 S_T S_S]^{a_4}$ to the neural data. We further posit that surprise combines multiplicatively across time scales, such that an event is surprising only if at all relevant time scales, allowing the model to learn periodic stimuli of various frequencies. One caveat with this approach is that stimuli which may fully adapt the surprise detectors at one time scale and yield zero surprise at that scale may effectively zero out potentially surprising events at other time scales. To address this, we introduce below an additional step just before the prior is updated, whereby the variance of the prior is slightly relaxed (increased) prior to the arrival of every video frame. A parameter ζ regulates the amount of relaxation. With $0 < \zeta < 1$, the variance of the prior will not settle, even if, e.g., the data is stationary, resulting in surprise values that are always non-zero. We finally assume that surprise sums across features, such that a location may be surprising by its color, motion, or other. It is interesting to note that other alternatives may be more desirable, e.g., a max operation across features (Li, 2002; Zhaoping & May, 2007). However, for the datasets evaluated here, this alternative yields lower ordinal dominance scores (Fig. 5).

The sum is then passed through a saturating sigmoidal nonlinearity to enforce plausible neuronal firing dynamics and yields the final master map used for comparison with human gaze behavior.

6.2.1. Gaussian data with Gaussian prior

With Gaussian models, the data from every low-level feature detector at every location in every video frame is assumed to have a Gaussian distribution with mean given by the feature detector's response to the current frame, and fixed variance which approximates to the observation noise at the feature detection stage (which, in turns, reflects the sensor's noise and possibly the neural noise inherent to the feature computation process).

Thus, for each of the 72 feature types (e.g., luminance, red–green opponency, etc.), video frame, spatial center-surround feature scale, 2D image location in the rescaled feature map, and temporal scale, we here use the unknown mean/known variance formulation of Section 3.3, Eq. (17). With the addition of the prior relaxation term $0 < \zeta < 1$ (which here simply divides the prior's variance σ_1^2, we obtain:

$$\mu_2 = \frac{\frac{\mu_1 \zeta}{\sigma_1^2} + \frac{N\bar{m}}{\sigma^2}}{\frac{\zeta}{\sigma_1^2} + \frac{N}{\sigma^2}} \quad \text{and} \quad \frac{1}{\sigma_2^2} = \frac{\zeta}{\sigma_1^2} + \frac{N}{\sigma^2}. \tag{48}$$

```
function S = SurpriseMap(Data)

{
input: Data <- 2D array of feature detector outputs for current video frame

static variables: Models[...] vector of 2D arrays of models

external function: surprise(Model, Data) updates 2D array Model of model
  prior probabilities into posterior probabilities using Data, and
  compute and return 2D array of surprise values

local variable: S <- 2D array of combined surprise values,
  initialized to ones

loop i over surprise time scales
  local variables: SL <- empty 2D array for local temporal surprise values
                   SS <- empty 2D array for spatial surprise values
                   Neigh <- compute array of neighborhood models from
                            Models[i-1] (or from Data if i == 0)

  SL <- surprise(Models[i], Data)   // local temporal surprise
  SS <- surprise(Neigh, Data)       // spatial surprise

  S <- S * (SL + SS/20)^(1/3)

  // the updated models provide input to the next time scale:
  Data <- Models[i]
end loop

return S
}
```

Fig. 3. Simplified algorithm to compute surprise over space and time at multiple scales. This algorithm is applied to each of the 72 feature maps. The incoming data is the array of raw feature detector values for the feature map's feature type and center-surround spatial scale.

We assume that $N = 1$ data samples are received for every new video frame, the data sample value (and $\bar{m}$ since $N = 1$) is given by the feature detector's response to the current video frame, and the variance σ^2 of the data is fixed and reflects noise in the sensor and the early stages of processing (we use $\bar{s} = 5$ given RGB pixel values in the $[0 \dots 255]^3$ range). Our simulations use $\zeta = 0.7$. Surprise is computed exactly per Eq. (18) (without the approximation for large N).

6.2.2. Poisson data with Gamma prior

In a somewhat more neurally-plausible implementation, we model data received from feature map f at location (x, y) and time t as Poisson distributions $M(\lambda)$ (which describe well cortical pyramidal cell firing statistics (Softky & Koch, 1993)), parameterized by firing rate $\lambda \geq 0$. λ is trivially estimated over the duration of each video frame as simply the feature detector's response $\bar{\lambda} = f(x, y, t)$. In contrast with the Gaussian case, here the variance of the sample is not arbitrarily fixed but directly determined by the assumed Poisson nature of the data samples, and thus is equal to the mean $\bar{\lambda}$.

As detailed in Section 3.2, the prior $P(M)$ satisfying the conjugate prior property on Poisson data is the Gamma probability density, Eq. (13). With the addition of the ζ prior relaxation term (which here again divides the prior's variance, by multiplying both a_1 and b_1), we slightly modify Eq. (14) and obtain:

$$a_2 = \zeta a_1 + N\bar{m} \quad \text{and} \quad b_2 = \zeta b_1 + N. \tag{49}$$

We here again assume that $N = 1$ data samples are received at every video frame. Our simulations here again use $\zeta = 0.7$. Surprise is computed exactly using Eq. (15).

7. An application of surprise to psychophysics and eye movements

Developing an ability to rapidly detect surprising events is crucial in allowing living systems to quickly identify potential predators, preys, or mates and in ensuring survival. It is reasonable to postulate that surprising events ought to attract attention mechanisms in living systems, and that the same principle may be useful in the design of artificial systems, for instance in computer vision and surveillance. Indeed, it has been noted that events previously described as novel or salient tend to attract attention (Ranganath & Rainer, 2003). To test the surprise theory, we here present an application to finding surprising objects and events in natural video streams. Using eye-tracking experiments with human subjects, we quantitatively evaluate the extent to which surprising visual events occurring in natural video stimuli may indeed capture the attention and gaze of human observers. The application presented here extends previous similar experimentations (Itti & Baldi, 2005, 2006, 2009) by adding a new dataset, comparing two different model classes, Gaussian data with a Gaussian prior, and Poisson data with a Gamma prior, and introducing a new temporal analysis which demonstrates a strong

time contingency between the onset of a surprising event in the stimuli and the initiation of a human eye movement towards that event.

7.1. Subjects, stimuli and gaze recording methods

Our experiments use a publicly available human eye movement dataset from the NSF-CRCNS data sharing project (crcns.org). This dataset contains two components.

In the first component, we recorded eye movements from eight naïve observers. Each watched a subset from 50 videoclips totaling over 25 min of playtime ("original" dataset). Clips comprised outdoors daytime and nighttime scenes of crowded environments, video games, and television broadcast including news, sports, and commercials. The right-eye position was tracked with a 240 Hz video-based device (see Itti and Baldi (2006) for additional methodological details). To maintain interest, observers were instructed to follow the stimuli's main actors and actions, as simple questions would be asked to them at the end of the experiment to test their understanding of the contents of the clips. Here we only retain those clips for which gaze recordings from at least 4 observers were available (to allow for the establishment of an upper bound inter-observer correlation performance, below). Two hundred calibrated eye movement traces (10,192 gaze shifts or "saccades") were analyzed, corresponding to four distinct observers for each of the 50 clips.

In the second component, the video clips were cut into 1–3 s short "clippets" which were then re-assembled in random order into an "MTV-style" set of clips (Carmi & Itti, 2006). Another set of 8 observers watched the clips and we here again only retain clips for which recordings are available for at least 4 observers. The goal of the random shuffling of clippets was to abolish some of the long-term cognitive influences on attention and gaze, so that observer's gaze allocation would be more strongly short-term. Here, this dataset is of particular interest because it will help us gauge the extent to which past events beyond a few seconds may influence our modeled low-level vision priors, surprise, and human gaze. In total, 6648 saccades were analyzed for the MTV-style dataset.

