Estimates of the information content and dimensionality of natural scenes from proximity distributions

Damon M. Chandler

School of Electrical and Computer Engineering, Oklahoma State University, Stillwater, Olahoma 74078, USA

David J. Field

Department of Psychology, Cornell University, Ithaca, New York 14853, USA

Received May 30, 2006; accepted September 29, 2006;
posted October 30, 2006 (Doc. ID 71420); published March 14, 2007

Natural scenes, like most all natural data sets, show considerable redundancy. Although many forms of redundancy have been investigated (e.g., pixel distributions, power spectra, contour relationships, etc.), estimates of the true entropy of natural scenes have been largely considered intractable. We describe a technique for estimating the entropy and relative dimensionality of image patches based on a function we call the proximity distribution (a nearest-neighbor technique). The advantage of this function over simple statistics such as the power spectrum is that the proximity distribution is dependent on all forms of redundancy. We demonstrate that this function can be used to estimate the entropy (redundancy) of 3×3 patches of known entropy as well as 8×8 patches of Gaussian white noise, natural scenes, and noise with the same power spectrum as natural scenes. The techniques are based on assumptions regarding the intrinsic dimensionality of the data, and although the estimates depend on an extrapolation model for images larger than 3×3, we argue that this approach provides the best current estimates of the entropy and compressibility of natural-scene patches and that it provides insights into the efficiency of any coding strategy that aims to reduce redundancy. We show that the sample of 8×8 patches of natural scenes used in this study has less than half the entropy of 8×8 white noise and less than 60% of the entropy of noise with the same power spectrum. In addition, given a finite number of samples ($<2^{20}$) drawn randomly from the space of 8×8 patches, the subspace of 8×8 natural-scene patches shows a dimensionality that depends on the sampling density and that for low densities is significantly lower dimensional than the space of 8×8 patches of white noise and noise with the same power spectrum.
© 2007 Optical Society of America
OCIS codes: 330.1880, 330.1800, 330.5510, 330.5020, 100.7410.

1. INTRODUCTION

To be efficient, any coding strategy must take into account the statistical redundancy of the signals that are to be encoded. Whether the purpose is to compress an image or to encode an image to facilitate recognition, it can be argued that one must take advantage of the redundant structure of the data. Recent studies of both the visual and auditory systems of vertebrates have argued that these sensory systems make use of the statistical redundancy of natural signals in an attempt to maximize coding efficiency[1–4] (see Ref. 5 for a review). However, in general, measuring the true entropy of a signal class has proved computationally intractable for all but extremely simple data sets. Without knowledge of this redundancy, it remains an open question of how the absolute efficiency of these sensory systems, or of any compression system, should be quantified.

Natural scenes have been studied extensively over the past two decades, and these studies have revealed that such images have a large number of statistical regularities. Kersten[3] was able to provide an upper bound on the entropy of coarse quantized images based on the ability of human observers to guess the values of missing pixels. However, there was no assumption that this approach converged on the true entropy. A variety of other efforts have measured particular forms of redundancy, including pairwise statistics as described by the power spectra and autocorrelation function (see Ref. 1) as well as a variety of other structures, including the contour structure (e.g., Ref. 6), the pairwise relations between nonlinear transforms of the image,[7,5] and the low-order pixel statistics.[8] Lee *et al.*[9] provided a detailed analysis of the statistical structure of 3×3 high-contrast patches of natural scenes. Although they did not provide a measure of entropy, they demonstrated that most of their natural-scene patches occupied only a small fraction of the measured space.

Sparse coding techniques (e.g., Ref. 10) and related independent components analysis (ICA) techniques (e.g., Ref. 11) search for solutions that attempt to minimize the dependencies between basis vectors. If such dependencies could be removed, then the response histograms of the vectors (the marginals) could be used to determine the entropy. However, despite the name, the independent components produced by ICA are far from independent. Similarly, compression techniques such as Joint Photographic Experts Group (JPEG)[12] and JPEG-2000[13] employ a discrete cosine transform (DCT) and discrete wavelet transform (DWT), respectively, to attempt to minimize depen-

dencies. The DCT–DWT basis coefficients are then processed by a Huffman or arithmetic encoding stage that attempts to remove redundancy and thus yield a highly compressed stream. Indeed, one can provide an estimate of entropy based on the average bit rate of this compressed stream for a particular class of input. However, such an approach assumes that the compression strategy is ideal, and thus the compressed stream is maximally compressed. In reality, the basis coefficients show marked statistical dependencies across space, scale, and orientation; and the majority of encoders cannot take into account all of these dependencies. As a result, current compression algorithms provide only an upper bound on the true image entropy.

This paper describes a technique to estimate entropy of a complex data set and applies this estimate to natural scenes. Although we focus on natural scenes, we emphasize that the techniques described here are by no means limited to visual signals; the methodology can be applied to any data that behave according to a specific assumption (described shortly).

The major difficulty in computing entropy is that the standard approach generally requires knowledge of the full probability distribution from which the data are realized. Consider, for example, a source that emits 8×8 pixel images $\mathbf{X}=[X_1,X_2,\ldots,X_{64}]$ in which each pixel X_i takes on one of $l=256$ shades of gray. In this case, there are $l^{64}=256^{64}=2^{512}$ possible 8×8 pixel patterns (equivalent to approximately 10^{154} or 10^{54} googol of images and roughly 10^{69} times the estimated 10^{85} estimated particles in the universe). To directly compute the entropy of a set of 8×8 natural images via the standard entropy equation, one must therefore obtain enough images to determine the probability distribution $p_{\mathbf{x}}$ over all 2^{512} images and then use this probability distribution to calculate the entropy.

In most cases, estimates of the entropy consider only the first-and sometimes second-order entropy of a particular population. Fourier spectral analysis, in particular, has proved useful for analyzing pairwise pixel-value relationships and has given rise to well-accepted properties such as the $1/f^2$ power spectrum (f=spatial frequency) of natural scenes.[1,14,15] In addition, marginal probability distributions of DCT and DWT coefficients are typically well modeled by using a leptokurtotic generalized Gaussian density, which has served as a cornerstone in the design of quantization and rate-control mechanisms of modern image compression standards.[12,13,16] Indeed, several investigators have shown correspondences between cortical simple-cell receptive fields and the basis functions achieved when one attempts to jointly optimize kurtosis–statistical independence and reconstruction accuracy. However, regardless of the (linear) basis set used to represent the data, unless the basis coefficients are truly independent (i.e., the joint distribution can be factorized into a product of marginal distributions), computing the redundancy of the data based on these marginals will lead to an overestimate of entropy (underestimate of redundancy). Although attempts have been made to model the dependencies that exist between basis coefficients,[4,6,7,17] the somewhat intractable combinatorics involved in such an approach limits the numbers of coefficients that can be used to derive the joint distributions. Indeed, we cannot determine the efficiency of any particular coding or compression algorithm without knowing the true entropy. And, without this estimate, we cannot determine how much of the redundancy has been exploited by any particular coding or compression algorithm.

In this paper, we take an alternative approach to estimating the redundancy of natural scenes that does not require a direct computation of the probability distribution of the data. Instead, the technique we employ borrows heavily from nearest-neighbor-based techniques that have previously been used to estimate entropy of relatively low-dimensional data.[18–20] We note the fact that images drawn from the natural environment are not random patterns; rather, natural scenes typically occupy a subspace of the space of all possible images.[1] The redundancy of the data is determined by the *size* of this subspace[21] (see also Ref. 22, Theorems 3.1.1 and 15.7.1; Ref. 23). Accordingly, we apply nearest-neighbor-based techniques to estimate the relative density of the space of natural scenes by measuring the distances between images as the sample size is increased. We extend the previous methodology to data with larger dimensionality and use this to calculate two properties of images:

1. The *entropy*, which specifies the effective (log) size of the subspace;
2. The *relative dimensionality* (RD), which specifies the dimensionality that the subspace appears to have given a limited number of samples.

This entropy measure can be likened to a kind of "reverse birthday problem." In the birthday problem, one estimates the probability (p) that two people have the same birthday given a group of people of size N and $l=365$ possible birthdays. With $N>23$, the probability $p>0.5$ that any two people will have the same birthday[24] (see also Ref. 25). In the reverse problem, one estimates the number of birthdays l from the probability of obtaining a pair of matching birthdays given a group of size N. This general approach has a long history and was used as far back as Ref. 26 to estimate the population of fish in a lake from samples taken from the lake. For our purposes, the argument is that the relative probability of co-occurrences can provide not only an estimate of the size of the population but also the entropy of the population. Indeed, the approach reveals the size of the population only if one has a known distribution (e.g., see Ref. 27); however, without knowledge of the distribution, the approach (as we will argue) can still provide a measure of entropy.

Furthermore, extending the reverse birthday problem by relaxing the perfect-match criterion to a match of within D days requires sampling only $[1.2\sqrt{365/(2D+1)}]=24$ people[24] to estimate the number of birthdays. In general, given N samples with sufficiently large N and two average nearest-neighbor distances D_A and D_B for data sets A and B, respectively, if $D_A>D_B$ we expect the entropy of A to be greater than the entropy of B. This is the basic technique that we employ: Given samples from a data set, we estimate the entropy of the data based on nearest-neighbor matches in which D is defined as the

Euclidean distance. This process is illustrated in Fig. 1 for samples consisting of natural-scene patches.

As shown in Fig. 1, images from a given data set were randomly divided into two groups: Group $\mathcal{T}$, which consisted of images whose patches served as the to-be-matched "target" samples, and group $\mathcal{N}$, which consisted of images whose patches served as the target's neighbors. The images were divided into $r \times r$ patches, and then for each target patch in group $\mathcal{T}$, an exhaustive search was performed to find the corresponding patch in group $\mathcal{N}$ closest in Euclidean distance to the target patch.

Following from the work by Kozachenko and Leonenko[18] and Victor,[19] we introduce a function that we call the "proximity distribution," which specifies the average (log) nearest-neighbor distance as a function of the number of samples (see Figs. 5, 6, 9, and 10 later in this paper). Our primary assumption is that given a sufficient number of samples, the proximity distribution behaves as a linear function of the (log) number of samples (i.e., the function has a fixed slope); this assumption holds for any data set that is subject to noise (e.g., digitized natural scenes). Thus, with a sufficient number of samples, Kozachenko and Leonenko[18] and Victor[19] argue that the proximity distribution can lead to an estimate of the entropy of the data. Even in cases in which only a portion of the proximity distribution function can be measured, we argue that rational extrapolations can be made that allow a reasonable estimate of entropy and that comparisons between entropy estimates of different image types (e.g., natural scenes versus noise with a $1/f^2$ power spectrum) can provide insights into the contributions of various forms of redundancy to the entropy.

As we demonstrate in Subsection 2.D, nearest-neighbor-based techniques can estimate the entropy of 3×3 natural images using fewer than $2^{18}=262,144$ samples. Furthermore, we present in Section 4 extrapolations of the proximity distribution functions that can be used to estimate entropy of 8×8 images using only 2^{18} samples. We demonstrate that this approach estimates the entropy of 8×8 patches drawn from natural scenes to be less than half the entropy of 8×8 patches of Gaussian white noise and less than 60% of the entropy of noise with the same power spectrum as natural scenes.

