Natural Scene Statistics at the Center of Gaze

Pamela Reinagel† and Anthony M. Zador‡

† Sloan Center for Theoretical Neuroscience
California Institute of Technology
Pasadena, CA 91125

and

‡ Sloan Center for Theoretical Neuroscience
Salk Institute for Biological Sciences
La Jolla, CA 92037

‡ To whom correspondence should be addressed
Tel: (619) 453-4100 x1404
Fax: (619) 450-2172
Email: zador@salk.edu

Humans tend to look at regions of a scene that are interesting or surprising at a cognitive level. While the importance of such top-down factors in guiding the gaze has long been recognized, the role of low-level image properties has not been established. We recorded eye positions of human subjects as they viewed a wide variety of natural images. We report that subjects selectively directed their gaze at regions of the image that have unusual spatial frequency composition, and at regions that are "surprising" according to a local low-level measure. These results suggest that when subjects are free to move their eyes, the distribution of neural responses even at early stages of visual processing could guide eye movements.

The visual world in which humans evolved was highly structured, unlike for example the random static observed on an untuned television monitor. This structure can be characterized at many levels. At a high-level, an image can be decomposed into a collection of objects such as rocks and trees. Alternatively, at a low level, an image could be decomposed into spatial frequency components [1, 2, 3, 4] or wavelets [5]. In contrast to high-level descriptions, these low-level descriptions could plausibly be computed locally at early stages of visual processing.

Low-level analyses reveal robust statistical invariances across natural images, such as a strong correlation between nearby points. The neural encoding of any stimulus ensemble is most efficient if the dynamic range of the response is allocated preferentially to those aspects of the stimuli that vary most. Thus the early visual system can exploit the characteristic structure of images to encode natural visual stimuli efficiently [6, 7, 8, 9]. For example, motion detectors in insects are matched to the speed of flight and thus to the temporal frequencies typically experienced [10].

We hypothesized that the early visual system of humans may be adapted to the statistics of its inputs. To test this idea it is important to know the statistics of natural images; but to stop there would entail a tacit assumption that visual world is sampled uniformly. Humans move their eyes several times a second when looking at a scene, thereby actively selecting visual stimuli for further processing. The portions of a scene that fall on the fovea are sampled at high spatial resolution, and are subject to a disproportionate fraction of subsequent cortical processing. Thus, we set out to determine how voluntary eye movements affect the low-level statistics of images falling on the fovea, and conversely, how low-level image statistics affect where subjects direct or maintain their gaze.

We recorded eye positions [11] from subjects as they viewed black-and-white images [12] presented on a computer monitor. An image from our ensemble, with the

References

[1] Field, D.J. *J. Opt. Soc. Amer. A.* **4**, 2379–2394 (1987).

[2] Ruderman, D.L. & Bialek, W. B. *Phys. Rev. Lett.* **73**, 814–817 (1994).

[3] Tadmor, Y. & Tolhurst, D.J. *Vision Res.* **34**, 541–554 (1994).

[4] van der Schaaf, A. & van Hateren, J.H. *Vision Res.* **28**, 814–817 (1996).

[5] Strang, G. & Nguyen, T. *Wavelets and Filter Banks* (Wellesley-Cambridge Press 1996). Wornell, G.W. *Proc. IEEE* **81**, 1428–1450 (1993).

[6] Atick, J.J. *Network* **3**, 213–251 (1992).

[7] Barlow, H.B. *Neural Computation* **1**, 295–311 (1989).

[8] Bialek, W. & Zee, A. *Phys. Rev. Let.* **61**, 1512–1515 (1988).

[9] Srinivasan, M.V., Laughlin S.B. & Dubs A. *Proc. Royal Soc. London, B* **215**, 427–459 (1982).

[10] O'Carroll, D.C., Bidwell, N.J., Laughlin, S.B., Warrant, E.J. *Nature* **382**, 63–66 (1996).

[11] An RK-416 infrared Pupil Tracking System (Iscan Inc.; Cambridge, MA) was used to record eye position every 20 msec. A bite bar was used to minimize head movement. Subjects viewed the images on a 21 inch monitor at 79 cm; the whole image subtended 28×21 degrees of visual angle (23 pixels/degree). Subjects were instructed to "study the images". Images were presented in seven blocks of eleven images each. Within a block a brief central fixation cue preceded each 10 sec image presentation. Raw eye positions were corrected by linear interpolation using a 5×5 calibration grid presented before each image block. The estimated tracking error was less than 0.5 degrees of visual angle. Analysis shown is based on eye positions from $0.4 - 4$ sec after presentation of the image.

