(ICNN&B'05 Keynote Speech)

# Attention: Bits versus Wows

*(Invited Paper)*

Pierre Baldi
Department of Computer Science
University of California, Irvine
Irvine, CA 92697
E-mail: pfbaldi@ics.uci.edu

Laurent Itti
Department of Computer Science
University of Southern California
Los Angeles, CA 90089
E-mail: itti@usc.edu

***Abstract*— The concept of surprise is central to sensory processing, adaptation and learning, attention, and decision making. Yet, no widely-accepted mathematical theory currently exists to quantitatively characterize surprise elicited by a stimulus or event, for observers that range from single neurons to complex natural or engineered systems. We describe a formal Bayesian definition of surprise that is the only consistent formulation under minimal axiomatic assumptions. Surprise quantifies how data affects a natural or artificial observer, by measuring the difference between posterior and prior beliefs of the observer. Using this framework we measure the extent to which humans direct their gaze towards surprising items while watching television and video games. Humans are strongly attracted to locations of high Bayesian surprise, with 72% of all human gaze shifts directed towards locations more surprising than the average, a figure which rises to 84% when considering only gaze targets simultaneously selected by all subjects. The resulting theory of surprise is applicable across different spatio-temporal scales, modalities, and levels of abstraction.**

## I. Introduction

Visual attention is deployed based on a combination of bottom-up cues derived from the visual stimuli, and top-down cues derived from volition, expectations, prior recognitions of objects and of scene contexts, and goals of the observer [1]. Through the interplay between bottom-up and top-down, attention operates a rapid pruning of vast amounts of incoming sensory information, to focus slower and more sophisticated analysis resources onto only a few most important subsets of the available data [2], [3]. In real-life environments, often there is no time for detailed and thorough analysis of all inputs: a savana monkey must typically take action faster than it can fully recognize a rapidly approaching leopard. Consequently, evolving rapid and computationally efficient approximations or heuristics to important information is key to predation, escape and mating. We show that current computational understanding of the nature of these heuristics is encompassed and extended by a simple Bayesian measure of intrinsic stimulus surprise.

A productive approach to studying attentional selection and to characterising the putative underlying heuristic computations that guide attention in complex scenes uses eye-tracking devices, to evaluate image statistics and neural responses at the locations visited by gaze [4], [5]. With static natural stimuli, human observers preferentially look at image locations where local contrast and entropy, edges and corners, and bottom-up stimulus saliency are significantly stronger than expected by chance [6], [7], [8], [9]. Employing synthetic laboratory stimuli, several psychophysical studies have recently established that transient and dynamic aspects of stimuli also strongly capture attention, namely flicker, onsets of novel stimuli, abrupt changes in luminance, and motion energy [10], [11], [12]. However, with dynamic natural scenes, to which we are confronted during most of our lives, it is not known whether these features remain strong attractors of attention and suitable indicators of important information.

Here we propose that attention is attracted by features that are “surprising” and that surprise is a general, information-theoretic concept that must be analytically formalized across spatio-temporal scales and data types [14], [15]. We propose a Bayesian definition of surprise and test the hypothesis that surprise attracts attention using psychophysical eye-tracking experiments in humans subjecs watching realistic stimuli.

## II. Theory

Surprise is fundamentally a property of *data* with respect to an observer. As such, its definition must be independent of the nature of the data or the observer. It must apply equally well to visual, olfactory, or digital data and to information processing observers that range from synapses, to neuronal circuits, to organisms, to computer devices. Surprise exists only in the presence of uncertain environments, and therefore its essence must be probabilistic. Consistently with the Bayesian approach to data modeling and inference, the background information of an observer is captured by his/her/its prior probability distribution over the current space of hypotheses or models $\mathcal{M}$. The fundamental effect of the data $D$ on the observer is to change the prior distribution $P(M)$ into the posterior distribution $P(M|D)$ via Bayes theorem $P(M|D) = P(M)P(D|M)/P(D)$. Thus surprise can be measured by the distance between the prior and posterior distributions, which is best done using the relative entropy or Kullback-Liebler $(KL)$ divergence [13]. In short, surprise is defined by the average of the log-odd ratio:

$$\begin{aligned} S(D,\mathcal{M}) &= KL(P(M|D),P(M)) \\ &= \int_{\mathcal{M}} P(M|D)\log\frac{P(M|D)}{P(M)}dM \end{aligned} \qquad (1)$$

taken with respect to the posterior distribution over the model class $\mathcal{M}$. Surprise can always be computed numerically, but

0-7803-9422-4/05/$20.00 ©2005 IEEE

also analytically in many practical cases, in particular those involving probability distributions in the exponential family [16] with conjugate or other priors.

