Statistics of Natural Images: Scaling in the Woods

Daniel L. Ruderman[1,2] and William Bialek[1]
[1]*NEC Research Institute, 4 Independence Way, Princeton, New Jersey 08540*
[2]*Department of Physics, University of California, Berkeley, California 94720*
(Received 23 March 1994)

We study the statistics of an ensemble of images taken in the woods. Distributions of local quantities such as contrast are scale invariant and have nearly exponential tails. Power spectra exhibit scaling with a nontrivial exponent. These data limit the information content of natural images and point to the importance of gain-control strategies in visual processing.

PACS numbers: 42.30.Yc, 05.70.Jk, 42.66.Lc

Efficient signal processing systems take advantage of statistical structure in their input signals, both to reduce the effects of noise and to generate compact representations of seemingly complex data. Since Barlow's discussion in 1959 [1], many authors have explored the possibility that biological vision systems are designed to exploit the statistics of natural images [2]. It is difficult, however, to compare these ideas with experiment, because we know relatively little about the statistics of natural scenes. The fact that objects can appear on all possible angular scales leads to the hypothesis that natural images should exhibit some form of scaling or self-similarity [3]. Here we analyze an ensemble of images taken in the woods of Hacklebarney State Park in central New Jersey. These provide strong evidence for nontrivial scaling in the sense of statistical mechanics, and we discuss some implications of these results for our understanding of early visual processing.

Imagine that we have taken a set of photographs which are sampled into pixels of angular dimension $l_0 \times l_0$. The scaling hypothesis states that the statistical structure of our ensemble of photographs is independent of the pixel size l_0. In studying images we can test for scaling explicitly by constructing block pixels. In Fig. 1(a) we show an example from our ensemble of natural images. We define the contrast in each pixel to be $\phi(\mathbf{x}) = \ln[I(\mathbf{x})/I_0]$, where I_0 is chosen for each image so that the average contrast is zero; note that ϕ is invariant to overall changes in brightness. To characterize the statistical structure of the images we examine the histogram of contrasts in each pixel, $P(\phi)$, shown in Fig. 1(b). The distribution of contrasts is far from Gaussian, with nearly exponential tails. We now construct 2×2 block pixels, renormalize ϕ so that the root-mean-square contrast is fixed, and look again at the histogram of contrasts in each pixel; the distribution is the same. We continue the blocking procedure, and in each case the distribution of (renormalized) contrast is the same.

The non-Gaussian character of the natural image ensemble can also be seen in the probability distribution for contrast gradients, $|\nabla\phi|$, in Fig. 1(c). Scaling is observed over a range of nearly 10^4 in probability. If $P[\phi(\mathbf{x})]$ were Gaussian, the distribution of $|\nabla\phi|$ would be Rayleigh, and in Fig. 1(c) we see that both small and large gradients are more likely than expected in a Gaussian world. Qualitatively, regions of small gradient are large and connected, interrupted by smaller regions of high gradient. This is similar to turbulent fluid flow, where images of the flow (or of the temperature in thermally driven turbulence) show a concentration of large gradients into small regions. These flows also exhibit strongly non-Gaussian probability distributions, as observed here for natural images [5].

The link between non-Gaussian distributions and the inhomogeneity of gradients can be seen by defining a local variance in the image, for example, the variance of ϕ values in an $N \times N$ block surrounding each point (Fig. 2). The distribution of local variance has a long tail, but if we normalize the deviation of each point from the local ($N \times N$ block) mean by the local standard deviation, then the histogram of pixel values has Gaussian tails (for $N = 5$), and the distribution of gradients in the "variance-normalized" images is almost exactly the Rayleigh distribution. Seemingly featureless regions of the original image reveal their textures in the variance-normalized images, while maps of the local variance reveal "ghosts" of objects.

We have found no local linear transformation on the images which produce Gaussian distributions. In particular, biologically motivated center-surround filtering [6] produces pixel histograms which are even more precisely exponential. Thus while filtering can reduce redundancy by decorrelating neighboring pixels [1] (analogous to the "q-u redundancy in English), these operations cannot remove the redundancy associated with the non-Gaussian distributions of local quantities. In contrast, the nonlinear operation of variance normalization returns a modified image in which local quantities have Gaussian distributions, and hence maximum entropy given their dynamic range.