In addition to estimating entropy, there exists a wide body of research developed to estimate the dimensionality of a data set[28–34] (see Ref. 35 for a review). Examples of such techniques include projection-based dimensionality-reduction methods such as principal components analysis, nonlinear methods based on local topology such as Isomap[33] and locally linear embedding (LLE),[34] and a variety of techniques based on nearest neighbors.[28–32] It should be emphasized that the definition of "dimension" varies in the literature; a number of dimensions have been reported, including correlation dimension, Hausdorff

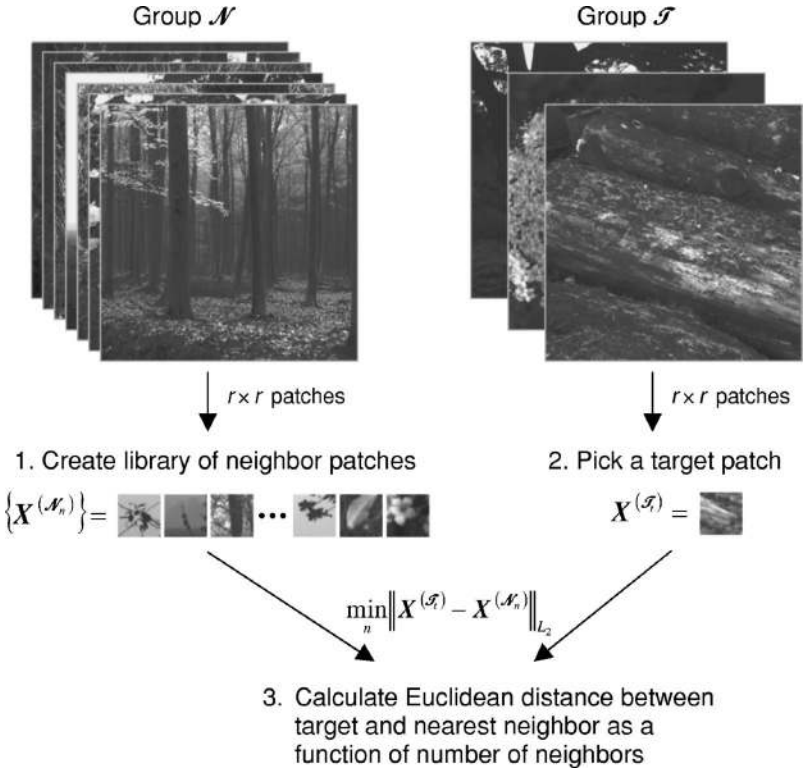

Fig. 1. Diagram of the procedure used in the experiments. Images from a given class were randomly divided into two groups: Group $\mathcal{T}$ containing the to-be-matched "target" samples, and group $\mathcal{N}$ containing the samples from the population. Patches of size $r \times r$ pixels were then extracted from the images in a nonoverlapping fashion. For each target patch in group $\mathcal{T}$, an exhaustive, brute-force search procedure was performed to find the patch in group $\mathcal{N}$ with the minimum Euclidean distance to the target patch (minimum L_2-norm of the difference). The average log nearest-neighbor distance was then estimated by computing the sample mean of the minimum Euclidean distances over all target patches; this process was then repeated for increasing numbers of samples to compute the average log nearest-neighbor distance as a function of the number of samples (the proximity distribution). See Figs. 5, 6, 9, and 10 later in this paper for examples of proximity distribution functions.

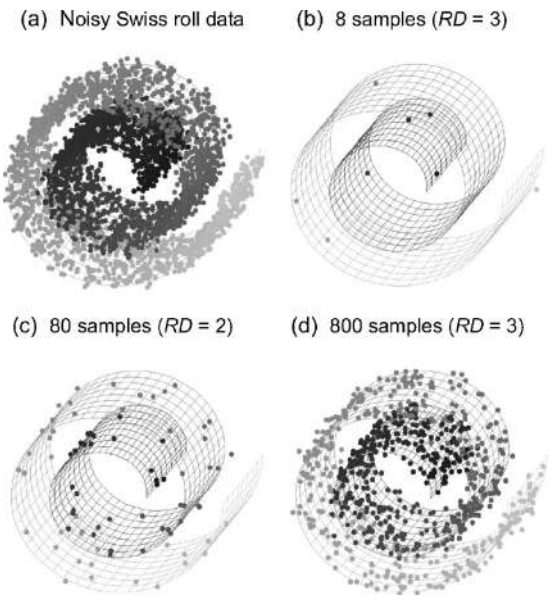

Fig. 2. (a) Swiss roll data to which Gaussian white noise has been added (here, showing 3200 samples), (b) eight random samples of the noisy Swiss roll data; here, there are too few samples to discern any particular geometry (RD=3), (c) 80 random samples of the noisy Swiss roll data; here, there are enough samples to begin to see a two-dimensional Swiss roll manifold (RD=2), (d) 800 random samples of the noisy Swiss roll data; here, there are enough samples to see that the roll actually has a thickness (RD=3).

dimension, pointwise dimension, and quantization dimension; see Ref. 36 for a review. Here, we borrow from these nearest-neighbor-based approaches and use our proximity distributions to estimate dimensionality. However, whereas the majority of dimensionality estimation techniques aim to estimate the *intrinsic* dimensionality of the data, here we do not focus on the intrinsic dimensionality for two reasons: (1) digitized natural images are both quantized and subject to noise, and we argue that the intrinsic dimensionality is equivalent to the dimensionality of the space in which the data are embedded (e.g., the intrinsic dimensionality is given by the number of pixels for digitized natural scenes); (2) real sensory systems cannot encode input signals in an error-free manner, and thus the error puts a limit on the entropy and dimensionality that is relevant to the sensory system. Accordingly, in this paper, we focus on the RD, defined as the dimensionality that the data appear to have, given a limited number of samples (the sampling density). As has been noted previously (see Ref. 35), and as we will confirm, the RD changes as a function of the sampling density. We emphasize that this dependence on sampling density can provide insights into the geometry of the data space (the manifold of natural scenes).

The RD is analogous to the dimensionality estimates given by techniques such as Isomap[33] and LLE.[34] However, the emphasis here is that the dimensionality is dependent on the number of samples and that this dependence can provide insight into the data space. For example, Fig. 2(a) depicts the classical Swiss roll data[33] to which Gaussian white noise has been added. Without the addition of noise, the data would have an intrinsic dimensionality of two; i.e., the data would lie on the two-dimensional surface (manifold) shown as a wireframe in Fig. 2(a), and thus with a proper transformation that "unrolls" the data, any data point could be described with only two coordinates. The addition of the noise, however, increases the intrinsic dimensionality of the data to three (three coordinates are required to specify any data point).

Clearly, with enough samples, one can readily visualize the intrinsic dimensionality of the data. However, given only a coarse sampling, the data might appear to have a vastly different dimensionality—a dimensionality that is *relative* to the number of samples and that may thus provide insight into the geometry of the space. For example, Figs. 2(b)–2(d) depict the noisy swiss roll data given only 8, 80, and 800 random samples, respectively. In Fig. 2(b), there is no clear geometry to the data; thus, given only 8 samples, one would estimate that the data are three dimensional (RD=3). In Fig. 2(c), given 80 samples, the swiss roll geometry begins to emerge, and one might guess that the data fall on this two-dimensional, swiss roll manifold (RD=2). In Fig. 2(d), given 800 samples, it becomes apparent that there is actually a thickness to the swiss roll, and thus the RD is equivalent to the intrinsic dimensionality of three.

Now consider the data shown in Fig. 3(b), which corresponds to an unrolled version of the noisy Swiss roll data. As with the noisy Swiss roll data, these unrolled data have an intrinsic dimensionality of three. However, because the data have been unrolled, as shown in Fig. 3(b), RD=2 given only eight samples. Similarly, given 80 samples [Fig. 3(c)], there are still an insufficient number of samples to discover the thickness of the plane (RD=2). At 800 samples [Fig. 3(d)], one begins to discover the third dimension (RD=3).

Thus, even though the noisy Swiss roll data and the unrolled version of the data have the same intrinsic dimensionality and entropy, the data sets have markedly different geometries. Accordingly, each data set gives rise to a different vector of RDs: [3, 2, 3] and [2, 2, 3] for the rolled and unrolled versions, respectively, given 8, 80, and 800 samples, respectively. In Subsection 2.D and Section 3, we use nearest-neighbor-based techniques to measure the

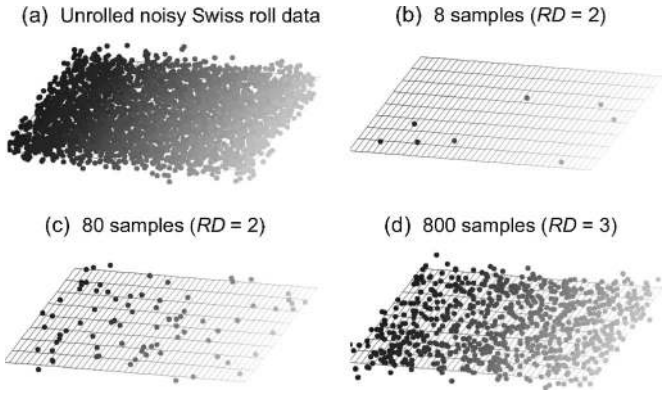

Fig. 3. (a) Unrolled version of the noisy Swiss roll data in Fig. 2. (b) Eight random samples of the unrolled data; here, there are too few samples to clearly expose the third dimension (RD=2); (c) 80 random samples of the unrolled data; here, there are still too few samples to clearly expose the third dimension (RD=2); (d) 800 random samples of the unrolled data; here, there are enough samples to see that the plane has a thickness (RD=3).

RD given $1-2^{18}$ samples of 3×3 and 8×8 images, respectively. We show that in this range of sample sizes the RD of natural-scene patches is significantly lower than the RD of Gaussian white-noise patches and that this difference in RD increases for larger patch sizes.

This paper is organized as follows: Section 2 describes the general methods used in the experiments performed to investigate entropy and RD, including details of the experimental stimuli and the theory behind the methods. Results and analyses of the experiments are provided throughout Sections 3–5. A discussion is provided in Section 6. General conclusions are provided in Section 7.

2. GENERAL METHODS

This section describes the experimental stimuli and procedures used in the experiments, an overview of the theory underlying the techniques, a derivation of the theory for Gaussian white noise, and a verification of the theory on 3×3 patches.

Three experiments were performed to estimate the entropy and dimensionality of various types of images. First, nearest-neighbor distances were measured for 8×8 patches cropped from images of various types; this experiment was designed to investigate the entropy and RD of a typical 8×8 image patch (Subsection 3.A). Next, nearest-neighbor distances were measured for 8×8 patches in which each patch was normalized for mean intensity and root-mean-square (RMS) contrast; this experiment was designed to investigate the entropy and RD of the *pattern* of a typical 8×8 image patch (Subsection 3.B). Finally, nearest-neighbor distances were measured for 16×16 patches to provide an estimate of how entropy and RD scale with patch size (Section 5).

A. Experimental Stimuli

Stimuli used in this study were $r\times r$ pixel patches cropped from 8-bit $R\times R$ pixel digitized and natively digital images[37] with pixel values in the range 0–255. Five types of images were used:

1. *Gaussian white noise*, in which each pixel was drawn independently from a common Gaussian distribution;
2. $1/f$ *noise* (amplitude spectrum), in which each Fourier component was drawn independently from a Gaussian distribution with standard deviation inversely proportional to the spatial frequency of the Fourier component;
3. $1/f^2$ *noise* (amplitude spectrum), in which each Fourier component was drawn independently from a Gaussian distribution with standard deviation inversely proportional to the squared spatial frequency of the Fourier component;
4. *Spectrum-equalized noise*, in which each Fourier component was drawn independently from a Gaussian distribution with variance proportional to the sample variance measured using a collection of natural scenes;
5. *Natural scenes* obtained from the van Hateren database.[38]

The (real-valued) pixels of all images were quantized to 8 bits (256 levels) of gray-scale resolution, as necessary, via uniform scalar quantization[39] in which real-valued pixel X was mapped to its quantized (discrete-valued) version X_Δ via $X_\Delta=\lfloor X+\frac{1}{2}\rfloor$, where $\lfloor\cdot\rfloor$ denotes the floor operator. The details of the image-generation process are as follows (experiment-specific details are provided throughout Sections 3 and 5).