Note that the $KL$ divergence is not symmetric but could easily be symmetrized using the the symmetric version $[KL(P(M), P(M|D)) + KL(P(M|D), P(M))]/2$. More importantly, the KL divergence has well-known theoretical advantages, including invariance with respect to reparameterizations.

A unit of surprise — a *"wow"* — may then be defined for a single model $M$ as the amount of surprise corresponding to a two-fold variation between $P(M|D)$ and $P(M)$, corresponding to $1 = \log P(M|D)/P(M)$ (with $\log$ taken in base 2). The total number of wows experienced when simultaneously considering all models is obtained through the integration in the definition of suprise.

## III. Analytical Computation of Surprise

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

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

$$P(d|w) = h(d)c(w)\exp\left(\sum_{i=1}^{k}\theta_i(w)t_i(d)\right) \qquad (2)$$

Most common distributions (Binomial, Poisson, Gaussian, etc.) are members of the exponential family. With $N$ independent data points ($D = d_1, \ldots, d_N$),

$$P(D|w) = [c(w)]^N[\prod_{j=1}^{N}h(d_j)]\exp\left(\sum_{i=1}^{k}\theta_i(w)T_i(D)\right) \qquad (3)$$

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

$$P(w;\alpha_i) = C\exp\left(\sum_{i=1}^{k}\alpha_i\theta_i(w)\right) \qquad (4)$$

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

$$S(D,\mathcal{M}) = \log\frac{C'}{C} - \sum_{i=1}^{k}T_i(D)E[\theta_i(w)] \qquad (5)$$

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

$$\begin{aligned} S(D,\mathcal{M}) &= N[\log c(w) + <\log h(d)> - <\log P(d)> \\ &- \sum_{i=1}^{k} <t_i(d)> E[\theta_i(w)]] \end{aligned} \qquad (6)$$

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

This general result for the exponential family can easily be specialized to particular cases. Consider, for example, the classical case of binary data modeled as a series of independent and identical coin tosses (Binomial Model). The family $\mathcal{M}$ of models is parameterized by the probability $0 \le w \le 1$ of observing "heads" on a coin toss, thus encompasses models of biased coins (small and large $w$ values) and of fair coins ($w \approx 0.5$). The conjugate prior is the Beta prior $P(w;\alpha,\beta) = Cw^{\alpha-1}(1-w)^{\beta-1}$ with $C = \Gamma(\alpha+\beta)/[\Gamma(\alpha)\Gamma(\beta)]$ and parameters $\alpha$, $\beta$. With a number $n$ of heads observed after tossing a coin $N$ times, the posterior is also a Beta distribution with $\alpha' = \alpha + n$ and $\beta' = \beta + (N - n)$. Integrating over models, surprise is

$$\begin{aligned} S(D,\mathcal{M}) &= \log\frac{C'}{C} - n[\Psi(\alpha+\beta+N) - \Psi(\alpha+n)] \\ &- (N-n)[\Psi(\alpha+\beta+N) - \Psi(\beta+N-n)] \end{aligned}$$

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

## IV. Experiments and Results

To test the surprise hypothesis — that surprise attracts human attention and gaze in natural scenes — we recorded eye movements from eight naïve observers (three females and five males, ages 23-32, normal or corrected-to-normal vision). Each watched a subset from 50 videoclips totaling over 25 minutes of playtime (46,489 video frames, $640 \times 480$, 60.27 Hz, mean screen luminance 30 cd/m$^2$, room 4 cd/m$^2$, viewing distance 80cm, field of view $28° \times 21°$). Clips comprised outdoors daytime and nighttime scenes of crowded environments, video games, and television broadcast including news, sports, and commercials. Right-eye position was tracked with a 240 Hz video-based device (ISCAN RK-464). Two hundred calibrated eye movement traces (10,192 saccades) were analyzed, corresponding to four distinct observers for each of the 50 clips. Figure 1 shows sample scanpaths for one videoclip.