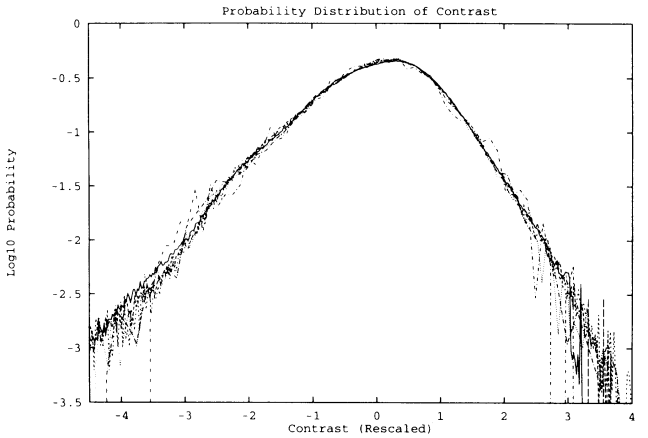
Another way of looking for scale invariance is to measure the power spectrum

$$S_\phi(\mathbf{k}) = \int d^2y \exp[i\mathbf{k}\cdot\mathbf{y}]\langle\phi(\mathbf{x})\phi(\mathbf{x}+\mathbf{y})\rangle, \qquad (1)$$

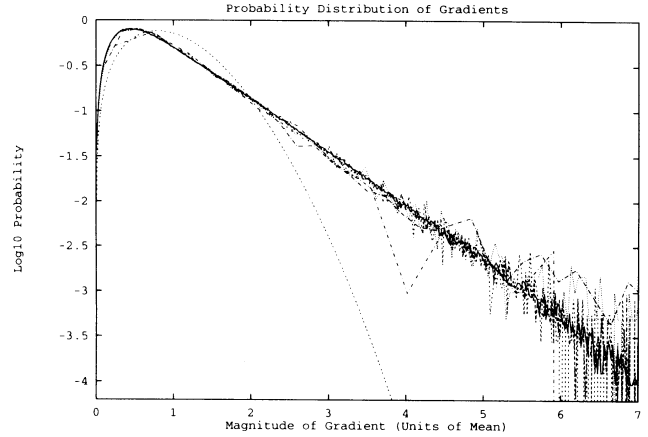
where $\langle\cdots\rangle$ denotes an average over images. We recall

(a)

(b)



(c)



that the power spectrum is a useful characterization only for ensembles which are translation invariant, so that the expectation value $\langle\phi(\mathbf{x})\phi(\mathbf{x}+\mathbf{y})\rangle$ is independent of the central position $\mathbf{x}$. This excludes environments where the horizon is prominent, and we know that organisms which live in such environments have eyes which sample the world in a strongly inhomogeneous manner [7].

The power spectrum, as shown in Fig. 3(a), is fit well by $S_\phi(k) = A/k^{2-\eta}$ over a range of eight octaves in spatial frequency k, which quantifies our intuition that structures occur on all possible scales. In a world with $\eta = 0$ each octave of spatial frequency contains the same amount of power [3]; with $\eta > 0$ there is proportionately more power in the short distance details. We find that for images in the woods $\eta = 0.19 \pm 0.01$.

The power spectrum by itself does not tell us very much about the statistics of natural images; rather it should be viewed as confirming the scaling which we observed in the contrast histograms. A Gaussian ensemble of images with the observed $S_\phi(k)$ is the maximum entropy ensemble consistent with the power spectrum of natural images, and hence a visual system which views this Gaussian ensemble collects more visual information than one which views the real world [8].

We imagine that an animal views the world through a lattice of receptor cells, each of which has independent contrast noise of variance σ^2, and that the optics of the eye consist of an ideal lens with a cutoff at spatial frequency k_c which matches the Nyquist frequency of the receptor lattice. Then by using the maximum entropy property of the Gaussian distribution we can show [9] that the information which the receptor array provides about a single snapshot is

$$I \leq \frac{\pi NG}{k_c^2}\int \frac{d^2k}{(2\pi)^2}\log_2\left[1+\frac{(1-k/k_c)^2}{2\pi G\sigma^2}k_c^2 S_\phi(k)\right], \quad (2)$$

where N is the total number of receptors and G is a geometrical factor; $G = \pi/2$ for a square lattice. Because of scale invariance, the noise level σ, the Nyquist frequency