Informed consent was obtained from all subjects prior to the experiments. Each subject watched a subset of the collection of video clips, so that eye movement traces from four distinct subjects were obtained for each clip. Video clips 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° × 21°). The clips comprised between 164 and 2814 frames or 5.5 s to 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).

The right-eye position was tracked at 240 Hz using a video-based device (ISCAN RK-464), which robustly estimates the 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°.

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. The random sampling process was repeated 100 times.

7.2. Gaze prediction results

We evaluate in Fig. 4 four different attention systems that predict human saccades: one using surprise with Gaussian data and prior (Section 6.2.1), one using surprise with Poisson data and Gamma prior (Section 6.2.2), a simple baseline control system which simply computes local pixel variance in small image patches, and a human inter-observer system—all further described below. In addition, in Fig. 5 we compare these systems with additional previously published systems and with variations on our surprise-based systems.

The Gaussian/Gaussian and Poisson/Gamma surprise systems are as described above. The Variance system simply computes the local variance of pixel luminance within 16 × 16 image patches. The resulting variance map has been suggested to already predict human gaze above chance (Reinagel & Zador, 1999); hence we use this very simple system as a baseline or lower bound. The inter-observer system is built by plotting a Gaussian blob with $\sigma = 3$ master map pixels (4.5°), continuously 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 values on this map would hence be present only at the locations currently gazed at by some human observer. This system allows us to establish an upper bound for how well the other systems might be expected to predict human gaze.

To characterize image regions selected by our observers, we process the video clips through the different systems. Each system outputs a topographic dynamic master response map, assigning in real-time a response value to every input location. A good master map should highlight, more than expected by chance, locations gazed to by our human observers. Hence to score and compare each system, we sample, at the onset of every human saccade made, the master map activity around the saccade's future endpoint, and around a uniformly random endpoint. Random sampling is repeated 100 times to evaluate variability. To quantify the extent to which humans may be attracted towards hotspots in the master maps, we use ordinal dominance analysis (Bamber, 1975). To this end, we first normalize master map values sampled at human and random saccade endpoints by the maximum activity in the master map at the time of the saccade. For each system, histograms of master map values at human and random saccade endpoints are then created (Fig. 4). In a manner similar to computing a Receiver Operating Characteristic (ROC) curve, we then sweep a threshold from 0 to 1; for each threshold value we count the percentage of eye positions and random positions which land above the threshold ("hits"). The ordinal dominance curve is then created similarly to an ROC curve, with the difference that it plots human hits vs. random hits. This curve indicates how well a simple binary threshold is able to discriminate the signal (master map values at human gaze locations) from noise (values at random locations). We finally score each system by computing the area under the curve (AUC) from the ordinal dominance curve. The process is repeated 100 times for the 100 random samples, which allows us to attach a confidence estimate to each AUC value.

An AUC value of 0.5 would indicate a system which is at chance in predicting where human observers looked. Our inter-observer system yields AUC scores of 0.805 ± 0.002 for the first ("original") dataset and 0.827 ± 0.002 for the second ("MTV-style") dataset. The surprise and variance systems are hence expected to score between 0.5 and 0.8, with higher scores indicating a better ability to predict human gaze. Scores are shown in Fig. 4. All three systems perform significantly above the chance level (AUC = 0.5; t-test, $p < 10^{-10}$ or better). Furthermore, both variants of the surprise-based system perform significantly better than the much simpler variance-based system (t-tests on the respective AUC scores, $p < 10^{-10}$ or better). The Poisson/Gamma surprise system exhibits

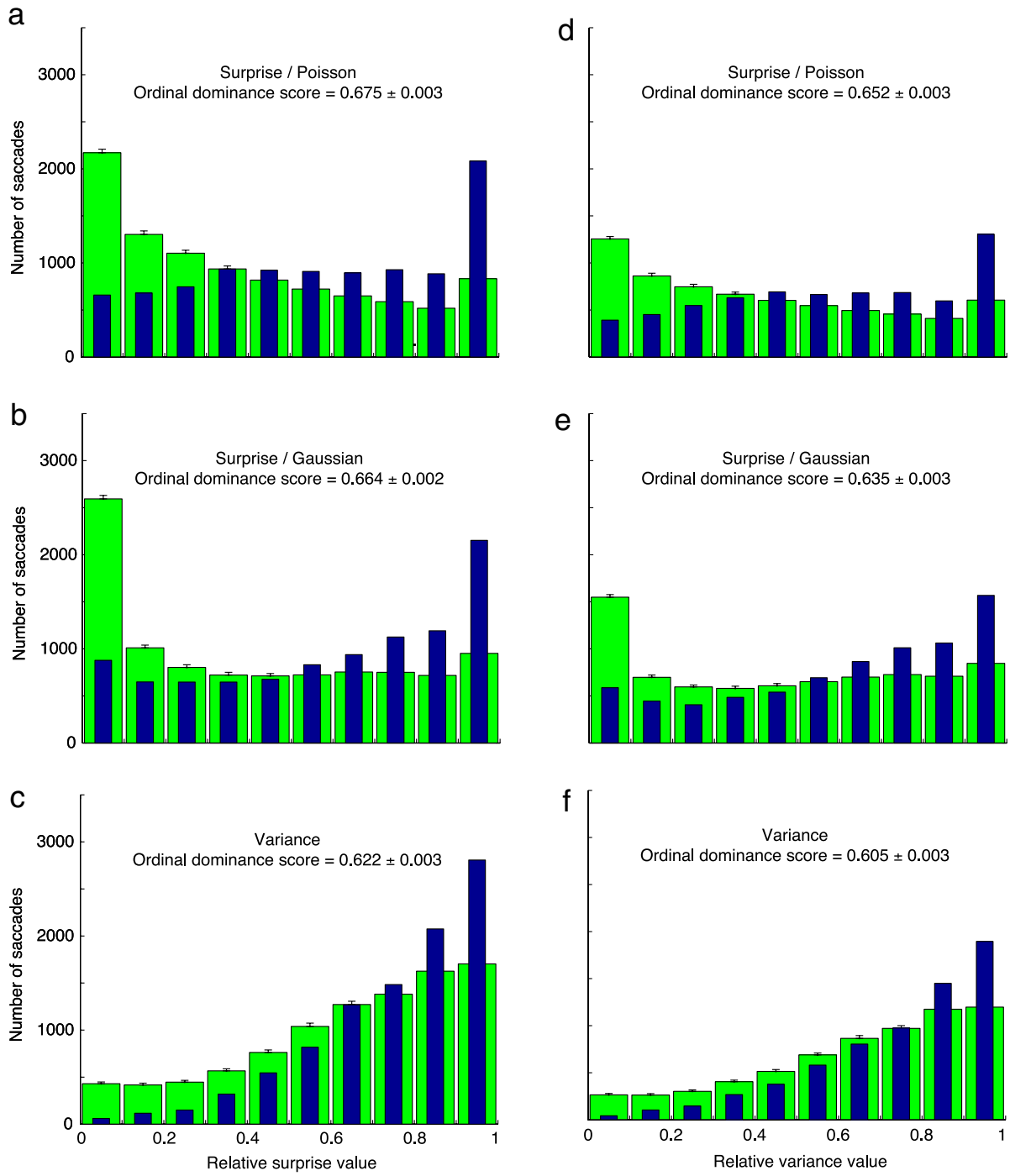

Fig. 4. Histograms of master map values at saccade endpoints for random (wide green bars) and humans (narrow blue bars), for the "original" (a, b, c) and "MTV-style" (d, e, f) datasets. Human histograms particularly differ from the random ones in that humans gaze towards low master map values (leftmost bins in each histogram) less often than expected by chance, while they gaze towards high mastermap values (rightmost bins) more often than expected by chance. AUC scores indicate significantly different performance levels, and a strict ranking of chance < variance < Gaussian/Gaaussian surprise < Poisson/Gamma surprise < inter-observer.

a small but significant advantage over the Gaussian/Gaussian surprise system in these experiments. These two surprise systems score about half-way between chance and inter-observer.