To characterize image regions selected by participants, we process videoclips through computational metrics that output

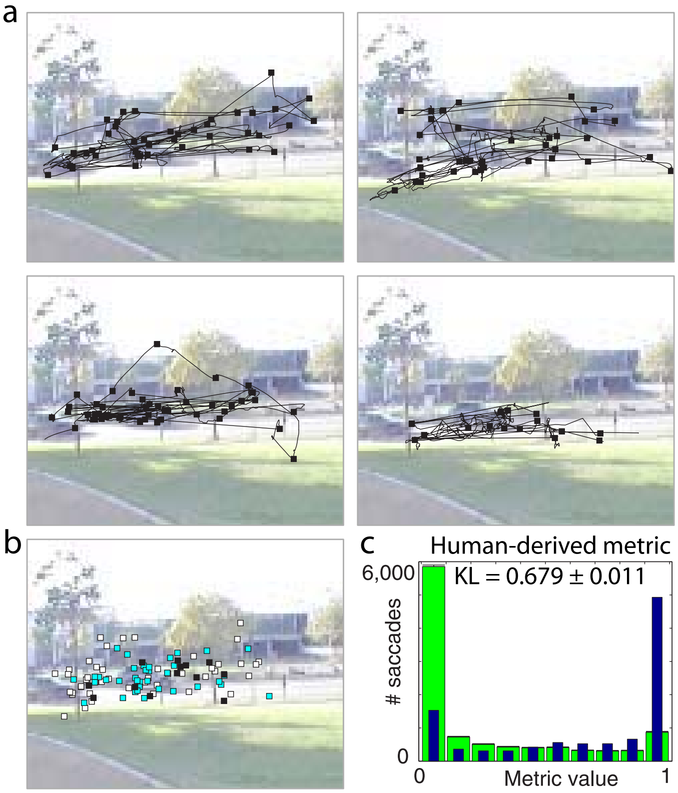

Fig. 1. **(a)** Sample eye movement traces from four observers (CZ, NM, RC, VN) watching one video clip (545 frames, 18.1s) that showed cars passing by on a fairly static background (most other clips tested in this study had dynamic backgrounds and camera motion). Squares denote locations segmented as saccade endpoints [17] (42, 36, 48, and 16 saccades for CZ, NM, RC, and VN). **(b)** Our data shows high inter-individual overlap of saccade targets, as shown here with the locations where one human saccade endpoint was nearby (within $5.6°$) the instantaneous eye position of one (white squares, 47 saccades), two (cyan squares, 36 saccades) or all three (black squares, 13 saccades) other humans. **(c)** Given this high overlap, a metric where the master map was created from the three eye movement traces other than that being tested yielded the highest $KL$ score of all metrics tested, as derived from the histograms of metric values at human (blue) and random (green) saccade targets. Indeed, this metric's map was very sparse, as demonstrated by the high number of random saccades landing on locations with near-zero metric response, yet humans preferentially saccaded towards the three active hotspots of that metric, as demonstrated by the high number of human saccades landing on locations with near-unity metric responses.

a topographic dynamic master response map, assigning in real-time a response value to every input location. A good master map would highlight, more than expected by chance, locations gazed to by observers. To score each metric we hence sample, at onset of every human saccade, master map activity around the saccade's future endpoint, and around a uniformly random endpoint (random sampling was repeated 100 times to evaluate variability). We quantify differences between histograms of master map samples collected from human and random saccades using again the Kullback-Leibler ($KL$) distance: metrics which better predict human scanpaths exhibit higher distances from random as, typically, observers non-uniformly gaze towards a minority of regions with highest metric responses while avoiding a majority of regions with low metric responses. This approach presents several advantages over simpler scoring schemes [6], [9], including agnosticity to putative mechanisms for generating saccades and the fact that applying any continuous nonlinearity to master map values would not affect scoring.