Gaussian white noise: The Gaussian white noise images were generated by drawing $R\times R$ independent realizations from the following Gaussian distribution:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-[(x-\mu)^2/2\sigma^2]}, \tag{1}$$

where the mean μ and standard deviation σ were set as described in Sections 3 and 5. The pixel values of the resulting images were quantized to 8 bits. Figure 4(a) depicts one of the white-noise images used in this study.

$1/f$ *and* $1/f^2$ *noise*: The $1/f$ noise images were generated by first creating an $R\times R$ Gaussian white-noise image via Eq. (1) and then filtering that image with a digital, finite-impulse response filter with the following frequency response:

$$H(u,v)=\begin{cases}1 & u=v=0\\ \dfrac{1}{\sqrt{u^2+v^2}} & \text{else}\end{cases}, \tag{2}$$

where $u,v\in[0,R-1]$. The $1/f^2$ noise images were generated in a similar fashion by creating an $R\times R$ Gaussian white-noise image [Eq. (1)], followed by filtering with a digital filter with the following frequency response:

$$H(u,v)=\begin{cases}1 & u=v=0\\ \dfrac{1}{u^2+v^2} & \text{else}\end{cases}, \tag{3}$$

where u, $v\in[0,R-1]$. The filtering was performed in the frequency domain by means of the discrete Fourier transform (DFT) and multiplication of frequency responses (DFT coefficients). The pixel values of the resulting images were offset and scaled to span the range 0–255 and then quantized to 8 bits. Figures 4(b) and 4(c) depict, respectively, sample $1/f$ and $1/f^2$ images used in this study.

Spectrum-equalized noise: The spectrum-equalized noise images were generated in a fashion similar to that used for the $1/f$ and $1/f^2$ noise images except that the filtering was applied separately to each $r\times r$ pixel patch and was performed by using an empirical $H(u,v)$ determined based on the spectra of a large collection of $r\times r$ pixel patches. Specifically, a Gaussian white-noise image was created via Eq. (1), and then frequency-domain filtering was performed by multiplying the spectrum of each $r\times r$ pixel patch with the following $r\times r$ element frequency response:

$$H(u,v)=\sqrt{\sigma_\Re^2(u,v)+\sigma_\Im^2(u,v)}, \tag{4}$$

where u, $v\in[0,r-1]$, and where $\sigma_\Re(u,v)$ and $\sigma_\Im(u,v)$ denote the sample standard deviations of the real and imaginary components, respectively, of the DFT coefficient corresponding to frequency u, v; the sample standard deviations $\sigma_\Re(u,v)$ and $\sigma_\Im(u,v)$ were measured based on a collection of $r\times r$ patches from 71 natural scenes (described next). The pixel values of the resulting

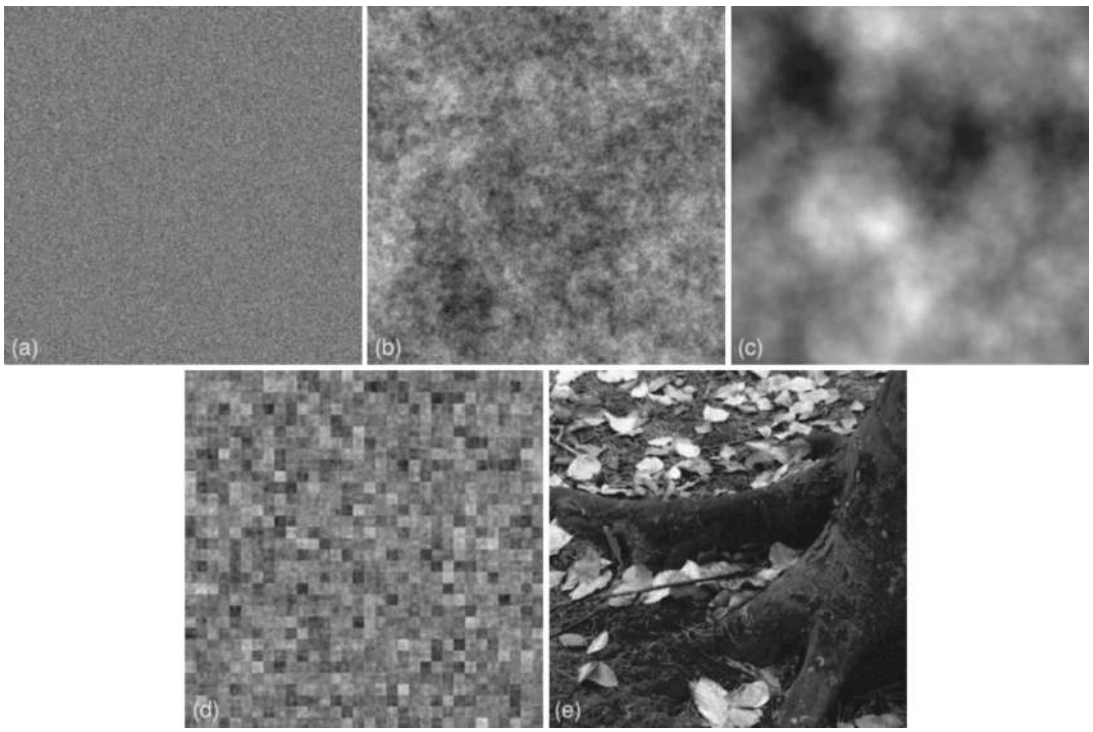

Fig. 4. Example stimuli used in the experiments ($R \times R = 1024 \times 1024$): (a) Gaussian white noise; (b) $1/f$ noise; (c) $1/f^2$ noise; (d) spectrum-equalized noise with $r \times r = 8 \times 8$; (e) natural scene cropped from image *imk*04103 of the van Hateren database [note that to promote visibility, the intensities of these images have been adjusted and (d) depicts only the top-left 256×256 section].

images were quantized to 8 bits. Figure 4(d) depicts one of the spectrum-equalized noise images used in this study.

Natural scenes: Seventy-one digitized natural scenes were selected at random from the van Hateren database.[38] The original images were of size 1536×1024 and contained 16-bit pixel values. A 1024×1024 section was cropped from each image, and then the pixel values of that 1024×1024 section were converted to a floating-point representation. The pixels were then offset, scaled to span the range 0–255, and quantized to 8 bits. Reference 40 provides further details regarding the specific images used in this study. Figure 4(e) depicts one of these images. We wish to emphasize that our estimates of entropy and RD are dependent on the particular class of images used here, and thus the results should not be considered universal for all natural scenes. The van Hateren database is attractive owing to its widespread use; however, factors such as camera blur, scene content, and noise all have a significant impact on the results.

B. Experimental Procedures

Let D_N^* denote the Euclidean distance between a patch and its nearest neighbor among N neighbors. The average log nearest-neighbor distance $E\{\log_2 D_N^*\}$ was estimated by using an exhaustive, brute-force search procedure. Each set of $R \times R$ images from each image class was randomly divided into two groups: Group $\mathcal{J}$, which consisted of images whose patches served as the to-be-matched "target" samples, and group $\mathcal{N}$, which consisted of images whose patches served as the target's neighbors. This division into targets and neighbors was used to avoid computing nearest-neighbor distances between patches from the same image. Patches of size $k = r \times r$ pixels were cropped from each $R \times R$ image in a nonoverlapping, sequential raster-scan order starting from the top-left corner of the image.

For each target patch in group $\mathcal{J}$, an exhaustive search was performed to find the corresponding patch in group $\mathcal{N}$ closest in Euclidean distance to the target patch. This procedure is illustrated in Fig. 1 and is formally defined as follows: Let $\mathbf{X}^{(\mathcal{J}_t)}$, $t \in [1, T]$, denote the tth target patch, and let $\mathbf{X}^{(\mathcal{N}_n)}$, $n \in [1, N]$, denote one of its neighbors. For each patch in group $\mathcal{J}$ and each value of N, the search procedure yields the Euclidean distance $D_{N,t}^*$ between $\mathbf{X}^{(\mathcal{J}_t)}$ and its nearest neighbor among N neighbors via

$$D_{N,t}^* = \min_{n \in [1,N]} \|\mathbf{X}^{(\mathcal{J}_t)} - \mathbf{X}^{(\mathcal{N}_n)}\|_{L_2} = \left(\min_{n \in [1,N]} \left\{ \sum_{i=1}^{k} (X_i^{(\mathcal{J}_t)} - X_i^{(\mathcal{N}_n)})^2 \right\} \right)^{1/2}, \quad (5)$$

where X_i denotes the ith pixel of $\mathbf{X}$. The search procedure was performed to compute $D_{N,t}^*$ for all T target patches, and then $E\{\log_2 D_N^*\}$ was estimated via the sample mean over all target patches, i.e., $E\{\log_2 D_N^*\} \approx (1/T) \Sigma_{t=1}^{T} \log_2 D_{N,t}^*$.

In all experiments, $D_{N,t}^*$ was measured at power-of-two values of N up to 2^K (i.e., $N = 1, 2, 4, \ldots, 2^K$), where K was determined by the total number of images in group $\mathcal{N}$, the latter of which was chosen based on the patch size (see Sections 3 and 5). This process was repeated for at least three trials for each patch in group $\mathcal{J}$. Owing to the enormous memory and processing-time requirements, the total number of patches in group $\mathcal{J}$ was selected based on initial runs and was varied across image classes and

patch size; further details regarding the total number of patches in groups $\mathcal{J}$ and $\mathcal{N}$ are provided throughout Sections 3 and 5.

C. Theory

In this paper, we estimate entropy and dimensionality based on nearest-neighbor distances. This section provides a brief outline of the mathematical theory upon which this technique is based. The estimation of entropy based on nearest-neighbor distances was initially proposed by Kozachenko and Leonenko[18] and was later applied to neural data by Victor[19] and subsequently to the estimation of mutual information by Kraskov *et al.*[20] and by Kybic.[41] This is a so-called *binless* estimator of differential entropy that operates by estimating $i_{\mathbf{X}}(\mathbf{x}) \triangleq -\log_2 f_{\mathbf{X}}(\mathbf{x})$ via nearest-neighbor distances, where $\mathbf{X}$ denotes a (possibly vector-valued) random variable with corresponding probability density function $f_{\mathbf{X}}(\mathbf{x})$. In this formulation, differential entropy, $h(\mathbf{X})$, is the expected value of $i_{\mathbf{X}}(\mathbf{x})$:

$$h(\mathbf{X}) \triangleq -\int_{\mathbf{x}\in\mathcal{A}} f_{\mathbf{X}}(\mathbf{x})\log_2 f_{\mathbf{X}}(\mathbf{x})\mathrm{d}\mathbf{x} = \int_{\mathbf{x}\in\mathcal{A}} f_{\mathbf{X}}(\mathbf{x}) i_{\mathbf{X}}(\mathbf{x})\mathrm{d}\mathbf{x} = E\{i_{\mathbf{X}}(\mathbf{x})\} \approx \frac{1}{M}\sum_{m=1}^{M} \hat{i}_{\mathbf{X}}(\mathbf{x}_m), \tag{6}$$

where the final relation approximates the expectation in the third relation with the sample mean computed using M observed samples, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M$, drawn according to $f_{\mathbf{X}}$. Specifically, the approximation results from (1) replacing the integral with a sum; (2) assuming $f_{\mathbf{X}}(\mathbf{x})\mathrm{d}\mathbf{x} \approx 1/M$, $\forall \mathbf{x}_m$; and (3) using $\hat{i}_{\mathbf{X}}(\mathbf{x}_m)$ as an estimator of $i_{\mathbf{X}}(\mathbf{x}_m)$.