FIG. 1. An example (a) of the images [4] in our ensemble, which consists of 45 images at focal length 15 mm and 25 images at 80 mm. We make no attempt to correct for limitations of the optics or camera noise, which are noticeable in the power spectra [Fig. 3(a)]. Probability distributions are shown only for the data at 15 mm. (b) Distribution of contrast ϕ, averaged over $N \times N$ pixel regions and normalized to unit variance. We see that distributions are identical for $N = 1, 2, 4, \ldots, 32$, with nearly exponential tails. (c) Distribution of magnitudes of the gradient, $|\nabla\phi|$. We define the gradient in discrete images simply by computing differences among neighboring pixels; to study scaling we first average $N \times N$, then apply the same procedure and normalize to unit mean. The tail of the distribution is quite precisely exponential, and contrasts strongly with the Rayleigh distribution expected for a Gaussian world.

k_c, and the strength of the power spectrum A can be combined into the signal-to-noise ratio (SNR) in a single receptor cell,

$$R_{\mathrm{SNR}} = \frac{1}{\sigma^2}\int \frac{d^2k}{(2\pi)^2}(1-k/k_c)^2 S_\phi(k), \tag{3}$$

$$I \le \frac{1}{2}NG\int_0^1 dx\, x \log_2\left[1+\frac{\eta(\eta+1)(\eta+2)}{2G}R_{\mathrm{SNR}}\frac{(1-x)^2}{x^{2-\eta}}\right]. \tag{4}$$

Figure 3(b) shows the available information I over the range of SNRs relevant to the primate fovea. We see that the information is of order 1 bit per receptor cell, or less since we have computed an upper bound. This is less than half what would be available from an array of independent cells, corresponding to a large redundancy.

Although it is well known that intensity histograms of individual images are non-Gaussian, and that individual scenes have roughly $1/k^2$ power spectra, there is also tremendous variability in these data from image to image [10]. We find, in contrast, that an *ensemble* of natural scenes has highly robust statistical features. The observation of precisely scale-invariant exponential tails in the distribution of contrast gradients seems especially significant. Scale-invariant correlations strongly limit the amount of information available in a single scene, while recent measurements indicate that sensory neurons can transmit much more information than previously thought [11]. Together these observations suggest that the optic nerve may be able to transmit all of the information provided by the photoreceptor array. Variance normalization seems crucial to an efficient representation of image data, and there may be a connection between variance normalization as defined here and the various types of contrast gain control observed throughout the visual pathways [12].

We thank H. B. Barlow, A. J. Libchaber, and R. R. de Ruyter van Steveninck for many helpful discussions, M. Potters and A. Schweitzer for their help in estimating receptor SNRs, and B. Gianulis and A. Schweitzer for

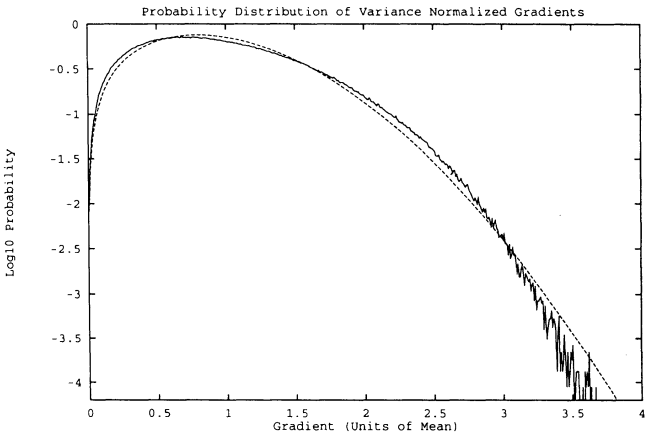


FIG. 2. The probability distribution of gradients $|\nabla\psi(x)|$ in the variance-normalized images, superposed with the Rayleigh distribution expected if $P[\psi(x)]$ is Gaussian. At each pixel x we define the local mean $\bar{\phi}_N(x)$ and local variance $\sigma_N^2(x)$ of the contrast values in the $N \times N$ pixels with x at the center. The variance-normalized image maps the deviation of $\phi(x)$ from the local mean in units of the local standard deviation, that is, $\psi(x) = [\phi(x) - \bar{\phi}_N(x)]/\sigma_N(x)$.