**Experimental results.** We test six computational metrics, encompassing and extending the state-of-the-art found in previous studies. The first three quantify static image properties (local intensity variance in $16 \times 16$ image patches [6]; local oriented edge density as measured with Gabor filters [3]; and local Shannon entropy in $16 \times 16$ image patches [7]). The remaining three metrics are more sensitive to dynamic events (local motion [3]; outlier-based saliency [3]; and surprise).

For all metrics, we find that humans are significantly attracted by image regions with higher metric responses. However, the static metrics typically respond vigorously at numerous visual locations (Figure 2), hence they are poorly specific and yield relatively low $KL$ scores between humans and random. The metrics sensitive to motion, outliers, and surprising events, in comparison, yield sparser maps and higher $KL$ scores.

The surprise metric of interest here quantifies low-level surprise in image patches over space and time, and at this point does not account for high-level or cognitive beliefs of our human observers. Rather, it assumes a family of simple models for image patches, each processed through 72 early feature detectors sensitive to color, orientation, motion, etc., and computes surprise from shifts in the distribution of beliefs about which models better describe the patches (all source code is freely available online at http://iLab.usc.edu/toolkit/). We find that the surprise metric significantly outperforms all other computational metrics ($p < 10^{-100}$ or better on $t$-tests for equality of $KL$ scores), scoring nearly 20% better than the second-best metric (saliency) and 60% better than the best static metric (entropy). Surprising stimuli often substantially differ from simple feature outliers; for example, a continually blinking light on a static background elicits sustained flicker due to its locally outlier temporal dynamics but is only surprising for a moment. Similarly, a shower of randomly-colored pixels continually excites all low-level feature detectors but rapidly becomes unsurprising.

**Strongest attractors of human attention.** Clearly, in our and previous eye-tracking experiments, in some situations potentially interesting targets were more numerous than in others. With many possible targets, different observers may orient towards different locations, making it more difficult for a single metric to accurately predict all observers. Hence we consider (Figure 3) subsets of human saccades where at least two, three, or all four observers simultaneously agreed on a gaze target. Observers could have agreed based on bottom-up factors (e.g., only one location had interesting visual appearance at that time), top-down factors (e.g., only one object was of current cognitive interest), or both (e.g., a single cognitively interesting object was present which also had distinctive appearance). Irrespectively of the cause for agreement, it indicates consolidated belief that a location was attractive. While the $KL$ scores of all metrics improved when