Fig. 5 shows an example video frame and corresponding master maps for the computational systems studied here, as well as for a few systems evaluated previously (Itti & Baldi, 2009), and variations on the surprise systems described below. In particular, the Michelson contrast system, as in Mannan, Ruddock, and Wooding (1996), is an interesting alternative to our local variance system. However, we find that in these datasets it actually performs slightly below chance level (corresponding to a score of 0.5), indicating that observers tended to preferentially look towards locations with a lower contrast than expected by chance. This is in agreement with the original results of Mannan et al., who conclude that a number of local features tested in their experiments (including Michelson contrast, edge density measures, and others) is a poor predictor of human fixations. The Entropy and DCT-based information systems are very simplified measures of information in local image patches, and have been previously proposed as gaze predictors (Itti & Baldi, 2009; Itti et al., 1998; Privitera & Stark, 2000). We find that they score above chance, but below surprise. Finally, we evaluated the hypothesis

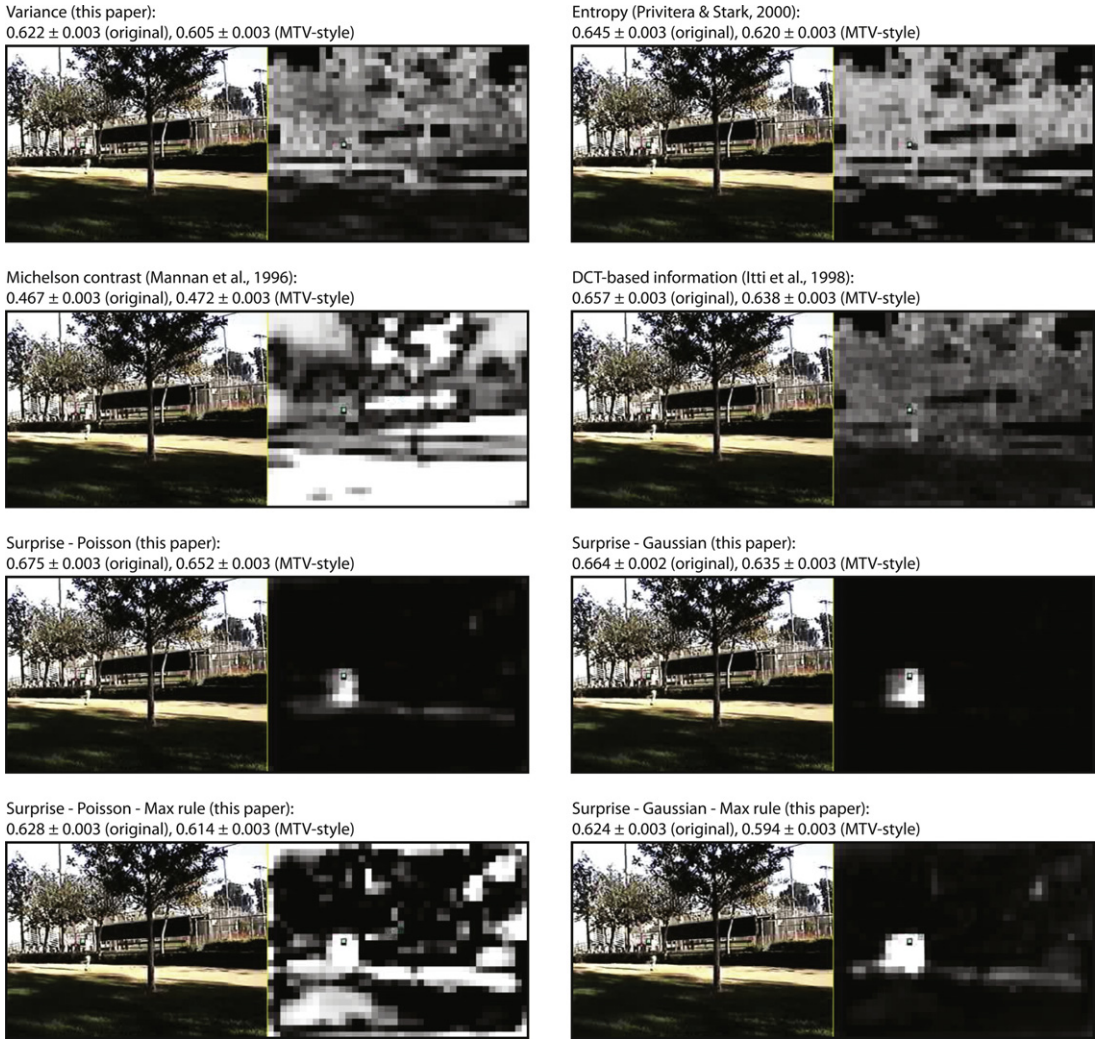

Fig. 5. Example maps generated by the computational systems tested. One video frame is shown (clip beverly08, frame 127) with the corresponding master map for each model. The current eye position of one human observer is indicated by the small cyan square (on the person running in the video clip). Ordinal dominance scores are indicated for each system, for both the original and MTV-style dataset. Higher scores indicate better systems, i.e. systems which predict high master map values near the locations selected by human gazes and low map values everywhere else.

of Zhaoping and colleagues that salience signals may combine across different feature channels according to a maximum rule rather than a sum rule (Li, 2002; Zhaoping & May, 2007). To this end, we implemented variants of the surprise systems where color, intensity, orientation, flicker and motion sources of surprise combine with a max rule. With the datasets evaluated here, this alternative yields lower AUC scores (Fig. 5). The reason for the lower scores when taking the maximum across features is that this often yields a lower gaze target to clutter signal-to-noise ratios. These intriguing results obtained with our free-viewing datasets should be evaluated further with future experiments aimed more directly at addressing the max vs. sum question.

To further investigate the extent to which events detected as surprising by our systems might attract human gaze, we recorded, at the landing location of every saccade, the history of master map values for up to 1000 ms preceding the initiation of the saccade of interest (Fig. 6). In both datasets, a sudden surge of surprise is seen at saccade landing points shortly before saccades are initiated. A similar but weaker surge is observed also for the Variance system, indicating that the sudden appearance of textured objects might also have attracted saccades. However, the surge for the variance system is significantly weaker than for any of the surprise systems. In the MTV-style experiments, a surge of surprise is also observed for random saccades around 500 ms prior to human saccades; this reflects the global surge of surprise which is observed at every abrupt jump-cut when the scene changes from one 1–3 s clippet to the next. The surge in the random saccade data suggests that humans tended to execute saccades in a manner that was somewhat time-locked to jump-cuts in the stimuli. This is also reflected by the larger surge observed for human saccades in the MTV-style compared to the original dataset.

8. Discussion

The Bayes theorem is the most fundamental theorem of learning and adaptation, quantifying how the prior distribution over the space of models or hypotheses ought to be revised into a posterior distribution as data is collected. Accordingly, the distance between the prior and posterior distributions is also bound to be a fundamental quantity, that may have escaped systematic attention. Here we have defined this quantity – surprise – and studied its properties systematically.