The estimator $\hat{i}_{\mathbf{X}}(\mathbf{x})$ is computed based on the Euclidean distance D_N^* between $\mathbf{x}$ and its nearest neighbor among the remaining $N=M-1$ observations as

$$\hat{i}_{\mathbf{X}}(\mathbf{x}) = kE\{\log_2 D_N^*\} + \log_2\left(\frac{A_k N}{k}\right) + \frac{\gamma}{\ln 2}, \tag{7}$$

where γ is the Euler constant, and where $A_k = k\pi^{k/2}/\Gamma(k/2+1)$ denotes the surface area of a k-dimensional hypersphere. Combining Eqs. (6) and (7), $h(\mathbf{X})$ is approximated by

$$h(\mathbf{X}) \approx \frac{k}{M}\sum_{m=1}^{M} \log_2 D_{N,m}^* + \log_2\left(\frac{A_k N}{k}\right) + \frac{\gamma}{\ln 2}, \tag{8}$$

where $D_{N,m}^*$ is the Euclidean distance between $\mathbf{x}_m$ and its nearest neighbor among the other $N=M-1$ observations.

For images in which the pixel values are drawn independently from a common Gaussian distribution, the pixels are independently and identically distributed (iid) Gaussian. The Gaussian distribution possess several favorable mathematical properties that facilitate an analysis of its nearest-neighbor-distance behavior and entropy. In particular, the differential entropy of a Gaussian random variable X can be computed directly via

$$h(X) = \frac{1}{2}\log_2(2\pi e\sigma^2) \text{ bits}, \tag{9}$$

where σ denotes the standard deviation of the Gaussian. Moreover, of all distributions with a given fixed variance, the Gaussian distribution maximizes differential entropy (Ref. 22, Theorem 9.6.5).

In addition, for iid Gaussian realizations there exists an analytical solution for the expected log nearest-neighbor distance among N neighbors ($E\{\log_2 D_N^*\}$). We show this by first deriving the distribution of Euclidean distances between two patches, and then we extend that result to the expected minimum distance among N patches.

Distribution of distances between two patches: Without loss of generality, we assume that each pixel is drawn from a zero-mean Gaussian distribution.[42] Let $X_i \sim \mathcal{N}(0,\sigma^2)$ and $Y_i \sim \mathcal{N}(0,\sigma^2)$ denote the ith pixel of image $\mathbf{X}$ and $\mathbf{Y}$, respectively. Clearly, $X_i - Y_i \sim \mathcal{N}(0,2\sigma^2)$. Thus, we can define a new random variable $\tilde{D} = (1/2\sigma^2)\Sigma_{i=1}^{k}(X_i - Y_i)^2$, which follows a χ^2 distribution with k degrees of freedom.[43] Observe that $\tilde{D}$ is $1/2\sigma^2$ times the squared Euclidean distance between $\mathbf{X}$ and $\mathbf{Y}$. Given that $\tilde{D} \sim \chi_k^2$, the cumulative distribution function is given by $F_{\tilde{D}}(d) = 1 - \Gamma(k/2, d^2/2)/\Gamma(k/2)$, where $\Gamma(a,x)$ and $\Gamma(a)$ are the upper incomplete and complete gamma functions, respectively; and the corresponding probability density function is given by $f_{\tilde{D}}(d) = (d^{k/2-1}e^{-d/2})/[2^{k/2}\Gamma(k/2)]$.

Expected nearest-neighbor distance among N patches: Let $\tilde{D}_N^*$ denote $1/2\sigma^2$ times the squared Euclidean distance between a patch and its nearest neighbor among N neighbors. The cumulative distribution function for $\tilde{D}_N^*$ is thus given by $F_{\tilde{D}_N^*}(d) = 1 - (1 - F_{\tilde{D}}(d))^N$, and the corresponding probability distribution function is given by $f_{\tilde{D}_N^*}(d) = N(1 - F_{\tilde{D}}(d))^{N-1} f_{\tilde{D}}(d)$. Note that the nearest-neighbor distance $D_N^* = (2\sigma^2\tilde{D}_N^*)^{1/2}$ and thus $\log_2 D_N^* = \frac{1}{2}\log_2(2\sigma^2\tilde{D}_N^*) = \frac{1}{2}\log_2(2\sigma^2) + \frac{1}{2}\log_2 \tilde{D}_N^*$. The expected log nearest-neighbor distance $E\{\log_2 D_N^*\}$ is therefore given by

$$\begin{aligned}
E\{\log_2 D_N^*\} &= \frac{1}{2}\log_2(2\sigma^2) + \frac{1}{2}E\{\log_2 \tilde{D}_N^*\} \\
&= \frac{1}{2}\log_2(2\sigma^2) + \frac{1}{2}\int_0^\infty f_{\tilde{D}_N^*}(\zeta)\log_2(\zeta)\mathrm{d}\zeta \\
&= \frac{1}{2}\log_2(2\sigma^2) + \frac{N}{2}\int_0^\infty (1 - F_{\tilde{D}}(\zeta))^{N-1} f_{\tilde{D}}(\zeta)\log_2(\zeta)\mathrm{d}\zeta \\
&= \frac{1}{2}\log_2(2\sigma^2) + \frac{N}{2}\int_0^\infty \left(\frac{\Gamma(k/2,\zeta^2/2)}{\Gamma(k/2)}\right)^{N-1} \\
&\quad \times \frac{1}{2^{k/2}\Gamma(k/2)}\zeta^{k/2-1}e^{-\zeta/2}\log_2(\zeta)\mathrm{d}\zeta \\
&= \frac{1}{2}\log_2(2\sigma^2) + \frac{N}{2^{k/2+1}\Gamma(k/2)^N}
\end{aligned}$$

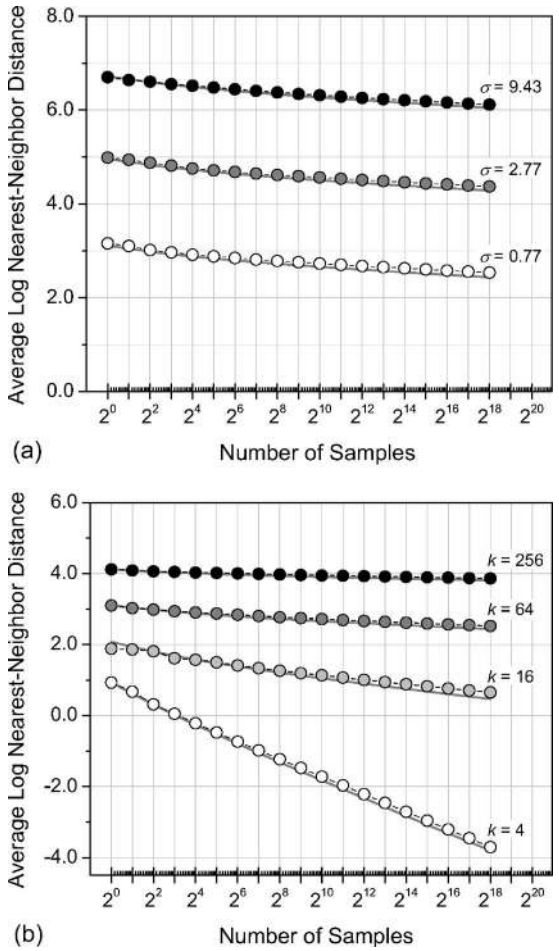

Fig. 5. Proximity distribution functions for iid Gaussian data computed via Eq. (10) (solid curves) and measured experimentally (circles). In each graph, the horizontal axis denotes the number of samples N; the vertical axis denotes the corresponding $E\{\log_2 D_N^*\}$ computed via Eq. (10). (a) Proximity distribution functions for a fixed dimensionality ($k=64$) and various values of standard deviation σ; (b) proximity distribution functions for a fixed standard deviation ($\sigma=0.77$) and various values of dimensionality k.

$$\times \int_0^\infty \Gamma\left(\frac{k}{2}, \frac{\zeta^2}{2}\right)^{N-1} \zeta^{k/2-1} e^{-\zeta/2} \log_2(\zeta) \mathrm{d}\zeta. \quad (10)$$

To verify that the experimental procedures described in Subsection 2.B yield results that are consistent with Eq. (10), 8×8 patches cropped from images in which the pixel values were iid Gaussian were used as a control condition. The Gaussian white-noise images were generated as described in Subsection 2.A via Eq. (1) with fixed mean $\mu=127.5$, and with three different standard deviations $\sigma=9.43$, $\sigma=2.77$, and $\sigma=0.77$. Seventeen images of size 1024×1024 pixels were generated for each standard deviation, one of which was placed into group $\mathcal{J}$, and the remaining 16 of which were placed into group $\mathcal{N}$. Thus, there were a total of $(1024\times1024)/(8\times8)=16{,}384$ target patches and $16\times(1024\times1024)/(8\times8)=262{,}144$ potential neighbors.

Figure 5 depicts proximity distribution functions computed via Eq. (10) (computed digitally via a summation-based approximation to the integral) and the corresponding data measured experimentally. In each graph, the vertical axis corresponds to $E\{\log_2 D_N^*\}$ and the horizontal axis corresponds to the number of samples (here, $N=1,2,4,\ldots,2^{18}$). Figure 5(a) depicts proximity distribution functions for a fixed dimensionality $k=64$ and various values of standard deviation σ. Figure 5(b) depicts proximity distribution functions for a fixed standard deviation $\sigma=0.77$ and various values of dimensionality k. Notice that the theoretical and experimental results are very much in agreement ($R^2>0.99$).

The trends in Fig. 5(a) demonstrate that for a fixed dimensionality ($k=64$), decreasing the standard deviation of the underlying Gaussian effects a downward shift in the proximity distribution function. Indeed, this observation follows directly from Eq. (10): Notice that only the left-hand portion of the sum depends on σ and that this portion depends *only* on σ. The trends in Fig. 5(b) demonstrate that when the standard deviation is fixed ($\sigma=0.77$), changing the dimensionality effects both a downward shift and a change in the slope.

D. Verification of the Theory on 3×3 Patches

When the slope of the proximity distribution function of a data set has converged (i.e., the RD has converged to the intrinsic dimensionality of the data), there are a sufficient number of samples to estimate the entropy. Here, we show this result experimentally by applying Eq. (8) to nearest-neighbor distances measured for 3×3 patches.