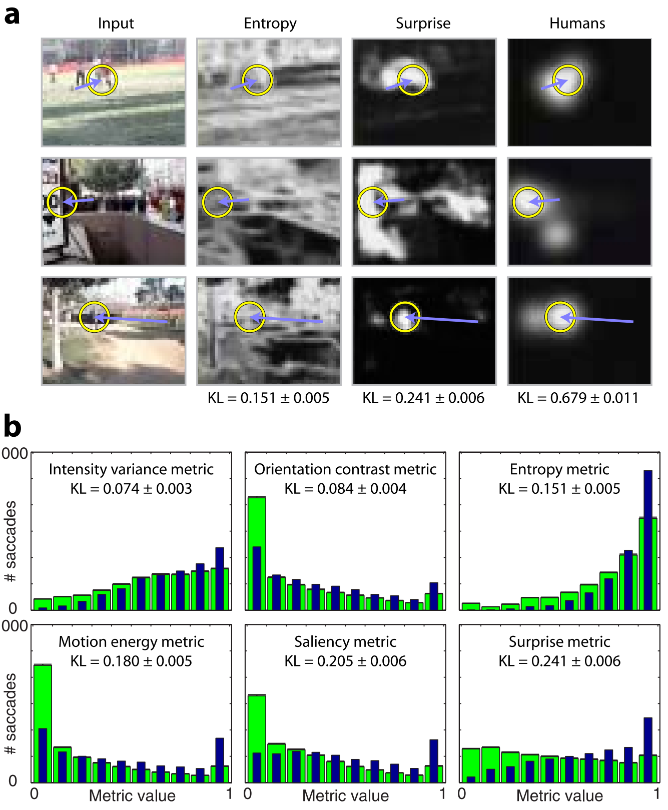

Fig. 2. **(a)** Sample frames from our video clips, with corresponding human saccades and predictions from the entropy, surprise, and human-derived metrics. Entropy maps, like intensity variance and orientation maps, exhibited many locations with high responses, hence had low specificity and were poorly discriminative. In contrast, motion, saliency, surprise, and human-derived maps were much sparser and more specific, with surprise significantly more often on target than motion and saliency. For three example frames (first column), saccades from one subject are shown (arrows) with corresponding apertures over which master map activity at the saccade endpoint was sampled (circles). Associated master maps exemplify the varying degrees of sparseness and specificity of the metrics tested. **(b)** $KL$ scores for these metrics indicate significantly different performance levels, and a strict ranking of variance $<$ orientation $<$ entropy $<$ motion $<$ saliency $<$ surprise $<$ human-derived. $KL$ scores were computed by comparing the number of human saccades landing onto each given range of master map values (narrow blue bars) to the number of random saccades hitting the same range (wider green bars). A score of zero would indicate equality between the human and random histograms, i.e., humans did not tend to hit various master map values any differently from expected by chance, or, the master map could not predict human saccades better than random saccades. Among the six computational metrics tested in total, surprise performed best, in that surprising locations were relatively few yet reliably gazed to by humans.

progressively focusing onto only those locations, dynamic metrics improved more steeply, indicating that stimuli which more reliably attracted all observers carried more motion, saliency, and surprise. Surprise remained significantly the best metric to characterize these agreed-upon attractors of human gaze ($p < 10^{-100}$ or better on $t$-tests for equality of $KL$ scores).

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

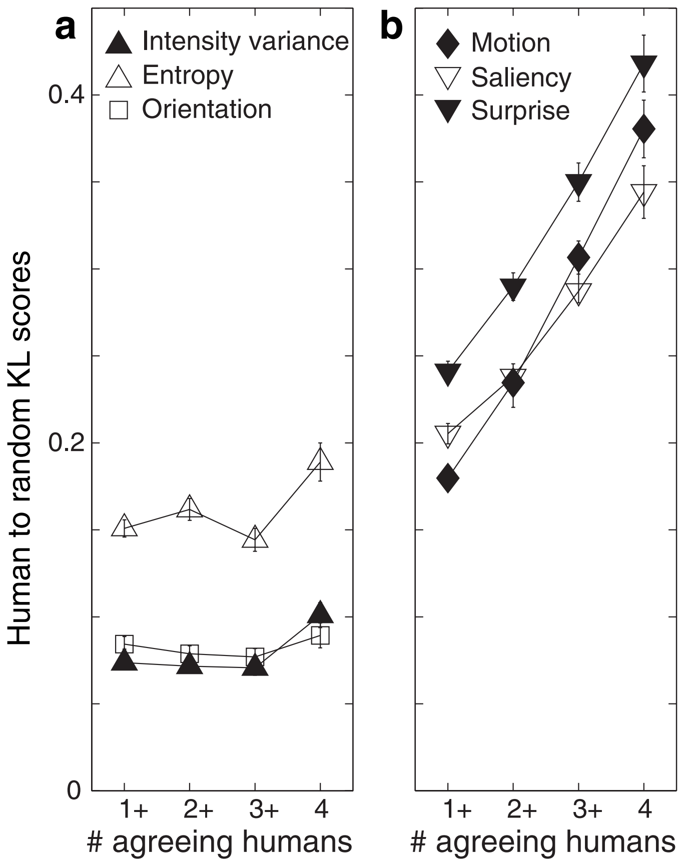

Fig. 3. $KL$ scores when considering only saccades where at least one (all 10,192 saccades), two (7,948 saccades), three (5,565 saccades), or all four (2,951 saccades) humans agreed on a common gaze location, for the static **(a)** and dynamic metrics **(b)**. Static metrics improved substantially when progressively focusing onto saccades with stronger inter-observer agreement (average slope $0.56 \pm 0.37$ percent $KL$ score units per 1,000 pruned saccades). Hence, when humans agreed on a location, they also tended to be more reliably predicted by the metrics. Furthermore, dynamic metrics improved $4.5$ times more steeply (slope $2.44 \pm 0.37$), suggesting a stronger role of dynamic events in attracting human attention. Surprising events were significantly the strongest ($t$-tests for equality of $KL$ scores between surprise and other metrics, $p < 10^{-100}$).