Surprise is different from Shannon's entropy, which it complements. Surprise is also different from several other definitions of information that have been proposed (Aczel & Daroczy, 1975) as alternatives to Shannon's entropy. Most alternative definitions of

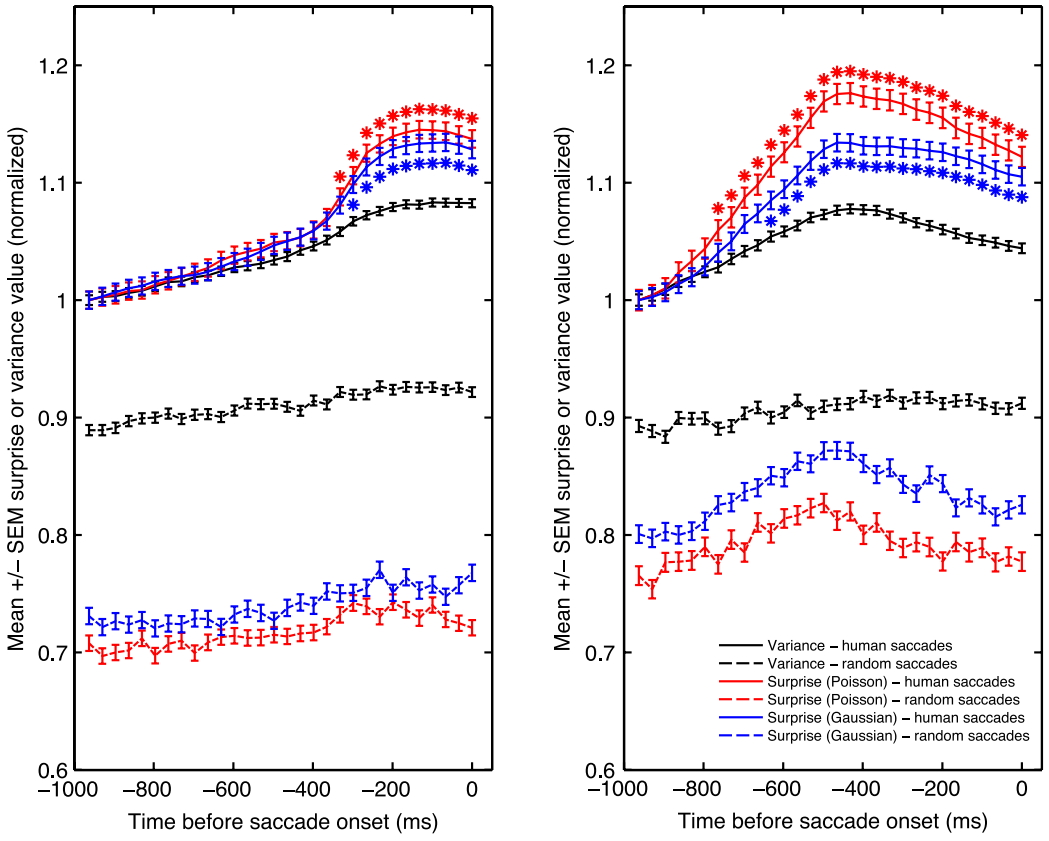

Fig. 6. History of master map values at locations selected by human saccades, for up to 1000 ms prior to the initiation of each saccade, for the "original" (left) and "MTV-style" (right) datasets. Curves are normalized in pairs (human saccades, solid lines, and random saccades, dashed lines) by the surprise value for human saccades at −1000 ms. This allows an easy comparison between the curves for the different systems. Stars indicate the time points where the surprise-based systems performed significantly differently from the variance-based system (t-tests, $p < 0.05$ after Bonferroni correction).

entropy, such as Rényi's entropies, are actually algebraic variations on Shannon's definition rather than conceptually different approaches.

To measure the effect of data on the observer's prior and posterior distributions, one could envision using the difference between the entropy of the prior and the entropy of the posterior. However, such a difference would only quantify the difference in uncertainty between the prior and posterior distributions. Unlike surprise which is always positive, such a difference could be either positive or negative and therefore less appealing as a measure. More fundamentally, the posterior could, for instance, be very different from the prior, but retain a similar level of entropy. Thus data greatly affecting the observer could appear insignificant by this measure.

In many important cases related to the exponential family of distributions, surprise can be computed analytically and efficiently, both in terms of exact and approximate formula. The analytical results presented here could be extended in several directions including non-conjugate and other prior distributions as well as more complex multidimensional distributions (e.g., inverse Wishart). In general, however, the computation of surprise can be expected to require numerical techniques including Monte Carlo methods to approximate integrals over model classes. In this respect, the computation of surprise should benefit from ongoing progress in Markov chain and other Monte Carlo methods, as well as progress in computing power.

The concept of surprise has its own limitations. In particular, it does not capture all the semantic/relevance aspects of data. If, while surfing the web in search of a car to purchase one stumbles on a picture of Marilyn Monroe, the picture may carry a low degree of relevance, a high degree of surprise, and a low-to-high amount of Shannon's information depending on the pixel structure. Thus, relevance, surprise, and Shannon's entropy are three different facets of information that can be present in different combinations. Although there have been several attempts (e.g. Jumarie (1990) and Tishby, Pereira, and Bialek (1999)), defining relevance remains a central open challenge. Surprise, however, appears remarkable for its simplicity and generality which ought to result in its applicability to areas as diverse as learning, data mining and compression, and the design or reverse engineering of natural or artificial sensory systems.

We have only touched upon the connection between surprise and statistical learning theory (Hastie, Tibshirani, & Friedman, 2001; Vapnik, 1995) by showing that surprise decreases as $1/N$ during sequential learning in simple cases. This analysis could be extended to more complex settings, such as artificial neural networks. At higher abstraction levels, informal ideas of novelty and surprise have been proposed that could capture attention and trigger learning (Grossberg, 2000; Ranganath & Rainer, 2003), which may now be formalized in terms of priors and posteriors. Highly surprising data could signal that learning is required and highly unsurprising data could signal that learning is completed, or adaptation no longer necessary.

A surprising training set is a prerequisite for learning. The amount of surprise in training data, however, should not be so excessive as to overwhelm the learning system. Thus information surprise in the training set ought to be calibrated to the capacity of the learning system. Furthermore, when the degree of surprise of the data with respect to the model class becomes low, the data is no longer informative for the given model class. This, however, does not necessarily imply that a good models of the data have been learnt since the model class itself could be unsatisfactory and in need of a complete overhaul. The process by which a learning

system realizes that a model class is unsatisfactory in an alternative free setting – the open-ended aspect of inference – has so far eluded precise formalizations and ought to be the object of future investigations.

As a side note, and to avoid any confusion, it is also worth noting that relative entropy has often been used as a training function for neural networks and other machine learning systems (e.g. Baldi and Brunak (2001) and references therein), but in a completely different sense. In a multinomial classification problem, for instance, model parameters can be adjusted in order to minimize the relative entropy $\sum t_i \log t_i - \sum t_i \log p_i$ aggregated over all training examples. For a single training example, t_i is the 0–1 target representing membership in the i-th class and p_i is the probability of membership in the i-th class computed by the learning system. This relative entropy between the vectors t and p has little to do with the relative entropy between the prior and posterior distributions over the parameters of the model that is used to compute p.

In data mining applications, surprise could be used to systematically detect novelty in areas ranging from surveillance to information retrieval. In data compression applications, surprise could be used to guide dynamical encoding of information, allocating more bits to surprising data. For sensory systems, we have described an application of surprise to computer vision and the analysis of human attentional gaze shifts. Human attention is an exceedingly complex phenomenon, under the control of both bottom up and top down influences. Our results are not intended in any way to prove that human brains compute surprise to control eye movements. It is however encouraging to see that the simple bottom-up version of surprise outperforms other state-of-the-art metrics in predicting gaze shifts and that, in principle, top down influences could be incorporated into the surprise framework, simply by modulating the prior distribution. A number of additional theoretical frameworks and computational models has been proposed to explain attention guidance and eye movements, using information-theoretic principles in very restricted scenarios, such as discriminating shape silhouettes or searching for a known target in a noisy environment (Najemnik & Geisler, 2005; Renninger, Coughlan, Verghese, & Malik, 2005). While interesting, these approaches still need additional development before they can be applied to the CRCNS eye-tracking datasets, or other kinds of "real data", for the purpose of comparison with surprise or other theories. Indeed, such theories and approaches are not yet able to make useful predictions for arbitrary image or video stimuli and observer tasks. This is a fundamental difference between these very interesting but more specialized theories and our new approach: Our surprise theory and associated computational model are capable of making predictions (good or bad) for any set of image or video stimuli and any set of associated eye movement traces (human or other) acquired under any observer task.