[12] Seventy-seven images were presented to 5 naive subjects. The image ensemble include 69 natural images, of which 38 depicted nature scenes, 17 depicted man-made objects such as building interiors and exteriors, and 14 included animals or humans. The remaining 8 synthetic images included 4 fractal images, 2 white noise images, and 2 phase-randomized images. The latter were obtained from the Fourier transforms of natural images by randomly reassigning the computed phases and then inversing the Fourier transform. All images were 640×480 pixels, with 256 gray levels. The natural images were amateur and professional black-and-white photographs from a variety of sources, scanned on an Scanjet 4c Digitizer (Hewlett Packard). All images were in sharp focus at all depths of field.

[13] Each 64×64 image patch was windowed with a conical window. We normalized each image patch to mean intensity as in [4]. The orientation-averaged power spectrum was

computed from the 2-dimensional Fourier transform. We used image patches of 64×64 pixels to compensate for the windowing; the central 1 degree of visual angle of the patch contributed most of the signal to the power spectrum.

[14] Westheimer, G. *Vision Res* **22**, 157-162 (1982).

[15] Yarbus, A.L. *[Eye movements and vision]* B. Haigh, trans. (New York: Plenum Press 1967).

[16] Loftus, G.R. & Mackworth, N.H. *J. Exp. Psych.: Human Perception & Performance* **4**, 565–572 (1978).

[17] Mackworth, N.H. & Morandi, A.J. *Perception and Psychophysics* **2**, 547–551 (1967).

[18] Shannon, C. *Bell Sys. Tech. Journal* **27**, 379–423 (1948).

[19] Bishop, C.M. *Neural networks for pattern recognition* (Clarendon Press ; New York : Oxford University Press, 1995).

[20] Daugman, J.G. *IEEE Trans. on Biomed. Engin* **36**, 107–114, (1989); D. J. Field, *Neural Computation* **6**, 559–601, (1994)

[21] The coefficients $c^m_{x,y}$ of the Haar wavelet transform of each image were computed using MATLAB (The MathWorks, Inc, Natick, MA). The coefficients are given by $c^m_{x,y} = \int\int I(x,y)\Phi^m_{x,y}(x,y)dxdy$, where m is the detail level, and the basis elements $\Phi^m_{x,y}(x,y)$ are translations and dilations of a mother wavelet $\Phi(x,y)$. For all the analyses shown, the number of detail levels $m = 5$, corresponding to images patches 32×32 pixels, or approximately 1 degree of visual angle.

[22] We computed the cross-entropy from the distribution of wavelet coefficients as follows. Let $P(c)$ denote the probability distribution of coefficients obtained from the wavelet decomposition of a single image, and let $P_s(c)$ and $P_u(c)$ denote the distributions of coefficients obtained from the decomposition of the subject-selected patches and randomly-selected from the same image, respectively. If wavelet coefficients were statistically independent, the Shannon entropy [18] of the entire image is given by $H = -\sum_c P(c) log_2 P(c)$. The quantity $-log_2 P(c)$ is the length (in bits) required to encode the coefficient c, using an optimal encoding scheme for the image, in which less frequent coefficients require logarithmically longer codewords. Because these less frequent coefficients are more "surprising", they are more informative. The cross-entropy of the subject-selected ensemble is defined as $H_s = -\sum_c P_s(c) log_2 P(c)$, and measures the total surprise of the coefficients in the ensemble, relative to the whole image. The cross-entropy of the randomly selected patches is similarly defined as $H_u = -\sum_c P_u(c) log_2 P(c)$; in the limit of infinite sampling and in the absence of edge effects, H_u would be exactly equal to H. The cross-entropy H_s of the subject-selected ensemble can be higher or lower than the Shannon entropy H, depending on whether it includes a disproportionate number of unusual or common coefficients.

[23] Field, D.J. *Neural Computation* **6**, 559–601 (1994).

[24] Olshausen, B.A. & Field, D.J. *Nature* **381**, 607–609 (1996).

[25] Rayner, K. & Duffy, S. *Memory and Cognition* **14**, 191–201 (1986).

[26] Zelinsky, G.J., Rao, R.P.N., Hayhoe, M.M., & Ballard, D.H. *Invest. Ophth. & Vis. Sci.* **37**,64–64 (1996).

[27] Kortum, P.T. & Geisler, W.S. *Invest. Ophth. & Vis. Sci.* **37**,1363–1363 (1996).

[28] J.M. Wolfe *Psych. Bull. Rev.* **1**, 202–238 (1994).

[29] Treisman, A. *Quart. J. Exp. Psych.* **12**, 97–136 (1988).

[30] Mannan, S. Ruddock, K. H. & Wooding, D. S. *Spatial Vision* **9**, 363–386 (1995).