Patches of size 3×3 pixels ($r=3$, $k=9$) drawn from Gaussian white-noise, spectrum-equalized noise, and natural scenes were used in this verification experiment. The Gaussian white-noise images were generated via Eq. (1) with $\mu=127.5$ and $\sigma=32$. For the Gaussian white-noise and spectrum-equalized noise images, 13 images of size $R\times R=513\times513$ pixels were generated as described in Subsection 2.A; for each image type, three images were placed into group $\mathcal{J}$, and the remaining 10 images were placed into group $\mathcal{N}$, resulting in $3\times(513\times513)/(3\times3)$

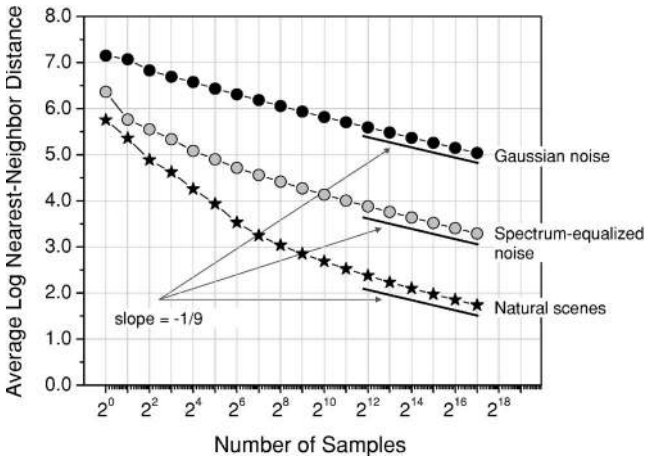

Fig. 6. Proximity distribution functions for 3×3 patches of Gaussian white noise, spectrum-equalized noise, and natural scenes. The horizontal axis denotes the number of samples N; the vertical axis denotes the corresponding $E\{\log_2 D_N^*\}$ estimated via a sample mean over all target patches. Black circles, Gaussian white noise; light-gray circles, spectrum-equalized noise; stars, natural scenes. The solid lines represent a slope of −1/9th; notice that all three curves eventually converge on this slope.

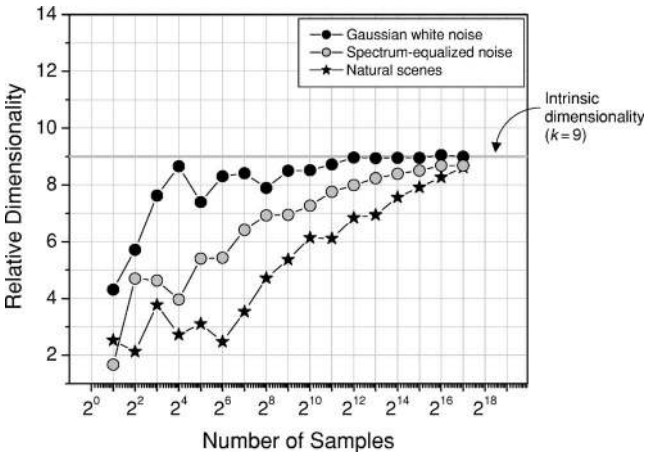

Fig. 7. RD curves for 3×3 patches of Gaussian white noise, spectrum-equalized noise, and natural scenes. The horizontal axis denotes the number of samples N; the vertical axis denotes the corresponding RD. Black circles, Gaussian white noise; light-gray circles, spectrum-equalized noise; stars, natural scenes. The solid gray line denotes the intrinsic dimensionality of $k=9$ for all three data sets (the natural scenes possess an intrinsic dimensionality of $k=9$ owing to photon noise).

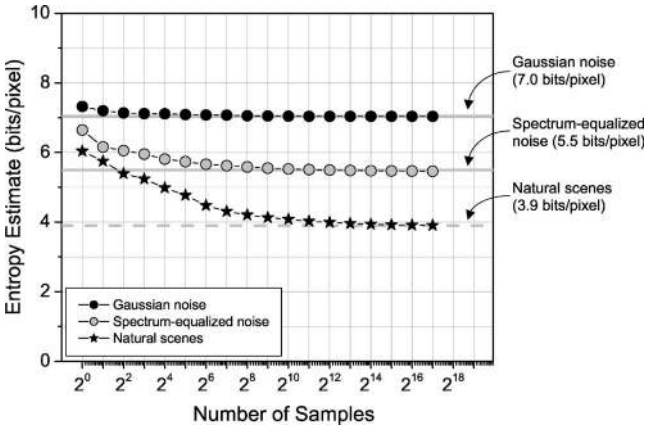

Fig. 8. Entropy estimates for 3×3 patches of Gaussian white noise, spectrum-equalized noise, and natural scenes. The horizontal axis denotes the number of samples N; the vertical axis denotes the entropy computed via Eq. (8) using the corresponding value of N. Black circles, Gaussian white noise (7.0 bits/pixel); light-gray circles, spectrum-equalized noise (5.5 bits/pixel); stars, natural scenes (3.9 bits/pixel). The solid gray lines indicate the actual entropies of Gaussian white noise and spectrum-equalized noise (7.0 and 5.5 bits/pixel, respectively) as computed via Eq. (9); the dashed line denotes the entropy estimate of 3.9 bits/pixel for natural scenes.

$=87{,}723$ target patches and $10\times(1024\times1024)/(8\times8)=292{,}410$ potential neighbors. For the natural scenes, images of size 1024×1024 pixels were obtained as described in Subsection 2.A, and patches were selected from the top-left 1023×1023 portion of each image. Five images were placed into group $\mathcal{J}$, and 66 images were placed into group $\mathcal{N}$, resulting in $5\times(1023\times1023)/(3\times3)=581{,}405$ target patches and $66\times(1023\times1023)/(3\times3)=7{,}674{,}546$ potential neighbors.

Figure 6 depicts the resulting proximity distribution functions; the horizontal axis denotes the number of samples N, and the vertical axis denotes the corresponding log nearest-neighbor distance averaged over all patches in group $\mathcal{J}$. RDs and estimates of entropy [computed via Eq. (8)] based on these proximity distribution data are provided in Figs. 7 and 8, respectively. The solid gray line in Fig. 7 denotes an intrinsic dimensionality of $k=9$; the solid gray lines in Fig. 8 denote the true values of entropy as computed via Eq. (9). Note that due to the presence of photon noise, the natural scenes also possess an intrinsic dimensionality of $k=9$.

Notice from the proximity distribution functions of Fig. 6 that for a given number of samples, spectrum-equalized noise exhibits a lower average log nearest-neighbor distance than Gaussian white noise, and natural scenes exhibits a lower average log nearest-neighbor distance than both Gaussian white noise and spectrum-equalized noise. Similarly, notice from Fig. 7 that although the RD curves for all three image types eventually converge to a dimensionality of approximately $k=9$, spectrum-equalized noise exhibits a lower RD than Gaussian white noise, and natural scenes exhibit a lower RD than both Gaussian white noise and spectrum-equalized noise.

Because the RD curves of Fig. 7 have approximately converged to a dimensionality of $k=9$ given $N=2^{17}$ samples, there are sufficient data to estimate entropy. The entropy estimates shown in Fig. 8 were obtained by using Eq. (8) with $k=9$. Indeed, for Gaussian white noise and spectrum-equalized noise, the estimates of entropy yield the correct values: 63 bits (7.0 bits/pixel) for Gaussian white noise and 49 bits (5.5 bits/pixel) for spectrum-equalized noise [the actual entropies were computed via Eq. (9); see Ref. 44]. Here, we obtain an estimate of 35 bits (3.9 bits/pixel) for the entropy of 3×3 natural scenes. We stress again that this result is not universal for all natural scenes; rather, it is dependent on the particular sample of images from the van Hateren database used here.

These results confirm that our main assumption holds for the images used here: Given a sufficient number of samples, the RD converges on the intrinsic dimensionality of the data, and thus the entropy estimate is close to the true entropy of the data. In the following sections, we investigate extensions of these estimators to 8×8 patches of various types of images for which there are an insufficient number of samples to directly apply the estimates.

3. RESULTS FOR 8×8 PATCHES

In Experiment 1, patches of size 8×8 pixels were used ($r=8$, $k=64$). Each patch $\mathbf{X}=[X_1,X_2,\ldots,X_{64}]$ can thus be viewed as a point in a 64-dimensional metric space $\mathcal{V}_{64}$ with distance function $d(\mathbf{X},\mathbf{Y})=\|\mathbf{X}-\mathbf{Y}\|_{L_2}=\sqrt{\Sigma_{i=1}^{64}(X_i-Y_i)^2}$. In our experiments, each patch was a discrete-valued random vector in which each pixel was limited to integer values in the range 0–255 ($l=256$ levels) as a result of the 8-bit quantization, and therefore the actual space is limited to $L=l^k=256^{64}=2^{512}$ possibilities.

A. 8×8 Patches

To serve as a control condition, patches of size 8×8 pixels were cropped from Gaussian white-noise images. Nineteen Gaussian white-noise images were gener-

ated as described in Subsection 2.A via Eq. (1) with $\mu = 127.5$ and $\sigma = 36$. Three images were placed into group $\mathcal{J}$, and the remaining 16 images were placed into group $\mathcal{N}$, resulting in $3 \times (1024 \times 1024)/(8 \times 8) = 49{,}152$ target patches and $16 \times (1024 \times 1024)/(8 \times 8) = 2^{18} = 262{,}144$ potential neighbors.

To investigate the effects of spatial correlations on nearest-neighbor distances, 8×8 patches cropped from images with $1/f$ and $1/f^2$ amplitude spectra ($1/f^2$ and $1/f^4$ power spectra, respectively) and from images with spectrum-equalized patches were used. In this paradigm, the image's DFT coefficients form a set of independent Gaussian random variables with standard deviations inversely proportional to spatial frequency. Nineteen $1/f$, $1/f^2$, and spectrum-equalized noise images of size 1024×1024 pixels were generated as described in Subsection 2.A. For each image type, three images were placed into group $\mathcal{J}$, and the remaining 16 images were placed into group $\mathcal{N}$, resulting in $3 \times (1024 \times 1024)/(8 \times 8) = 49{,}152$ target patches and $16 \times (1024 \times 1024)/(8 \times 8) = 2^{18} = 262{,}144$ potential neighbors.

In addition, to investigate the effects of the statistical properties of natural scenes on nearest-neighbor distances, 8×8 patches cropped from images obtained from the van Hateren database[38] were used. Seventy-one natural scenes were obtained as described in Subsection 2.A, five of which (chosen at random) were placed into group $\mathcal{J}$, and the remaining 66 of which were placed into group $\mathcal{N}$. Thus, there were a total of 81,920 target patches and a total of 1,081,344 potential neighbors.

Figure 9(a) depicts the proximity distribution functions for the patches taken from the $1/f$, $1/f^2$, and spectrum-equalized noise images (gray, white, and light-gray circles, respectively) and from the natural scenes (stars), along with the proximity distribution function for Gaussian white-noise patches (black circles). The horizontal axis denotes the number of samples N, and the vertical axis denotes the corresponding log nearest-neighbor distance averaged over all patches in group $\mathcal{J}$.

Images that possess power spectra that follow $1/f^{\alpha}$ demonstrate greater degrees of pairwise pixel correlations for increasing values of α. Gaussian white-noise images, which contain uncorrelated pixels, possess an amplitude spectrum in which $\alpha = 0$. The $1/f$ and $1/f^2$ images possess power spectra in which $\alpha = 2$ and $\alpha = 4$, respectively (amplitude spectra in which $\alpha = 1$ and $\alpha = 2$, respectively). The 8×8 patches of the spectrum-equalized noise images possess a power spectrum in which $\alpha \approx 2.8$. Thus, the proximity distribution functions of Fig. 9(a) demonstrate that for a fixed variance, increasing pairwise correlations between pixels increases the magnitude of the slope of the proximity distribution functions, which therefore suggests a lower entropy state.

The data of Fig. 9(a) also show that the proximity distribution function for the patches of natural scenes lies below the proximity distribution function for the patches of spectrum-equalized noise, despite the fact that the power spectra for these image types are equalized. These data confirm that the presence of spatial correlations does not provide a complete account for the redundancy (lower entropy) of natural scenes.