## V. Discussion: Bits and Wows

Much research has employed Shannon entropy and other inherently static measures of information to analyze neuronal processing and behavior [25], [26], [27]. Previous research has shown with either static scenes or dynamic synthetic stimuli that humans preferentially fixate regions of high entropy [7], contrast [6], saliency [9], flicker [10], or motion [11]. However, even casual observation suggests that these alone fail to capture neural response transients and adaptation.

Here, by explicitly considering internal models and changing beliefs and developing new tools to quantify bottom-up influences on attention in dynamic scenes, we find that humans fixate surprising locations even more reliably, making surprise the strongest of all algorithmic metrics tested. These conclusions were made possible by developing new tools to quantify what attracts human gaze over space and time in dynamic natural scenes. Surprise explains best where humans look when considering all saccades, and even more so when restricting the analysis to only those saccades for which human observers tended to agree. Hence, surprise represents an inexpensive, easily computable shortcut to important information [28].

Shannon's theory of communication focuses on "reproducing at one point either exactly or approximately a message selected at another point [29]." Although eminently successful for the development of modern computer and telecommunication technologies, such definition does not capture subjective and semantic aspects of information. This is clearly exemplified by the paradox that random snow, the most boring of all television programs, carries the largest amount of Shannon information. At onset, snow carries both surprise and information. Indeed snow may signal storm, earthquake, toddler's curiosity, or military putsch. After a moment, however, the viewer's model of the image shifts towards a random pixel model, prior and posterior become identical, and additional snow frames carry zero surprise albeit megabytes of Shannon information. Indeed, our video clips, presumably of interest to millions of television watchers and gamers, carried $\approx 0.3$ megabytes of Shannon information per second once compressed to constant-quality MPEG4. This was significantly lower than $\approx 5.0$ megabytes/s for matched MPEG4-compressed uniform snow clips, probably of interest only to a few engineers [30]. Thus entropy and surprise are two dual facets of information. Shannon's entropy

$$H(\mathcal{D}) = -\int_{\mathcal{D}} P(D) \log P(D) dD, \tag{7}$$

requires integration over the space of data. Suprise instead requires integration over the space of models.

Under a small set of axioms [31], [32], [33] the Bayesian definition of probability provides the only consistent approach to inference and learning. Likewise, in the same framework, surprise is the only consistent measure of novelty. Other measures of novelty, for instance in terms of outliers [34], can be viewed as approximations to surprise which can be flawed in some extreme cases.

In the absence of quantitative tools to measure surprise, most experimental and modeling work to date has adopted the approximation that novel events are surprising, and has focused on experimental scenarios which are simple enough to ensure an overlap between informal notions of novelty and surprise: for example, a stimulus is novel during testing if it has not been seen during training [38]. Our definition opens new avenues for more sophisticated experiments, where surprise elicited by different stimuli can be precisely compared and calibrated, yielding predictions at the single-unit as well as behavioral levels.

The definition of surprise — as the distance between the posterior and prior distributions of beliefs over models — is entirely general and readily applicable to the analysis of auditory, olfactory, gustatory, or somatosensory data. While here we have focused on behavior rather than detailed biophysical implementation, it is worth noting that detecting surprise in neural spike trains does not require semantic understanding of the data carried by the spike trains, and thus could provide guiding signals during self-organization and development of sensory areas. At higher processing levels, top-down cues and task demands are known to combine with stimulus novelty in capturing attention and triggering learning [36], [39], ideas which may now be formalized and quantified in terms of priors, posteriors, and surprise. Surprise, indeed, inherently depends on uncertainty and on prior beliefs. Hence surprise theory can further be tested and utilized in experiments where the prior is biased, for example by top-down instructions or prior exposures to stimuli [39]. In addition, simple surprise-based behavioral measures such as the eye-tracking one used here may prove useful for early diagnostic of human conditions including autism and attention-deficit hyperactive disorder, as well as for quantitative comparison between humans and animals which may have lower or different priors, including monkeys, frogs, and flies.