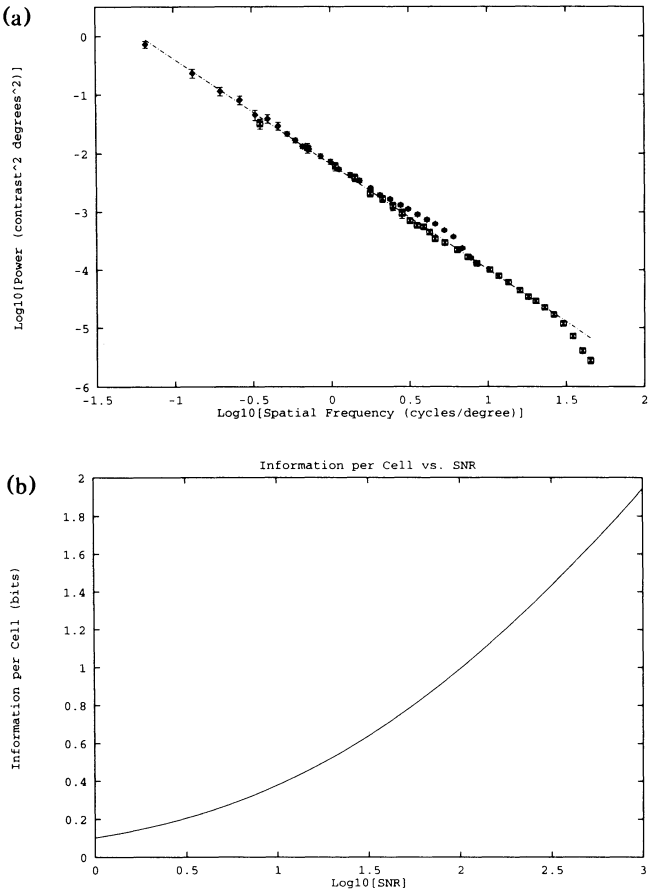


FIG. 3. Power spectra (a) are computed by Fourier transforming each image, taking the absolute square of each Fourier component, then averaging over the ensemble; spectra are averaged also over orientation, so that there is just one spatial frequency variable k. Overlapping data correspond to images collected at different focal lengths. The fitted line is $S_\phi(k) = A/k^{2-\eta}$ with $\eta = 0.19 \pm 0.01$ and $A = 6.47 \times 10^{-3}$ $(\mathrm{deg})^{0.19}$. Information per receptor cell (b) is bounded by Eq. (4). The range of signal-to-noise ratios shown corresponds to our best estimates for cells in primate fovea [13].

their assistance in image acquisition. Work in Berkeley was supported in part by a fellowship from the John and Fannie Hertz Foundation to D.L.R.

[1] H. B. Barlow, in *Sensory Communication*, edited by W. Rosenblith (MIT Press, Cambridge, MA, 1961), pp. 217–234.

[2] For reviews see the articles by J. J. Atick and W. Bialek, in *Princeton Lectures on Biophysics*, edited by W. Bialek (World Scientific, Singapore, 1992).

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

[4] We used a Sony Mavica MVC-5000 2-CCD still video camera fitted with a 9.5–123.5 mm zoom lens at $f/5.6$. Images were taken at two focal lengths in an effort to span a wide range of spatial frequencies and stored on analog still video floppy disks in "frame" mode, then extracted using an MVR-6500 playback unit whose RGB outputs fed into a Videolab board on a Silicon Graphics Crimson workstation; this delivers R, G, and B signals quantized to 8 bits. We convert to grey scale images by computing the standard CIE luminance as $Y = 0.59G+0.30R+0.11B$. Signals were averaged over 32 frame capture runs to reduce the effects of playback noise and calibrated against grey cards of known reflectance. Images are then restricted to the central 256×256 pixels of the frame; the angular spacing of pixels was calibrated for each focal length.

[5] C. W. van Atta and W. Y. Chen, J. Fluid Mech. **44**, 145–159 (1970); R. Antonia, E. Hopfinger, Y. Gagne, and F. Anselmet, Phys. Rev. A **30**, 2704–2707 (1984); F. Heslot, B. Castaing, and A. Libchaber, Phys. Rev. A **36**, 5870–5873 (1987).

[6] H. B. Barlow, J. Physiol. **119**, 69–88 (1953); S. W. Kuffler, J. Neurophys. **16**, 37–68 (1953).

[7] M. F. Land, in *Vision: Coding and Efficiency*, edited by C. Blakemore (Cambridge University Press, Cambridge, United Kingdom, 1990), pp. 55–64.

[8] C. E. Shannon, Proc. IRE **37**, 10–21 (1949); L. Brillouin, *Science and Information Theory* (Academic Press, New York, 1962).

[9] D. L. Ruderman, dissertation, University of California at Berkeley, 1993.

[10] D. J. Tolhurst, Y. Tadmor, and T. Chao, Opthal. Physiol. Opt. **12**, 229–232 (1992).