Finally, in sensory systems and beyond the visual attention application studied here, surprise may be computed on auditory, olfactory, gustative, somatosensory, or other features, using exactly the same definition. Surprise may be computed at different temporal and spatial scales, and different levels of abstraction. Indeed, detecting surprise in neuronal spike trains or other data streams is a very general operation that does not require understanding the "meaning" carried by the data and therefore may be suitable for learning in deep architectures and other self-organization processes.

Acknowledgements

The work of PB is supported by a Laurel Wilkening Faculty Innovation Award and grants from NSF and NIH. The work of LI is supported by grants from NSF, NGA, HFSP, ONR and DARPA. We wish to thank Mario Blaum and Robert McEliece for encouragement and feedback.

Appendix A. Discrete case

A.1. Multinomial case (Dirichlet prior)

In the two-dimensional case, consider two Dirichlet distributions $D_1 = D(a_1, b_1)(x) = C_1 x^{a_1-1}(1-x)^{b_1-1}$ and $D_2 = D(a_2, b_2)(x) = C_2 x^{a_2-1}(1-x)^{b_2-1}$, with $C_1 = \Gamma(a_1+b_1)/\Gamma(a_1)\Gamma(b_1)$, and similarly for C_2. To calculate the relative entropy in the two dimensional case, we use the formula (Gradshteyn & Ryzhik, 1980)

$$\int_0^1 x^{u-1}(1-x)^{v-1}\log x\,\mathrm{d}x = B(u,v)[\Psi(u) - \Psi(u+v)] \tag{50}$$

where $B(u, v)$ is the beta function $B(u,v) = \int_0^1 x^{u-1}(1-x)^{v-1}\mathrm{d}x = \Gamma(u)\Gamma(v)/\Gamma(u+v)$ and $\Psi(x)$ is the derivative of the logarithm of the gamma function $\Psi(x) = d(\log\Gamma(x))/\mathrm{d}x$. A cross term of the form

$$F(D_1, D_2) = \int_0^1 C_1 x^{a_1-1}(1-x)^{b_1-1}[\log C_2 + (a_2-1)\log x + (b_2-1)\log(1-x)]\mathrm{d}x \tag{51}$$

is equal to

$$F(D_1, D_2) = \log C_2 + (a_2-1)[\Psi(a_1) - \Psi(a_1+b_1)] + (b_2-1)[\Psi(b_1) - \Psi(a_1+b_1)] \tag{52}$$

using the fact that $C_1 B(a_1, b_1) = 1$. In particular, the entropy of a two-dimensional Dirichlet distribution such as D_1 is obtained by taking: $-F(D_1, D_1)$. With some algebra, the Kullback–Liebler divergence between any two Dirichlet distributions is finally given by:

$$K(D_1, D_2) = \log\frac{C_1}{C_2} + (a_1-a_2)[\Psi(a_1) - \Psi(a_1+b_1)] + (b_1-b_2)[\Psi(b_1) - \Psi(a_1+b_1)]. \tag{53}$$

With n successes in the sequence D, the posterior is a Dirichlet distribution $D_2(a_2, b_2)$ with Baldi and Brunak (2001)

$$a_2 = a_1 + n \quad \text{and} \quad b_2 = b_1 + (N-n). \tag{54}$$

Using this relation between the prior and the posterior, we get the surprise

$$S(D_1, D_2) = \log\frac{C_1}{C_2} + n[\Psi(a_1+b_1) - \Psi(a_1)] + (N-n)[\Psi(a_1+b_1) - \Psi(b_1)]. \tag{55}$$

Using the general fact that $\Psi(x) - \Psi(y) = \sum_{k=0}^{\infty}(\frac{1}{y+k} - \frac{1}{x+k})$, which implies $\Psi(x+n) - \Psi(x) = \sum_{k=0}^{n-1}\frac{1}{x+k}$ when n is an integer, we get

$$S(D_1, D_2) = \log\frac{C_1}{C_2} + n\left(\sum_{k=0}^{\infty}\frac{1}{a_1+k} - \frac{1}{a_1+b_1+k}\right) + (N-n)\left(\sum_{k=0}^{\infty}\frac{1}{b_1+k} - \frac{1}{a_1+b_1+k}\right). \tag{56}$$

Now we have

$$\sum_{k=0}^{\infty}\left(\frac{1}{a_1+k} - \frac{1}{a_1+b_1+k}\right) = \sum_{k=0}^{\lfloor b_1\rfloor-1}\left(\frac{1}{a_1+k}\right) + \text{Rest} \tag{57}$$

where

$$0 \le \text{Rest} = \sum_{k=0}^{\infty}\left(\frac{1}{a_1+\lfloor b_1\rfloor+k} - \frac{1}{a_1+b_1+k}\right) \le (b_1 - \lfloor b_1\rfloor)\sum_{k=0}^{\infty}\frac{1}{(a_1+\lfloor b_1\rfloor+k)^2} \tag{58}$$

and similarly for the symmetric term. The rest is exactly 0 when a_1 and b_1 (and hence a_2 and b_2) are integers, and in general decreases with the size of a_1 and b_1. This yields the approximation

$$S(D_1, D_2) \approx \log \frac{C_1}{C_2} + n \left(\sum_{k=0}^{\lfloor b_1 \rfloor - 1} \frac{1}{a_1 + k} \right) + (N - n) \left(\sum_{k=0}^{\lfloor a_1 \rfloor - 1} \frac{1}{b_1 + k} \right). \quad (59)$$

This approximation is *exact* when a_1 and b_1 are integers. Now for $x > 0$ we have $\log((x+n)/x) < \sum_{k=0}^{n-1} 1/(x+k) < \log((x+n-1)/x) + 1/x$ or $0 < \sum_{k=0}^{n-1} 1/(x+k) - \log((x+n)/x) < 1/x$. Thus,

$$S(D_1, D_2) \approx \log \frac{C_1}{C_2} + n \log \frac{a_1 + b_1}{a_1} + (N - n) \log \frac{a_1 + b_1}{b_1}. \quad (60)$$

Now we have,

$$\log \frac{C_1}{C_2} = \log \frac{\Gamma(a_1 + b_1)\Gamma(a_2)\Gamma(b_2)}{\Gamma(a_1)\Gamma(b_1)\Gamma(a_2 + b_2)} = \log \frac{(a_1 + n - 1)(a_1 + n - 2) \dots a_1 (b_1 + N - n - 1) \dots b_1}{(a_1 + b_1 + N - 1)(a_1 + b_1 + N - 2) \dots (a_1 + b_1)}.$$

We can use bounds of the form $\log a + \int_{a_1}^{a_1+n+1} \log x \mathrm{d}x < \log a_1 + \cdots \log(a_1 + n - 1) \leq \log a_1 \int_{a_1+1}^{a_1+n} \log x \mathrm{d}x$ to estimate this term. Alternatively, one can assume that a_1 and b_1 are integers and use binomial coefficient approximations, such as those in Bollobas (1985). In all cases, neglecting constant terms and terms of order $\log N$, if we let $n = pN$ $(0 < p < 1)$ and N go to infinity we have

$$\log \frac{C_1}{C_2} \approx -\log \binom{N}{n} \approx -NH(p) \quad (61)$$

where $H(p)$ is the entropy of the (p, q) distribution with $q = 1 - p$. Thus when $N \to \infty$, and $n = pN$ with $0 < p < 1$ we have

$$S(D_1, D_2) \approx N \left(p \log \frac{a_1 + b_1}{a_1} + q \log \frac{a_1 + b_1}{b_1} - H(p) \right) \approx N\mathrm{K}(p, a_1) \quad (62)$$

where $K(p, a_1)$ is the relative entropy between the empirical distribution (p, q) and the expectation of the prior $(\frac{a_1}{a_1+b_1}, \frac{b_1}{a_1+b_1})$. Thus, asymptotically surprise grows linearly with the number of data points with a proportionality coefficient that depends on the discrepancy between the expectation of the prior and the observed distribution. The same relationship can be expected to be true in the case of a multinomial model.