Beyond sensory biology, computable surprise could guide the development of data mining and compression systems (giving more bits to surprising regions of interest), to find surprising agents in crowds, surprising sentences in books or speeches, surprising sequences in genomes, surprising medical symptoms, surprising odors in airport luggage racks, surprising documents on the world-wide-web, or to design surprising advertisements.

## VI. Appendix: Additional Experimental Details

**Methods:** Subjects were USC students and staff, three females and five males, ages 23-32, normal or corrected-to-normal vision. Informed consent was obtained from all subjects prior to the experiments. Each subject watched a subset of the collection of videoclips, so that eye movement traces from four distinct subjects were obtained for each clip.

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 $\tau = 500ms$. The random sampling process was repeated 100 times, yielding the (very small) error bars of the random histograms of Figures 1 and 2.

**Human-derived metric:** A Gaussian blob with $\sigma = 3$ master map pixels was continuously painted at each of the eye positions of the three observers other than that under test, with some forgetting provided by the master map's temporal low-pass filtering. High metric responses were hence sampled if and only if a saccade of the observer under test was aimed to approximately a location where other observer(s) were currently looking. Because this metric is not predictive like the others, sampling was performed when a saccade ended (and other humans were expected to also be reaching the endpoint) rather than when it started (and other humans possibly also started).

**Static metrics:** The variance metric computes local variance of pixel luminance within $16 \times 16$ image patches [6], [21]. The Shannon entropy metric computes the entropy of the local histogram of grey-levels in $16 \times 16$ image patches [7]. The DCT-based (Discrete Cosine Transform) information metric similarly computes in image patches the number of DCT coefficients above detection threshold, for the luminance and two chrominance channels [37].

The colour, intensity and orientation contrast metrics are derived from reduced versions of our previously proposed bottom-up saliency metric [37], [3]. They compute local contrast in each feature dimension using difference-of-Gaussian centre-surround contrast detectors operating at six different spatial scales.

**Dynamic metrics:** The flicker and motion metrics rely on the same centre-surround architecture as the color, intensity and orientation metrics. The saliency metric combines intensity contrast (six feature maps), red/green and blue/yellow colour opponencies (12 maps), four orientation contrasts (24 maps), temporal onset/offset (six maps) and motion energy in four directions (24 maps), totalling 72 feature maps. Central to the saliency and the color, intensity, orientation, flicker and motion metrics is non-classical spatial competition for saliency [37], by which distant active locations in each feature map inhibit each other, giving rise to pop-out and attentional capture [22]. Thus, these metrics are not necessarily attracted to information-rich image regions, as many highly informative regions will be discarded if they resemble their neighbours, yielding sparser maps than the contrast, entropy and DCT-based information metrics which are purely local.

The surprise metric retains the 72 raw feature detection mechanisms of the saliency metric (but without the non-classical competition for saliency), and attaches local surprise detectors to each location in each of the 72 feature maps. Surprise detectors compute both local temporal surprise (or local temporal novelty) and spatial surprise (or spatial saliency). Further details on the implementation of this metric have been described previously [40], and are also available with our source code, distributed freely at http://iLab.usc.edu/toolkit/.

## ACKNOWLEDGEMENT

Supported by HFSP and NGA (L.I.), NIH and NSF (P.B.). We thank UCI's Institute for Genomics and Bioinformatics (www.igb.uci.edu) and USC's Center High Performance Computing and Communications (www.usc.edu/hpcc) for access to their computing clusters.

## REFERENCES

[1] W. James, *The Principles of Psychology* (Harvard University Press, Cambridge, MA, 1890/1981).

[2] J. K. Tsotsos, *Behav Brain Sci* **14**, 506 (1991).

[3] L. Itti, C. Koch, *Nat Rev Neurosci* **2**, 194 (2001).

[4] B. C. Motter, E. J. Belky, *Vision Res* **38**, 1805 (1998).

[5] J. Shen, E. M. Reingold, M. Pomplun, *Perception* **29**, 241 (2000).

[6] P. Reinagel, A. M. Zador, *Network* **10**, 341 (1999).

[7] C. M. Privitera, L. W. Stark, *IEEE Trans Patt Anal Mach Intell* **22**, 970 (2000).

[8] E. Barth, C. Zetzsche, I. Rentschler, *J. Opt. Soc. Am. A Opt. Image. Sci. Vis.* **15**, 1723 (1998).