Figure 9(b) depicts the RD curves for these images computed as the magnitude of the inverse of the instantaneous slope between successive pairs of measured values of $E\{\log_2 D_N^*\}$ [i.e., $-d \log_2(N)/dE\{\log_2 D_N^*\}$]. The horizontal axis denotes the number of samples N, and the vertical axis denotes the corresponding dimensionality given N samples. Notice that for most values of N (in particular, for $N > 16$), $1/f^2$ noise exhibits the lowest RD, natural scenes exhibit a slightly greater RD than $1/f^2$ noise, spectrum-equalized noise exhibits an even greater RD, followed by $1/f$ noise, and then Gaussian white noise. At $N = 2^{18}$ samples, the dimensionalities are approximately 13, 17, 27, 34, and 45 for $1/f^2$ noise, natural scenes, spectrum-equalized noise, $1/f$ noise, and Gaussian white noise, respectively. Clearly, many more samples are needed before these RD curves converge on the intrinsic dimensionality of $k = 64$, and thus $N = 2^{18}$ is an insufficient number of samples to produce a direct estimate of the entropy via Eq. (8). In Section 4, we discuss extrapolation techniques that attempt to overcome this limitation.

B. Mean- and Contrast-Normalized 8×8 Patches

Part of the redundancy in natural scenes can be attributed to the fact that natural scenes contain many low-

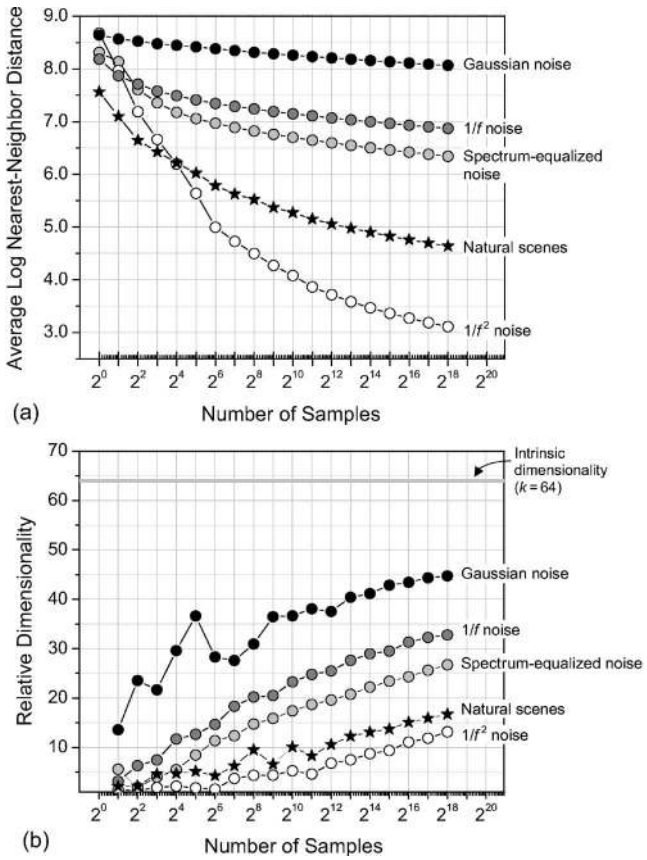

Fig. 9. (a) Proximity distribution and (b) RD curves for 8×8 patches. In both graphs, the horizontal axis denotes the number of samples N. The vertical axis in (a) denotes the corresponding $E\{\log_2 D_N^*\}$ estimated via a sample mean over all target patches; the vertical axis in (b) denotes the corresponding RD. Black circles, Gaussian white noise; gray circles, $1/f$ noise; light-gray circles, spectrum-equalized noise; white circles, $1/f^2$ noise; stars, natural scenes. The solid gray line in (b) denotes the intrinsic dimensionality of $k = 64$ for all data sets.

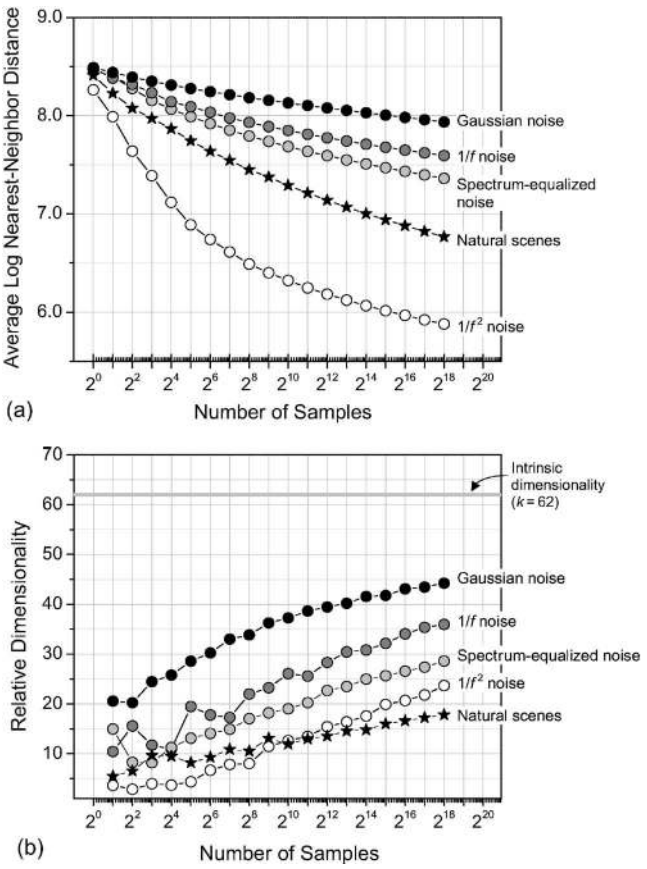

Fig. 10. (a) Proximity distribution and (b) RD curves for mean- and contrast-normalized 8×8 patches. The horizontal axis in both graphs denotes the number of samples N. The vertical axis in (a) denotes the corresponding $E\{\log_2 D_N^*\}$ estimated via a sample mean over all target patches; the vertical axis in (b) denotes the corresponding RD. Black circles, Gaussian white noise; gray circles, $1/f$ noise; light-gray circles, spectrum-equalized noise; white circles, $1/f^2$ noise; stars, natural scenes. The solid gray line in (b) denotes the intrinsic dimensionality of $k=62$ for all data sets.

contrast regions (e.g., in sky), whereas noise images such as spectrum-equalized noise only seldomly contain such low-contrast regions. To determine whether this prevalence of low-contrast patches can account for the differences in the proximity distribution functions, Experiment 2 investigated the nearest-neighbor-distance behavior of the underlying *patterns* by first normalizing the image patches for absolute luminance and RMS contrast. Specifically, each patch $\mathbf{X}$ was adjusted to have a zero mean and unity vector norm (L_2 norm) via

1. $\mathbf{X} := \mathbf{X} - (1/64)\Sigma_{i=1}^{64} X_i$,
2. $\mathbf{X} := 255\mathbf{X}/\sqrt{\Sigma_{i=1}^{64} X_i^2}$,

where X_i denotes the ith pixel of $\mathbf{X}$. Here, we limited the analysis to those patches with variance (after Step 1, above) of $(1/64)\Sigma_{i=1}^{64} X_i^2 > 2$ to prevent both division by zero (in Step 2, above) and amplification of noise.

The 1024×1024 images were randomly divided into groups $\mathcal{J}$ (containing the to-be-matched, target patches) and $\mathcal{N}$ (containing the neighbors). For the Gaussian white-noise images, only a single standard deviation $\sigma = 36$ was tested; Groups $\mathcal{J}$ and $\mathcal{N}$ consisted of 16,384 and 262,144 patches, respectively. For the $1/f$, $1/f^2$, and spectrum-equalized noise images, groups $\mathcal{J}$ and $\mathcal{N}$ consisted of 49,152 and 262,144 patches, respectively. For the natural scenes, groups $\mathcal{J}$ and $\mathcal{N}$ consisted of approximately 68,000 and 900,000 patches, respectively.

Figures 10(a) and 10(b) depict the resulting proximity distribution and RD, respectively, curves for Gaussian white noise (black circles), $1/f$ noise (gray circles), $1/f^2$ noise (white circles), spectrum-equalized noise (light-gray circles), and natural scenes (stars). In comparison with the proximity distribution functions in Fig. 9(a), notice that the curves for these mean- and contrast-normalized data all demonstrate a decrease in slope, suggesting that more templates are needed to describe the high-contrast patches to the same level of accuracy (vector norm of the difference) as that achieved when patches of all contrasts are considered. However, notice from Fig. 10(a) that the relative nearest-neighbor-distance behavior (rank order) of these various curves remains intact; in particular, natural scenes still fall below spectrum-equalized noise. Thus, the redundancy found in natural scenes cannot be

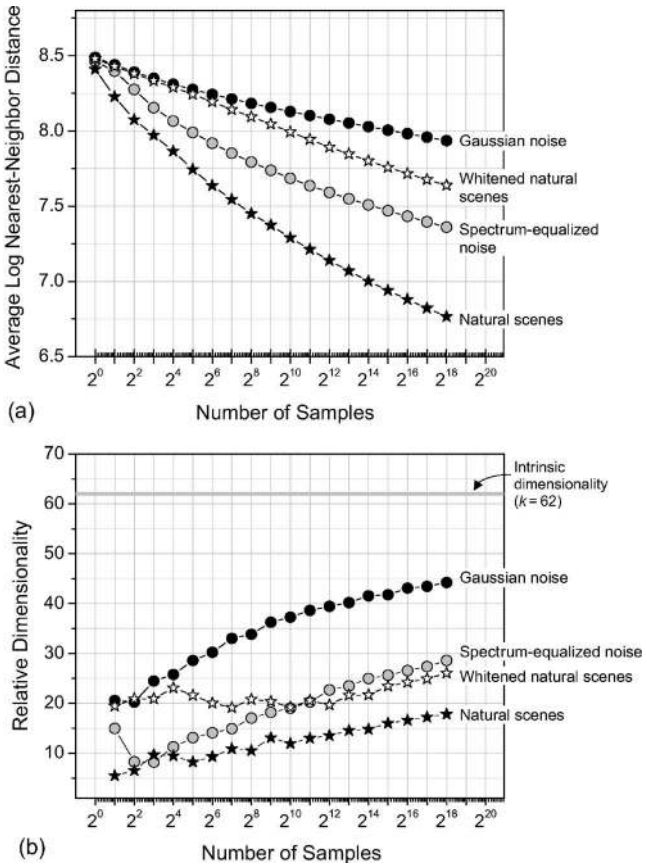

Fig. 11. (a) Proximity distribution and (b) RD curves for mean- and contrast-normalized 8×8 patches of whitened natural scenes and of Gaussian white noise, spectrum-equalized noise, and natural scenes (replotted from Fig. 10). In both graphs, the horizontal axis denotes the number of samples N; the vertical axis in (a) denotes the corresponding $E\{\log_2 D_N^*\}$ estimated via a sample mean over all target patches; and the vertical axis in (b) denotes the corresponding RD. Black circles, Gaussian white noise; light-gray circles, $1/f$ noise; black stars, natural scenes; white stars, whitened natural scenes. The solid gray line in (b) denotes the intrinsic dimensionality of $k=62$ for all data sets.

attributed solely to the power spectrum, nor can it attributed to the combination of the power spectrum and the prevalence of low-contrast patches.