A.2. Symmetric prior ($a_1 = b_1$)

Consider now the case of a symmetric prior, then

$$S(D_1, D_2) = \log \frac{C_1}{C_2} + N[\Psi(2a_1) - \Psi(a_1)]. \quad (63)$$

Using formulas in Gradshteyn and Ryzhik (1980), $\Psi(2a_1) - \Psi(a_1) = \sum_{k=0}^{\infty} \frac{(-1)^k}{2a_1 + k} + \log 2$, thus

$$S(D_1, D_2) = \log \frac{C_1}{C_2} + N \sum_{k=0}^{\infty} \frac{(-1)^k}{2a_1 + k} + \log 2 \approx N \left(\sum_{k=0}^{\infty} \frac{(-1)^k}{2a_1 + k} + \log 2 - H(p) \right) \quad (64)$$

the approximation being in the regime $n = pN$ and $N \to \infty$. When a_1 is an integer, we also have $\Psi(2a_1) - \Psi(a_1) = \sum_{k=1}^{2a_1-1} (-1)^{k+1}/k = \sum_{k=a_1}^{2a_1-1} 1/k$. Thus when a_1 is an integer

$$S(D_1, D_2) = N \left[\sum_{k=a_1}^{2a_1-1} \frac{1}{k} \right] + \log \frac{(2a_1 - 1)\binom{2a_1-2}{a_1-1}}{(2a_1 + N - 1)\binom{N+2a_1-2}{n+a_1-1}}. \quad (65)$$

As $N \to \infty$ with $0 < p < 1$

$$S(D_1, D_2) \approx N \left[\sum_{k=a_1}^{2a_1-1} \frac{1}{k} \right] - \log \binom{N + 2a_1 - 2}{n + a_1 - 1} \approx N \left[\sum_{k=a_1}^{2a_1-1} \frac{1}{k} \right] - \log \binom{N}{n} \quad (66)$$

and therefore

$$S(D_1, D_2) \approx N \left[\sum_{k=a_1}^{2a_1-1} \frac{1}{k} - H(p) \right]. \quad (67)$$

For instance, when $a_1 = b_1 = 1$, this gives:

$$S(D_1, D_2) = N - \log(N + 1) - \log \binom{N}{n} \quad (68)$$

with the asymptotic form

$$S(D_1, D_2) \approx N(1 - H(p)) + \log \frac{\sqrt{2N\pi pq}}{N + 1} \approx N(1 - H(p)). \quad (69)$$

With a uniform symmetric prior, the empirical distribution with maximal entropy brings the least information. When $a_1 = b_1 = 5$ this gives $R(D_1, D_2) \approx N[0.746 - H(p)]$. As we increase $a_1 + b_1$, keeping $a_1 = b_1$, the constant $\sum_{a_1}^{2a_1-1}(1/k)$ decreases to its asymptotic value $\log 2$ which corresponds to the asymptotic form $S(D_1, D_2) \approx NK(p, 0.5)$. The stronger the strength of the uniform prior (the larger $a_1 + b_1$), the smaller the surprise created by a die with maximum entropy.

A.3. Poisson case (Gamma prior)

We consider two Gamma distributions $\Gamma_1 = \Gamma_1(a_1, b_1)x = C_1 x^{a_1-1} \mathrm{e}^{-b_1 x}$ and $\Gamma_2 = \Gamma_2(a_2, b_2)(x) = C_2 x^{a_2-1} \mathrm{e}^{-b_2 x}$ with $C_1 = b_1^{a_1}/\Gamma(a_1)$, and similarly for C_2. To calculate the relative entropy, we use the formula (Gradshteyn & Ryzhik, 1980)

$$\int_0^{\infty} x^{u-1} \mathrm{e}^{-vx} \log x \mathrm{d}x = \frac{\Gamma(u)}{v^u} [\Psi(u) - \log v]. \quad (70)$$

A cross term $F(\Gamma_1, \Gamma_2)$

$$F(\Gamma_1, \Gamma_2) = \int_0^{\infty} C_1 x^{a_1-1} \mathrm{e}^{-b_1 x} [\log C_2 + (a_2 - 1) \log x - b_2 x] \mathrm{d}x \quad (71)$$

is then equal to:

$$F(\Gamma_1, \Gamma_2) = a_2 \log b_2 - \log \Gamma(a_2) - b_2 \frac{a_1}{b_1} + (a_2 - 1)[\Psi(a_1) - \log b_1]. \quad (72)$$

With some algebra, the KL divergence between two Gamma distributions is given by

$$K(\Gamma_1, \Gamma_2) = \log \frac{b_1^{a_1}}{a_2^{b_2}} + \log \frac{\Gamma(a_2)}{\Gamma(a_1)} + b_2 \frac{a_1}{b_1} - a_1 + (a_1 - a_2)[\Psi(a_1) - \log b_1]. \quad (73)$$

With N observations in D, the posterior is Gamma and satisfies

$$a_2 = a_1 + N\bar{m} \qquad b_2 = b_1 + N. \quad (74)$$

With these values, this finally yields the surprise $S(\Gamma_1, \Gamma_2)$

$$S(D, \mathcal{M}) = K(\Gamma_1, \Gamma_2) = a_1 \log \frac{b_1}{b_1 + N} - N\bar{m} \log(b_1 + N) + \log \frac{\Gamma(a_1 + N\bar{m})}{\Gamma(a_1)} + \frac{Na_1}{b_1} + N\bar{m}[\log b_1 - \Psi(a_1)]. \quad (75)$$

When N is large, using Stirling's formula, the dominant terms in $N \log N$ cancel leaving the approximation

$$S(D, \mathcal{M}) \approx N \left(\frac{a_1}{b_1} - \bar{m}[1 - \log \bar{m} + \Psi(a_1) - \log b_1] \right). \quad (76)$$

Appendix B. Continuous case

B.1. Unknown mean/known variance

Consider now two Gaussians $G_1(\mu_1, \sigma_1)$ and $G_2(\mu_2, \sigma_2)$. Then, after some algebra, the cross term is given by

$$F(G_1, G_2) = \int_{-\infty}^{+\infty} G_1 \log G_2 \mathrm{d}x = -\frac{1}{2} \log(2\pi\sigma_2^2) - \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} \quad (77)$$

here using for simplicity natural logarithms. $F(G, G) = \frac{1}{2} \log [2\pi e \sigma^2] = H(G)$ is the entropy. The Kullback–Liebler divergence can then be obtained

$$K(G_1, G_2) = -\frac{1}{2} + \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2}. \quad (78)$$

Consider now a data set with N points $d_1, \ldots, d_N$ with empirical mean $\bar{m}$. If the data has known variance σ^2, then the posterior parameters are given by:

$$\mu_2 = \frac{\frac{\mu_1}{\sigma_1^2} + \frac{N\bar{m}}{\sigma^2}}{\frac{1}{\sigma_1^2} + \frac{N}{\sigma^2}} \quad \text{and} \quad \frac{1}{\sigma_2^2} = \frac{1}{\sigma_1^2} + \frac{N}{\sigma^2}. \quad (79)$$

In the general case

$$S(G_1, G_2) = \log \frac{\sigma}{\sqrt{\sigma^2 + N\sigma_1^2}} + N \frac{\sigma_1^2}{2\sigma^2} + \frac{N^2 \sigma_1^2 (\mu_1 - \bar{m})^2}{2\sigma^2(\sigma^2 + N\sigma_1^2)} \approx \frac{N}{2\sigma^2}[\sigma_1^2 + (\mu_1 - \bar{m})^2] \quad (80)$$

when N is large. In the special case where the prior has the same variance has the data $\sigma_1 = \sigma$ then the formula simplifies a little and yields

$$S(G_1, G_2) = \frac{N}{2} - \frac{1}{2} \log(N + 1) + \frac{N^2(\mu_1 - \bar{m})^2}{2(N + 1)\sigma^2} \approx \frac{N}{2\sigma^2}[\sigma^2 + (\mu_1 - \bar{m})^2] \quad (81)$$

when N is large. In any case, surprise grows linearly with N with a coefficient that is the sum of the prior variance and the square difference between the expected mean and the empirical mean scaled by the variance of the data.