[9] D. Parkhurst, K. Law, E. Niebur, *Vision Res* **42**, 107 (2002).

[10] J. Theeuwes, *Percept Psychophys* **57**, 637 (1995).

[11] R. A. Abrams, S. E. Christ, *Psychol Sci* **14**, 427 (2003).

[12] S. L. Franconeri, A. Hollingworth, D. J. Simons, *Psychological Science* (in press).

[13] S. Kullback, *Information Theory and Statistics* (Wiley, New York:New York, 1959).

[14] P. Baldi, in *Information, Coding, and Mathematics*, M. Blaum Editor, Kluwer, 1-25, 2002.

[15] P. Baldi, in *Neurobiology of Attention*, L. Itti, G. Rees, and J. Tsotsos Editors, Elsevier, San Diego, CA, 24-28, 2005.

[16] L. D. Brown, *Fundamentals of Statistical Exponential Families* (Institute of Mathematical Statistics, Hayward, CA, 1986).

[17] Saccades were defined by a velocity threshold of $20^\circ/s$ and amplitude threshold of $2^\circ$.

[18] Computing the master maps for all models tested here and evaluating them against human eye movement traces required in excess of seven CPU-months, which we distributed over the high-performance computation clusters available at UCI and USC.

[19] Computing the Kullback-Leibler distance between empirical distributions of metric values at human and random saccade endppoints presents several definitive advantages over simpler scoring schemes used previously. These include its agnosticity to putative mechanisms for generating saccades based on the instantaneous distribution of metric responses over visual space, and its invariance to reparameterizations, so that applying any continuous nonlinearity to master map values would not affect scoring.

[20] R. B. Goldstein, E. Peli, S. Lerner, G. Luo, *J Vision* **4**, 643a (2004).

[21] D. J. Parkhurst, E. Niebur, *Spat Vis* **16**, 125 (2003).

[22] J. Wolfe, *Attention*, H. Pashler, ed. (University College London Press, London, UK, 1998).

[23] Two humans were considered in agreement on an important location at a given time if their sampling apertures (diameter $5.6^\circ$) intersected.

[24] Indeed, it is well known that continuous flicker does not continually attract attention [**?**].

[25] H. B. Barlow, *Proc. Symposium on the Mechanization of Thought Processes*, D. V. Blake, A. M. Utlley, eds. (HM Stationery Office, London, 1959), vol. 2, pp. 537–574.

[26] J. J. Atick, *Network* **3**, 213 (1992).

[27] E. P. Simoncelli, B. A. Olshausen, *Annu Rev Neurosci* **24**, 1193 (2001).

[28] Eearly sensory neurons are known to rapidly adapt and signal novel stimuli more vigorously, in a manner compatible with our surprise metric's behaviour (Müller *et al.*, 1999). Less evidence exists that single neurons may compute variance or Shannon entropy over spatially extended patches of inputs.

[29] C. E. Shannon, *Bell Syst Tech J* **27**, 379 (1948).

[30] This readily suggests a dynamic compression algorithm where fewer bits are used to encode unsurprising regions, and more bits are used to encode surprising regions.

[31] R. T. Cox, *Am. J. Phys.* **14**, 1 (1964).

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

[33] E. T. Jaynes, *Probability Theory. The Logic of Science* (Cambridge University Press, 2003).

[34] M. Markou, S. Singh, *Signal Processing* **83**, 2481 (2003).

[35] N. Friedman, S. Russell, *Annual Conference on Uncertainty in Artificial Intelligence* (1997), pp. 175–181.

[36] C. Ranganath, G. Rainer, *Nat Rev Neurosci* **4**, 193 (2003).

[37] L. Itti, C. Koch, E. Niebur, *IEEE Trans Patt Anal Mach Intell* **20**, 1254 (1998).

[38] J. H. Fecteau and D. P. Munoz, *Nat Rev Neurosci* **4**, 435 (2003).

[39] J. M. Wolfe and T. S. Horowitz, *Nat Rev Neurosci* **5**, 495 (2004).

[40] L. Itti and P. Baldi, *Proceedings of IEEE CVPR*, in press, (2005).