To further investigate the effects of the power spectrum on nearest-neighbor distances, proximity distribution functions were measured for whitened natural scenes. To each of the 71 natural scenes (obtained as described in Subsection 2.A), the following whitening filter was applied:

$$H(u,v) = (\sqrt{u^2 + v^2})^{1.38}, \tag{11}$$

where $u, v \in [0, 1023]$, and where the exponent 1.38 was measured by linearly regressing log magnitude (of the DFT coefficients averaged over all orientations) on log frequency (radial distance from zero frequency) using all 71 images. The filtering was performed in the frequency domain by means of the DFT and multiplication of spectra. The pixel values of the resulting images were offset and scaled to span the range 0–255 and then quantized to 8 bits. The images chosen for groups $\mathcal{J}$ and $\mathcal{N}$ were whitened versions of the same images used for these groups in Experiment 1 and in Subsection 3.B; thus, groups $\mathcal{J}$ and $\mathcal{N}$ consisted of approximately 68,000 and 900,000 patches, respectively.

Figures 11(a) and 11(b) depict the resulting proximity distribution and RD curves, respectively, for Gaussian white noise (black circles), spectrum-equalized noise (light-gray circles), natural scenes (black stars), and whitened natural scenes (white stars). The application of a whitening filter serves to remove average pairwise spatial correlations; thus, if the redundancy in the high-contrast patches of natural scenes were due solely to these correlations, we would expect the nearest-neighbor-distance behavior of whitened natural scenes to be identical to that of Gaussian white noise. Instead, we find the proximity distribution function for whitened natural scenes falls below the proximity distribution function for Gaussian white noise, which indicates that fewer templates are required (on average) to describe a whitened natural scene to the same level of accuracy as that achieved for Gaussian white noise. These data further suggest that the redundancy of natural scenes cannot be attributed solely to a combination of the spectrum and the prevalence of low-contrast patches.

4. ENTROPY EXTRAPOLATIONS (XENTROPY)

For the 3×3 patches analyzed in Subsection 2.D, the RD curves had converged to the intrinsic dimensionality of the data ($k=9$), whereas the RD curves shown in Fig. 9(b) for the 8×8 patches requires a prohibitively large number of samples to converge to a dimensionality of $k=64$. As a result, applying Eq. (8) using the corresponding proximity distribution data would result in a poor estimate of entropy. To overcome this limitation, we explore three techniques for extrapolating the proximity distribution data and thereby estimating entropy based on the extrapolations. We define the term XEntropy to denote this extrapolated entropy estimate and to reinforce the notion that these are only estimates of the entropy based on extrapolations.

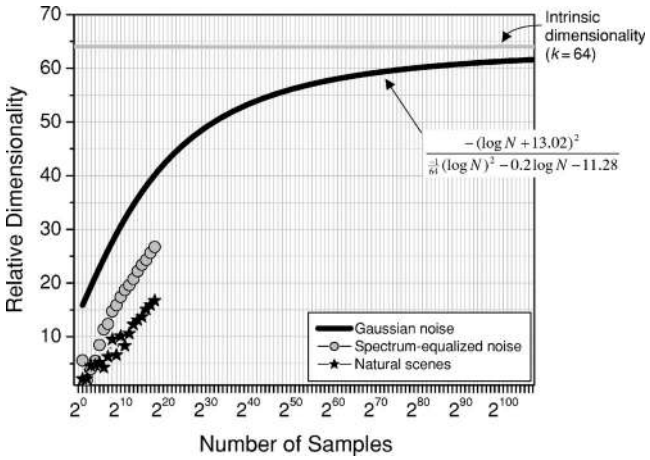

Fig. 12. RD curves for 8×8 Gaussian white noise (black curve), spectrum-equalized noise (light-gray circles), and natural scenes (stars). The RD curve for the Gaussian white noise was computed at values of $N \in [1, 2^{50}]$ via Eq. (10), and the remainder of the curve was fitted with $relativedimensionality(N) = -(\log N + b_0)^2/(a_2[\log N]^2 + 2a_2 b_0 \log N + a_1 b_0 - a_0)$, where $a_2 = -1/64$, $a_1 = 4.13$, $a_0 = 65.05$, and $b_0 = 13.02$ were computed via the Nelder–Mead simplex method. The data for the spectrum-equalized noise and natural scenes are replotted from Fig. 9(b). The solid gray line denotes the intrinsic dimensionality of $k=64$ for all data sets.

We use two constraints to aid in the extrapolations: (1) The expected log nearest-neighbor distance is a monotonically decreasing function of the number of samples; and (2) the RD curves for $r \times r$ white noise, spectrum-equalized noise, and natural scene patches must eventually converge to the intrinsic dimensionality of $k = r^2$. The first constraint specifies that the proximity distribution is necessarily a monotonically decreasing function (i.e., the slopes must be less than zero and therefore the RD functions must be greater than zero). The second constraint specifies that for 8×8 patches, the RD curves will eventually converge to a value of 64. Thus, we extrapolate the proximity distribution data by extrapolating the corresponding RD data.

Figure 12 depicts the RD curves for 8×8 patches of Gaussian white noise, spectrum-equalized noise, and natural scenes. The RD curves for spectrum-equalized noise and natural scenes are replotted from Fig. 9(b). The RD curve for Gaussian white noise (indicated by the black curve in Fig. 12) was computed via Eq. (10) for $N \in [1, 2^{50}]$, and the remainder of the curve was fitted with $RD(N) = -(\log N + b_0)^2/(a_2[\log N]^2 + 2a_2 b_0 \log N + a_1 b_0 - a_0)$, where the parameters $a_2 = -1/64$, $a_1 = 4.13$, $a_0 = 65.05$, and $b_0 = 13.02$ were computed via the Nelder–Mead simplex method.[45] The form of this function is by no means optimal; it was chosen (1) for its relative simplicity (it is a rational function in $\log N$); (2) for the fact that in the limit of large N, $RD(N) = 1/a_2 = 64$; and (3) because it provides decent fits to all three data sets (noise, spectrum-equalized noise, and natural scenes). However, although we believe that this is a rational extrapolation, we also believe that future work will allow more theoretically accurate predictions. In particular, we believe that it is possible to use the known statistics (e.g., the power spectra) to guide the bounds of the extrapolations. The corresponding entropy estimate for Gaussian white noise obtained by using Eq.

(8) with $k=64$ and $N=2^{300}$ yields a value of 449 bits (7.0 bits/pixel); the actual entropy computed via Eq. (9) is 462 bits (7.2 bits/pixel).

In the following sections, we assume three different forms of the RD curves of spectrum-equalized noise and natural scenes—form A, form B, and form C—which give rise to three corresponding techniques for extrapolating the proximity distribution data and thereby give rise to three entropy estimates. These extrapolated entropy estimates are denoted as XEntropy A, XEntropy B, and XEntropy C, respectively. We describe each of these extrapolation techniques so as to make explicit how different assumptions can lead to different estimates of entropy. Although XEntropy A and XEntropy B provide useful upper bounds with simple assumptions, we believe our best estimates of the true entropy are derived from XEntropy C. However, we are also confident that future work can improve on these estimates.

A. XEntropy A

Form A of the RD curves for spectrum-equalized noise and natural scenes assumes that the curves follow a straight line (in $\log N$) until they reach a dimensionality of 64 and thereafter remain at that value. Figure 13(a) depicts the resulting RD curves under this assumption. The linear portion of each extrapolated RD curve was obtained by fitting the last five measured data points with a first-degree polynomial in $\log N$: $RD(N)=1.13\log N+6.45$ for spectrum-equalized noise, $RD(N)=0.94\log N-0.19$ for natural scenes. The resulting extrapolated proximity distribution functions are shown in Fig. 13(b).

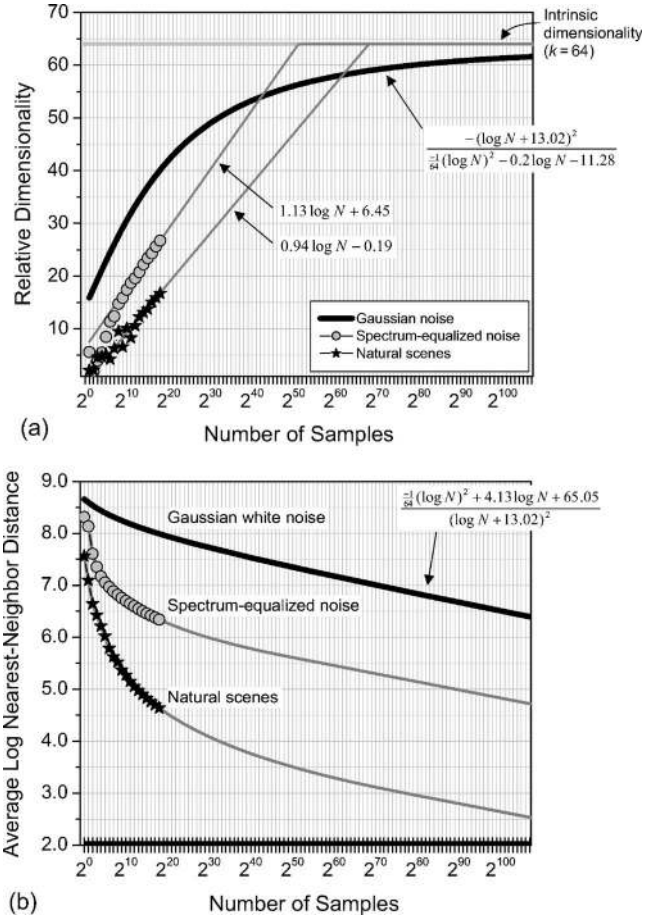

Fig. 13. (a) RD curves and (b) proximity distribution functions for 8×8 Gaussian white noise (black curve), and (a) extrapolated RD and (b) proximity distribution curves for spectrum-equalized noise (gray circles) and natural scenes (stars) by assuming form A of the RD curves. Under from A, the RD curves follow a straight line (in $\log N$) until they hit the dimensionality value of 64.

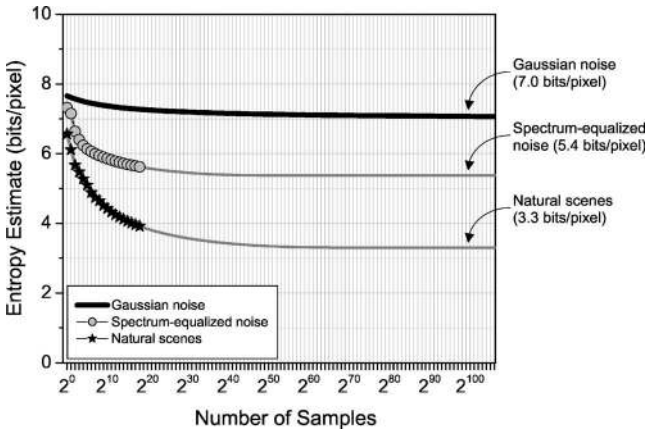

Fig. 14. Entropy estimate for Gaussian white noise and extrapolated entropy estimates (XEn curves) assuming form A of the RD curves (XEntropy A) for spectrum-equalized noise and natural scenes. The entropy estimates computed by using Eq. (8) with $k=64$ and $N=2^{300}$ are 5.4 bits/pixel and 3.3 bits/pixel for spectrum-equalized noise and natural scenes, respectively; the true entropy of spectrum-equalized noise computed via Eq. (9) is 5.1 bits/pixel.

Figure 14 shows the entropy estimate for Gaussian white noise (449 bits; 7.0 bits/pixel) and extrapolated entropy estimates (XEntropy A) for spectrum-equalized noise and natural scenes computed by using Eq. (8) with $k=64$ and $N=2^{300}$. For spectrum-equalized noise, the XEntropy A is 344 bits (5.4 bits/pixel); the actual entropy is 328 bits [5.1 bits/pixel; computed via Eq. (9); see Ref. 44]. For the sample of natural scenes used here, XEntropy A is 212 bits (3.3 bits/pixel).