ACKNOWLEDGEMENTS. We thank T. Albright, C. Koch, and T. Sejnowski for comments and suggestions and T. Albright for use of eye tracking equipment. Supported by the HHMI (A.Z.), the NSF Engineering Research Center at The California Institue of Technology (P.R.), and by The Sloan Centers for Theoretical Neuroscience at The Salk Institute (A.Z.) and The California Institue of Technology (P.R.).

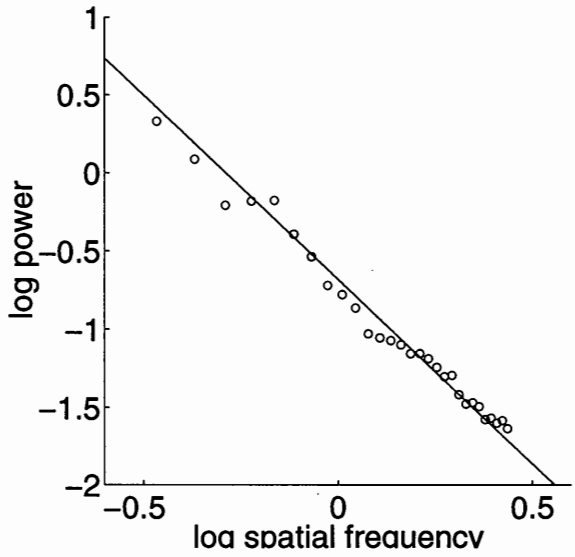
CORRESPONDENCE and requests for materials should be addressed to A.Z. (email: zador@salk.edu).

Fig. 1 A representative image from our ensemble with eye positions of one subject superimposed. Circles indicate the position of the center of gaze recorded at 20 msec intervals.

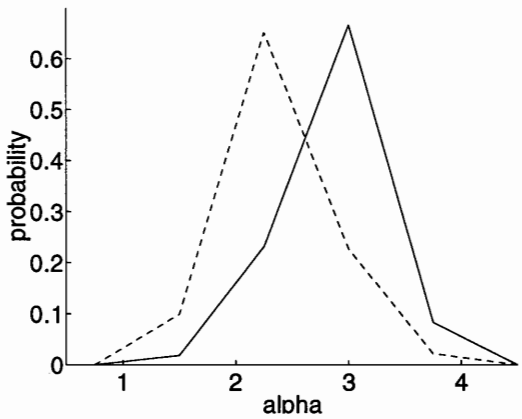
Fig. 2. (a) The orientation-averaged power spectrum of a single 64×64 pixel image patch extracted from this image. The exponent α, where $P(f) \propto 1/f^{\alpha}$, was obtained by fitting a line to the data plotted on a log-log scale; α corresponds to minus the slope of this line. For this image patch, $\alpha = 2.36$. Spatial frequency f is in units of cycles/degree. **(b)** The distribution of α for the subject-selected ensemble $\mathcal{S}$ (*solid*) and randomly-selected ensemble $\mathcal{U}$ (*dashed*) for the image and eye positions in *a*. For the ensemble $\mathcal{S}$ the mean $\bar{\alpha}_S = 2.36$; for the ensemble $\mathcal{U}$ the mean $\bar{\alpha}_U = 2.96$. **(c)** The difference $\bar{\alpha}_S - \bar{\alpha}_U$ between the mean subject-selected ($\bar{\alpha}_S$) and randomly-selected ($\bar{\alpha}_U$) image patches is shown for each image. For this subject $\bar{\alpha}_S > \bar{\alpha}_U$ in $\phi = 75\%$ of images. In 5/5 subjects, the fraction ϕ was significantly greater than 0.5 ($p < 0.01$).

Fig. 3. (a) The distribution of scaled coefficient values for a 5-level Haar wavelet transform of the image $I(x, y)$ in Fig. 1**a**. When scaled by 2^{-m} as shown, the 5 distributions are nearly identical, reflecting the scale-invariance of natural images. **(b)** The distribution of entropy H of level 3 wavelet coefficients in 32×32 pixel image patches for the subject-selected ensemble $\mathcal{S}$ (*solid*) and the randomly-selected ensemble $\mathcal{U}$ (*dashed*) for the image and eye positions in Fig. 1**a**. For the ensemble $\mathcal{S}$ the mean $\bar{H}_S = 4.22$ bits/coefficient; for the ensemble $\mathcal{U}$ the mean $\bar{H}_U = 3.52$ bits/coefficient. **(c)** The difference $\bar{H}_S - \bar{H}_U$ between the mean cross-entropy of level 3 wavelet coefficients between subject-selected and randomly-selected image patches. For this subject, $\bar{H}_S > \bar{H}_U$ in $\phi = 81\%$ of the natural images. In 5/5 subjects, the fraction ϕ was significantly greater than 0.5 ($p < 0.001$).