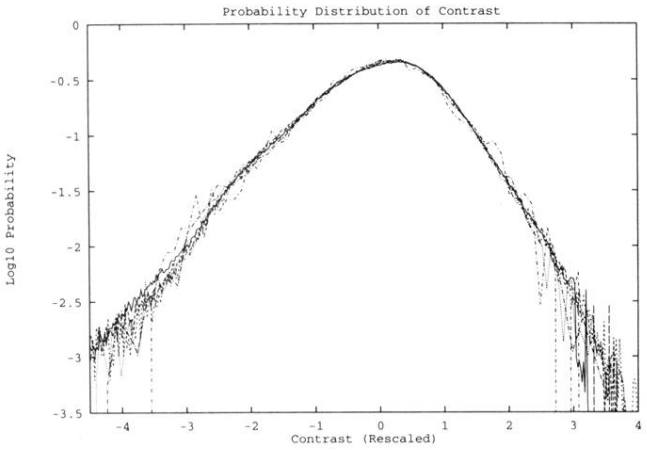
[11] F. Rieke, D. Warland, and W. Bialek, Europhys. Lett. **22**, 151–156 (1993).

[12] C. Enroth-Cugell and R. M. Shapley, Prog. Retinal Res. **3**, 263–346 (1984); I. Ohzawa, G. Sclar, and R. D. Freeman, J. Neurophys. **54**, 651–667 (1985); A. B. Bonds, Vis. Neurosci. **2**, 41–55 (1991); E. A. Benardete, E. Kaplan, and B. W. Knight, Vis. Neurosci. **8**, 483–486 (1992).

[13] Data collected by M. F. Land, Handbk. Sens. Physiol. **VII/6B**, 471–592 (1981), allow us to estimate the photon counting rate for single foveal cones looking at a large white card. If the mean reflectance of the environment is half that of the white card, and half the photons are lost in transmission through the eye, then we find counting rates from $R = 400\ \mathrm{s}^{-1}$ (at sunset with an 8 mm pupil) to $R = 5 \times 10^5\ \mathrm{s}^{-1}$ (bright sunlight with a 2 mm pupil). The signal-to-noise ratio (SNR) in one cell is $R_{\mathrm{SNR}} = RC^2\tau/F$, where C is the root-mean-square contrast of images as seen though the receptor aperture, τ is the integration time, and F is the factor by which the receptor cell noise power exceeds photon shot noise. Our data, as well as that of Laughlin [2], indicate $C \sim 0.3$, while for foveal cones $\tau \sim 0.05$–0.1 s and $F \sim 2$–3 [J. L. Schnapf, B. J. Nunn, M. Meister, and D. A. Baylor, J. Physiol. **427**, 681–713 (1990)]. Putting all the factors together we find that $1 < R_{\mathrm{SNR}} < 10^3$.

(a)

(b)



(c)

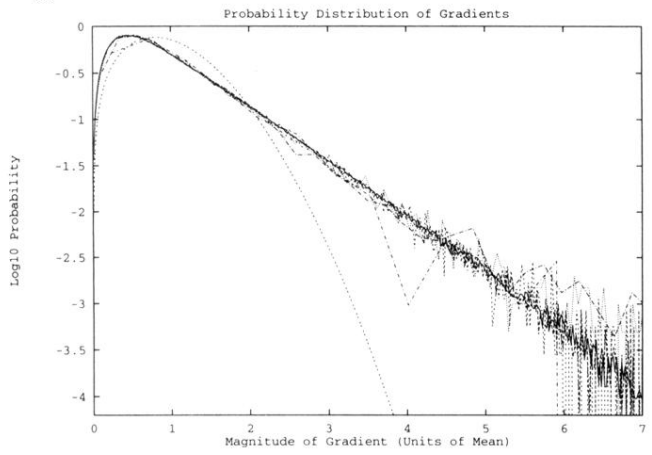


FIG. 1. An example (a) of the images [4] in our ensemble, which consists of 45 images at focal length 15 mm and 25 images at 80 mm. We make no attempt to correct for limitations of the optics or camera noise, which are noticeable in the power spectra [Fig. 3(a)]. Probability distributions are shown only for the data at 15 mm. (b) Distribution of contrast ϕ, averaged over $N \times N$ pixel regions and normalized to unit variance. We see that distributions are identical for $N = 1, 2, 4, \ldots, 32$, with nearly exponential tails. (c) Distribution of magnitudes of the gradient, $\mid \nabla\phi \mid$. We define the gradient in discrete images simply by computing differences among neighboring pixels; to study scaling we first average $N \times N$, then apply the same procedure and normalize to unit mean. The tail of the distribution is quite precisely exponential, and contrasts strongly with the Rayleigh distribution expected for a Gaussian world.