B.2. Unknown variance/known mean

In the case of unknown variance and known mean, we have a family $M(\sigma^2)$ of models with a conjugate prior for σ^2 that is a scaled inverse gamma distribution (Gelman et al., 1995)

$$\Gamma_1(\nu_1, s_1) = \frac{(\frac{\nu_1}{2})^{\nu_1/2} s_1^{\nu_1}}{\Gamma(\frac{\nu_1}{2})} (\sigma^2)^{-(\frac{\nu_1}{2}+1)} \mathrm{e}^{-\frac{\nu_1 s_1^2}{2\sigma^2}} \mathrm{d}\sigma^2 = C_1 (\sigma^2)^{-(\frac{\nu_1}{2}+1)} \mathrm{e}^{-\frac{\nu_1 s_1^2}{2\sigma^2}} \mathrm{d}\sigma^2 \quad (82)$$

with $\nu_1 > 0$ degrees of freedom and scale $s_1 > 0$. F can be computed expanding the integrals and using the fact that $\int_0^{+\infty} x^{\nu/2-1} \mathrm{e}^{-x} \log x = \Gamma(\frac{\nu}{2}) \Psi(\frac{\nu}{2})$. This yields:

$$F(\nu_1, s_1; \nu_2, s_2) = \log \frac{(\nu_2/2)^{\nu_2/2} s_2^{\nu_2}}{\Gamma(\frac{\nu_2}{2})} + \left(\frac{\nu_2}{2} + 1\right) \left[\Psi\left(\frac{\nu_1}{2}\right) + \log \frac{2}{\nu_1 s_1^2} \right] - \frac{\nu_2 s_2^2}{2 s_1^2}. \quad (83)$$

The posterior is then a scaled inverse gamma distribution (Gelman et al., 1995) with

$$\nu_2 = \nu_1 + N \quad \text{and} \quad s_2^2 = \frac{\nu_1 s_1^2 + N\bar{\sigma}^2}{\nu_1 + N} \quad (84)$$

where $\bar{\sigma}^2$ is the empirical variance $\bar{\sigma}^2 = \sum_i (x_i - m)^2 / N$, based on the known mean m. The surprise is given by

$$S(\Gamma_1, \Gamma_2) = \log \frac{C_1}{C_2} - \frac{N}{2} \left(\Psi\left(\frac{\nu_1}{2}\right) + \log \frac{2}{\nu_1 s_1^2} \right) + \frac{N\bar{\sigma}^2}{2 s_1^2}. \quad (85)$$

For large values of N, taking only the leading terms

$$S(\Gamma_1, \Gamma_2) \approx \frac{N}{2} \left(\frac{\bar{\sigma}^2}{s_1^2} + \log \frac{\nu_1 s_1^2}{2} - \Psi\left(\frac{\nu_1}{2}\right) \right) + \log \Gamma\left(\frac{\nu_1 + N}{2}\right) - \frac{\nu_1 + N}{2} \log \frac{\nu_1 + N}{2} - \frac{(\nu_1 + N)}{2} \log \frac{\nu_1 s_1^2 + N\bar{\sigma}^2}{\nu_1 + N} \quad (86)$$

$$S(\Gamma_1, \Gamma_2) \approx \frac{N}{2} \left(\frac{\bar{\sigma}^2}{s_1^2} + \log \frac{\nu_1 s_1^2}{2\bar{\sigma}^2} - \Psi\left(\frac{\nu_1}{2}\right) \right). \quad (87)$$

Thus surprise information scales linearly with N, with a coefficient of proportionality that typically depends mostly on the ratio of the empirical variance to the scale parameters s_1^2, which is roughly the expectation of the prior [the expectation of the prior is $\nu_1 s_1^2/(\nu_1 - 2)$ provided $\nu_1 > 2$]. The effects of very large of very small values of $\bar{\sigma}$, or ν_1 can also be seen in the formula above. In particular, surprise is largest when the empirical variance $\bar{\sigma}^2$ goes to 0 or infinity, i.e. is very different from the prior expectation.

B.3. Unknown mean/unknown variance

In the case of unknown mean and unknown variance, we have a family $M(\mu, \sigma^2)$ of models with a conjugate prior of the form $G_1\Gamma_1 = P(\mu|\sigma^2)P(\sigma^2) = G_1(\mu_1, \sigma^2/\kappa_1)\Gamma_1(\nu_1, s_1)$. Thus the prior has four parameters $(\mu_1, \kappa_1, \nu_1, s_1)$, with $\kappa_1 > 0$, $\nu_1 > 0$, and $s_1 > 0$. The conjugate posterior has the same form, with similar parameters $(\mu_2, \kappa_2, \nu_2, s_2)$ satisfying (see for instance (Gelman et al., 1995))

$$\begin{aligned} \mu_2 &= \frac{\kappa_1}{\kappa_1 + N} \mu_1 + \frac{N}{\kappa_1 + N} \bar{m} \\ \kappa_2 &= \kappa_1 + N \\ \nu_2 &= \nu_1 + N \\ \nu_2 s_2^2 &= \nu_1 s_1^2 + (N - 1)\bar{\sigma}^2 + \frac{\kappa_1 N}{\kappa_1 + N} (\bar{m} - \mu_1)^2 \end{aligned} \quad (88)$$

with $\bar{m} = \sum x_i / N$ and $\bar{\sigma}^2 = \sum (x_i - \bar{m})^2 / (N - 1)$. Computation of $F = F(\mu_1, \kappa_1, \nu_1, s_1; \mu_2, \kappa_2, \nu_2, s_2)$ is similar to the two cases treated above and yields:

$$F(\mu_1, \kappa_1, \nu_1, s_1; \mu_2, \kappa_2, \nu_2, s_2) = -\frac{1}{2} \left[\log \frac{2\pi}{\kappa_2} + \frac{\kappa_2}{\kappa_1} + \log \frac{\nu_1 s_1^2}{2} \right.$$

$$- \Psi\left(\frac{\nu_1}{2}\right) + \kappa_2(\mu_2-\mu_1)^2 s_1^{-2}\Bigg] + \frac{\log(\frac{\nu_2}{2})^{\nu_2/2} s_2^{\nu_2}}{\Gamma\left(\frac{\nu_2}{2}\right)} + \left(\frac{\nu_2}{2}+1\right)\left[\Psi\left(\frac{\nu_1}{2}\right) + \log\frac{2}{\nu_1 s_1^2}\right] - \frac{\nu_2 s_2^2}{2s_1^2}. \quad (89)$$

From Eq. (89), we can derive the surprise

$$S(G_1\Gamma_1, G_2\Gamma_2) = \frac{1}{2}\left[\log\frac{\kappa_1}{\kappa_2} - 1 + \frac{\kappa_2}{\kappa_1} + \kappa_2(\mu_2-\mu_1)^2 s_1^{-2}\right] + \log\frac{C_1}{C_2} + \left(\frac{\nu_1-\nu_2}{2}\right)\left[\Psi\left(\frac{\nu_1}{2}\right) + \log\frac{2}{\nu_1 s_1^2}\right] + \frac{\nu_2 s_2^2 - \nu_1 s_1^2}{2s_1^2}. \quad (90)$$