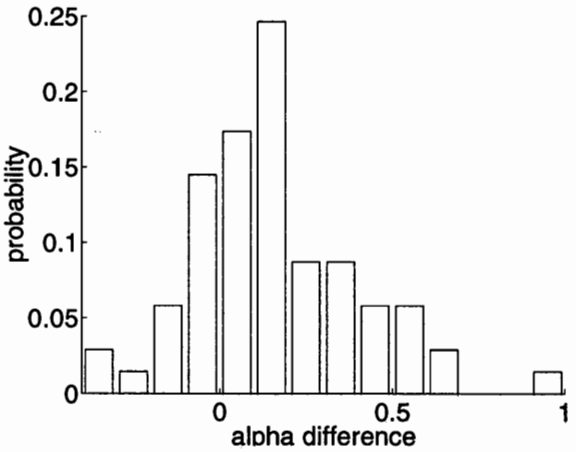
1


log power
log spatial frequency
1
0.5
0
−0.5
−1
−1.5
−2
−0.5
0
0.5

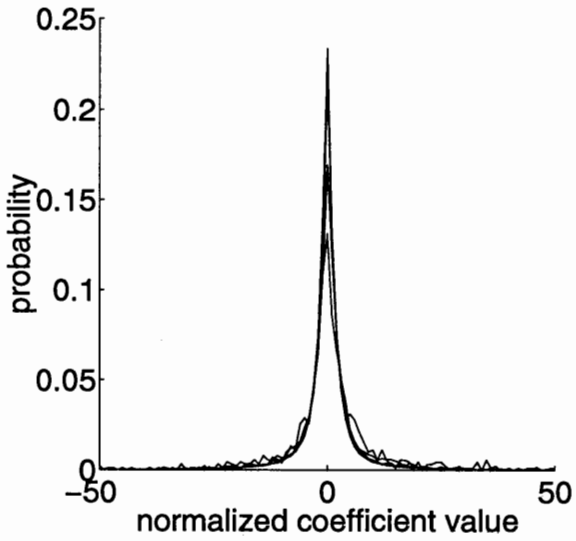
2*a*


probability
alpha
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4

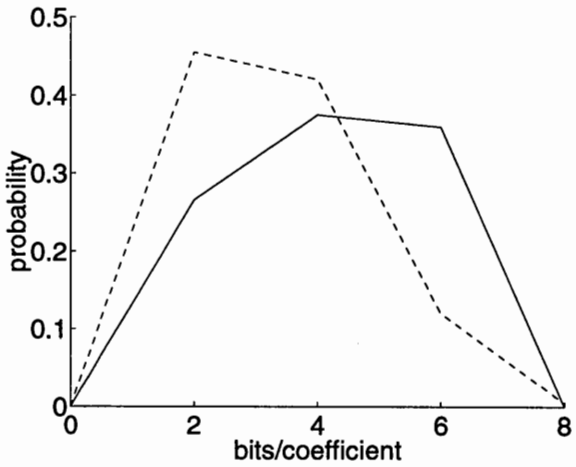
2*b*


0.25
0.2
0.15
0.1
0.05
0
probability
0
0.5
1
alpha difference

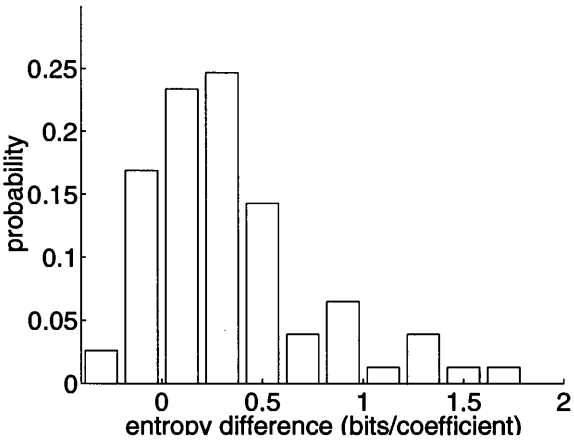
2*c*


0.25
0.2
0.15
0.1
0.05
0
probability
−50
0
50
normalized coefficient value

3*a*


0.5
0.4
0.3
0.2
0.1
0
probability
0
2
4
6
8
bits/coefficient

3*b*


0.25
0.2
0.15
0.1
0.05
0
probability
0
0.5
1
1.5
2
entropy difference (bits/coefficient)

3*c*