Substituting the value of the posterior parameters

$$S(G_1\Gamma_1, G_2\Gamma_2) = \frac{1}{2}\log\frac{\kappa_1}{\kappa_1+N} + \frac{N}{2\kappa_1} + \frac{\kappa_1+N}{2}\left[\frac{N(\bar{m}-\mu_1)}{(\kappa_1+N)s_1}\right]^2 + \log\frac{C_1}{C_2} - \frac{N}{2}\left[\Psi\left(\frac{\nu_1}{2}\right) + \log\frac{2}{\nu_1 s_1^2}\right] + \frac{(N-1)\bar{\sigma}^2 + \frac{\kappa_1 N}{\kappa_1+N}(\bar{m}-\mu_1)^2}{2s_1^2}. \quad (91)$$

For simplicity, we can consider the case where $\mu_1 = \bar{m}$. Then

$$S(G_1\Gamma_1, G_2\Gamma_2) = \frac{1}{2}\log\frac{\kappa_1}{\kappa_1+N} + \frac{N}{2\kappa_1} + \log\frac{C_1}{C_2} - \frac{N}{2}\left[\Psi\left(\frac{\nu_1}{2}\right) + \log\frac{2}{\nu_1 s_1^2}\right] + \frac{(N-1)\bar{\sigma}^2}{2s_1^2}. \quad (92)$$

In all cases, for large values of N we always have the approximation

$$S(G_1\Gamma_1, G_2\Gamma_2) \approx \frac{N}{2}\left(\frac{1}{\kappa_1} + \frac{\bar{\sigma}^2}{s_1^2} + \log\frac{\nu_1 s_1^2}{2\bar{\sigma}^2} - \Psi\left(\frac{\nu_1}{2}\right) + \frac{(\bar{m}-\mu_1)^2}{s_1^2}\right). \quad (93)$$

Surprise is linear in N with a coefficient that is essentially the sum of the coefficients derived in the unknown mean and unknown variance partial cases.

Appendix C. Exponential families with conjugate priors

Let A^1 and A^2 denote the parameters of two distributions P_1 and P_2 in the exponential family. Simple integration yields

$$F(P_1, P_2) = \log C_2 + \sum_{i=1}^{k} A_i^2 E_{A^1}[w_i(\theta)] \quad (94)$$

where $E_{A^1}[w_i(\theta)]$ denotes expectation with respect to P_1. Surprise S is then derived from $F(P_1, P_1) - F(P_1, P_2)$.

References

Aczel, J., & Daroczy, Z. (1975). *On measures of information and their characterizations*. New York: Academic Press.

Baldi, P. (2002). A computational theory of surprise. In M. Blaum, P. G. Farrell, & H. C. A. van Tilborg (Eds.), *Information, coding, and Mathematics*. Boston: Kluwer Academic Publishers.

Baldi, P., & Brunak, S. (2001). *Bioinformatics: The machine learning approach* (2nd ed.). Cambridge, MA: MIT Press.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, *12*, 387–415.

Bartlett, M. S. (1952). The statistical significance of odd bits of information. *Biometrika*, *39*, 228–237.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. New York: Springer-Verlag.

Blahut, R. E. (1987). *Principles and practice of information theory*. Reading, MA: Addison-Wesley.

Bollobas, B. (1985). *Random graphs*. London: Academic Press.

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

Carmi, R., & Itti, L. (2006). The role of memory in guiding attention during natural vision. *Journal of Vision*, *6*(9), 898–914.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: John Wiley.

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

Evans, M. (1997). Bayesian inference procedures derived via the concept of relative surprise. *Communications in Statistics*, *26*(5), 1125–1143.

Evans, M., Guttman, I., & Swartz, T. (2006). Optimality and computations for relative surprise inferences. *Canadian Journal of Statistics*, *34*(1), 113–129.

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

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman and Hall.

Good, I. (1956). The surprise index for the multivariate normal distribution. *Annals of Mathematical Statistics*, 1130–1135.

Gradshteyn, I. S., & Ryzhik, I. M. (1980). *Table of integrals, series, and products*. New York: Academic Press.

Grimson, W.E.L., Stauffer, C., Romano, R., & Lee, L. (1998). Using adaptive tracking to classify and monitor activities in a site. In *Proc. CVPR*.

Grossberg, S. (2000). How hallucinations may arise from brain mechanisms of learning, attention, and volition. *Journal of the International Neuropsychological Society*, *6*, 583–592.

Grossberg, S., & Raizada, R. (2000). Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex. *Vision Research*, *40*(10–12), 1413–1432.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning. Data mining, inference, and prediction*. New York, NY: Springer.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology (London)*, *160*, 106–154.

Itti, L., & Baldi, P. 2005. A principled approach to detecting surprising events in video. In *Proc. IEEE conference on computer vision and pattern recognition* (pp. 631–637).

Itti, L., & Baldi, P. (2006). Bayesian surprise attracts human attention. In *Advances in neural information processing systems*: *Vol. 19* (pp. 1–8). Cambridge, MA: MIT Press.

Itti, L., & Baldi, P. F. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295–1306.

Itti, L., Dhavale, N., & Pighin, F. (2003). Realistic avatar eye and head animation using a neurobiological model of visual attention. In B. Bosacchi, D. B. Fogel, & J. C. Bezdek (Eds.), *Proc. SPIE 48th annual international symposium on optical science and technology: Vol. 5200* (pp. 64–78). Bellingham, WA: SPIE Press.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10–12), 1489–1506.

Itti, L., & Koch, C. (2001). Computational modeling of visual attention. *Natures Review 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. (1986). Bayesian methods: General background. In J. Justice (Ed.)., *Maximum entropy and Bayesian methods in statistics* (pp. 1–25). Cambridge: Cambridge University Press.

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

Jumarie, G. (1990). *Relative information*. New York: Springer Verlag.

Kullback, S. (1968). *Information theory and statistics*. New York: Dover, First edition in 1959.

Kvalseth, T. (1987). Stimulus probability, surprise and reaction time. *Proceedings of the Human Factor Society*, *1*, 147–150.

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

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, *37*(1), 145–151.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision*, *10*(3), 165–188.

McEliece, R. J. (1977). *The theory of information and coding*. Reading, MA: Addison-Wesley Publishing Company.

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.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387–391.

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. *Natures Review Neuroscience*, *4*(3), 193–202.

Redheffer, R. M. (1951). A note on the surprise index. *Annals of Mathematical Statistics*, *22*, 128–130.

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

Renninger, L. W., Coughlan, J., Verghese, P., & Malik, J. (2005). An information maximization model of eye movements. *Advances in Neural Information Processing Systems*, *17*, 1121–1128.
Renyi, A. (1961). On measures of information and entropy. In *Proceedings of the 4th Berkeley symposium on mathematics, statistics and probability* (pp. 547–561).
Savage, L. J. (1972). *The foundations of statistics*. New York: Dover, First edition in 1954.
Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423. 623–656.
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.
Suder, K., & Worgotter, F. (2000). The control of low-level information flow in the visual system. *Reviews in the Neurosciences*, *11*(2–3), 127–146.
Tishby, N., Pereira, F., & Bialek, W. (1999). The information bottleneck method. In B. Hajek, & R. S. Sreenivas (Eds.), *Proceedings of the 37th annual allerton conference on communication, control, and computing* (pp. 368–377). University of Illinois.
Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer Verlag.
Weaver, W. (1948). Probability, rarity, interest and surprise. *Scientific Monthly*, *67*(6), 390–392.
Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Natures Review Neuroscience*, *5*(6), 495–501.
Wong, A. K. C., & You, M. (1985). Entropy distance of random graphs with application to structural pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *7*, 599–609.
Zhaoping, L., & May, K. A. (2007). Psychophysical tests of the hypothesis of a bottom-up saliency map in primary visual cortex. *PLoS Computational Biology*, *3*(4), e62.