B. XEntropy B

Form B of the RD curves for spectrum-equalized noise and natural scenes assumes that the curves follow a straight line (in $\log N$) until they intersect with the RD curve for Gaussian white noise, whereupon all subsequent RD values are equivalent to the RD values for Gaussian white noise. Figure 15(a) depicts the resulting RD curves under this assumption; the linear portion of each extrapolated RD curve was obtained as described in Subsection 4.A. The resulting extrapolated proximity distribution functions are shown in Fig. 15(b). Essentially, XEntropy B relies on the idea that the slowest possible falloff in the proximity distribution occurs for Gaussian noise. Therefore, if we assume that the proximity distribution for natural scenes does not decrease any faster that that determined by the linear portion in Fig. 15(a), then XEntropyB provides an upper bound on the entropy. We believe this is a rational assumption and provides a clear bound. However, as we will demonstrate in the following section, we believe that we can provide an extrapolation that provides a more accurate estimate.

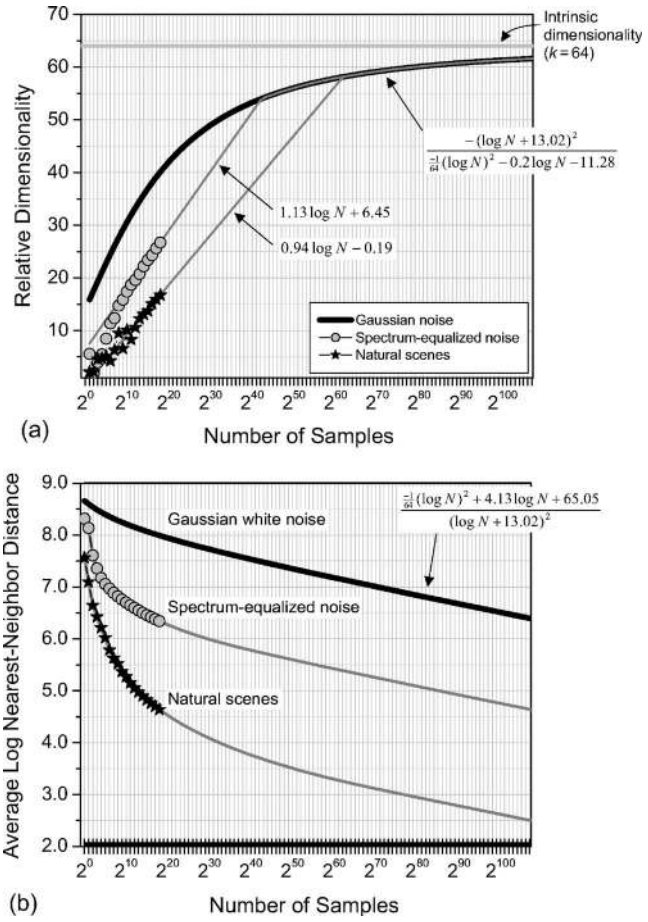

Fig. 15. (a) RD curves and (b) proximity distribution functions for 8×8 Gaussian white noise (black curve), and (a) extrapolated RD and (b) proximity distribution curves for spectrum-equalized noise (gray circles) and natural scenes (stars) by assuming form B of the RD curves. Under form B, the RD curves follow a straight line until they intersect with the RD curve for Gaussian white noise, whereupon all subsequent RD values are equivalent to the RD values for Gaussian white noise.

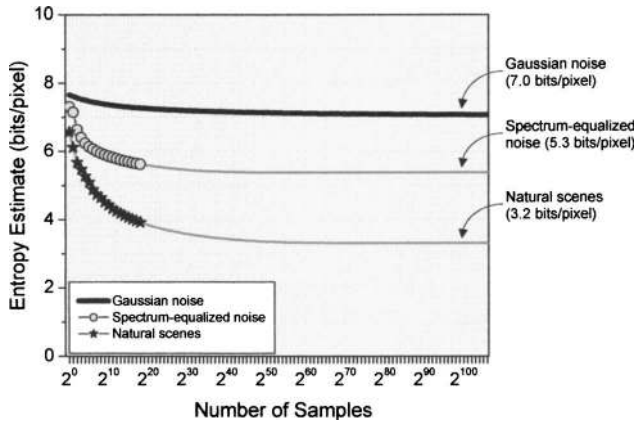

Fig. 16. Entropy estimate for Gaussian white noise and extrapolated entropy estimates (XEn curves) assuming form B of the RD curves (XEntropy B) for spectrum-equalized noise and natural scenes. The entropy estimates computed by using Eq. (8) with $k=64$ and $N=2^{300}$ are 5.3 bits/pixel and 3.2 bits/pixel for spectrum-equalized noise and natural scenes, respectively; the true entropy of spectrum-equalized noise computed via Eq. (9) is 5.1 bits/pixel.

Figure 16 shows the entropy estimate for Gaussian white noise replotted from Fig. 14 (449 bits; 7.0 bits/pixel) and the extrapolated entropy estimates (XEntropy B) for spectrum-equalized noise and natural scenes computed by using Eq. (8) with $k=64$ and $N=2^{300}$. For spectrum-equalized noise, XEntropy B is 337 bits (5.3 bits/pixel); the actual entropy is 328 bits [5.1 bits/pixel; computed via Eq. (9)]. For the sample of natural scenes used here, XEntropy B is 206 bits (3.2 bits/pixel).

C. XEntropy C

While the previous two measures provide upper bounds on the entropy, XEntropy C incorporates our best attempts to extrapolate to the true entropy of the data. Form C of the RD curves for spectrum-equalized noise and natural scenes assumes that the curves are described by the same functional form as the RD curve for Gaussian white noise,

$$RD(N) = -(\log N + b_0)^2/(a_2[\log N]^2 + 2a_2 b_0 \log N + a_1 b_0 - a_0), \quad (12)$$

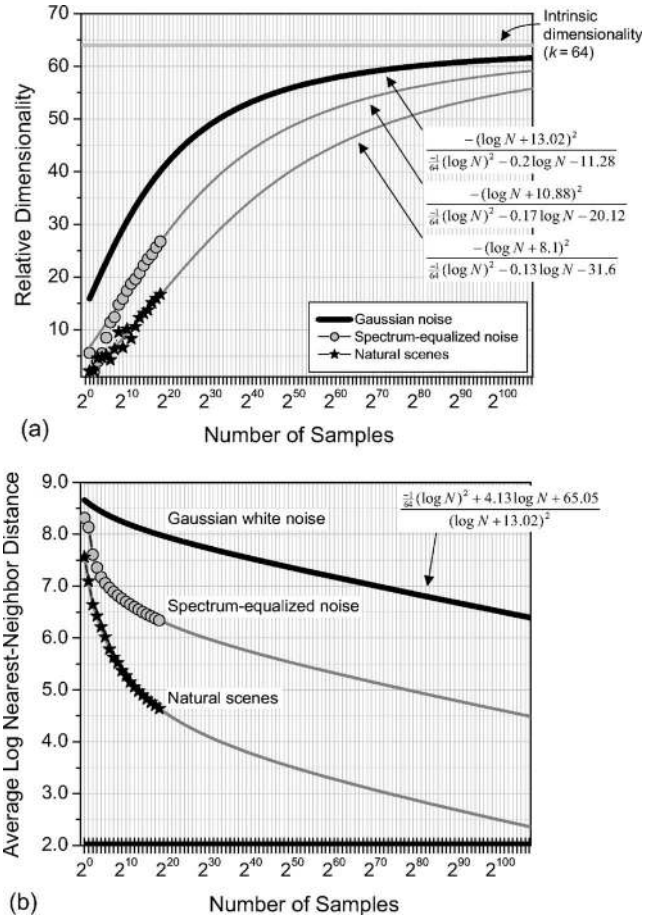

Fig. 17. (a) RD curves and (b) proximity distribution functions for 8×8 Gaussian white noise (black curve), and (a) extrapolated RD and (b) proximity distribution curves for spectrum-equalized noise (gray circles) and natural scenes (stars) by assuming form C of the RD curves. Under form C, the RD curves assume the same functional form as the RD curve for Gaussian white noise [Eq. (12)], where $a_2=-1/64$, $a_1=4.13$, $a_0=65.05$, and the parameter b_0 was adjusted to fit the measured data ($b_0=10.88$ for spectrum-equalized noise, $b_0=8.10$ for natural scenes).

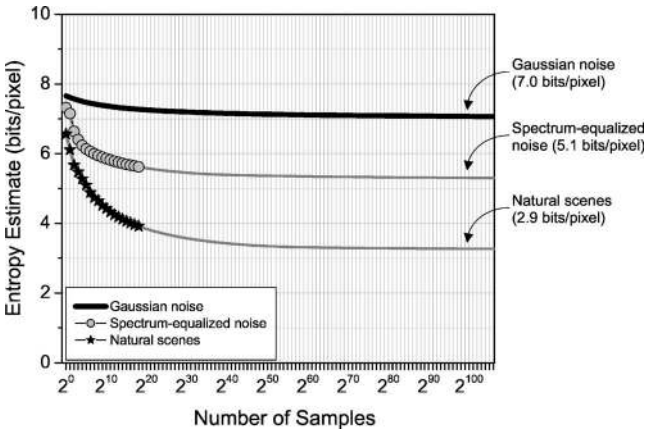

Fig. 18. Entropy estimate for Gaussian white noise and extrapolated entropy estimates (XEn curves) assuming form C of the RD curves (XEntropy C) for spectrum-equalized noise and natural scenes. The entropy estimates computed by using Eq. (8) with $k=64$ and $N=2^{300}$ are 5.1 bits/pixel and 2.9 bits/pixel for spectrum-equalized noise and natural scenes, respectively; the true entropy of spectrum-equalized noise computed via Eq. (9) is 5.1 bits/pixel.

where $a_2=-1/64$, $a_1=4.129$, $a_0=65.05$, and the parameter b_0 was adjusted to fit the measured data. For spectrum-equalized noise $b_0=10.88$, and for natural scenes $b_0=8.10$ as determined by the Nelder–Mead simplex method. Figure 17(a) depicts the resulting RD curves under this assumption. The resulting extrapolated proximity distribution functions are shown in Fig. 17(b).

Figure 18 shows the entropy estimate for Gaussian white noise replotted from Fig. 14 (449 bits; 7.0 bits/pixel) and extrapolated entropy estimates (XEntropy C) for spectrum-equalized noise and natural scenes computed by using Eq. (8) with $k=64$ and $N=2^{300}$. For spectrum-equalized noise, XEntropy C is 324 bits (5.1 bits/pixel), which is very close to the actual entropy of 328 bits [5.1 bits/pixel; computed via Eq. (9)]. For the sample of natural scenes used here, XEntropy C is 184 bits (2.9 bits/pixel).

In summary, among the three extrapolation techniques examined here, we believe that XEntropy C provides the best estimate of entropy. Clearly, a goal of future research is to improve on both the accuracy and robustness of these estimates. Furthermore, the XEntropy C estimate of 184 bits (2.9 bits/pixel) is dependent on the particular sample of images used here. A more extensive sample of natural scenes will certainly give rise to a better estimate of the entropy of natural scenes.

5. OTHER PATCH SIZES

In the previous experiments, patches of size 8×8 pixels were used. To investigate the effects of patch size on entropy, we measured proximity distribution functions for 16×16 patches of Gaussian white noise and natural scenes. Indeed, if the 8×8 subpatches of a 16×16 patch are independent, then one would expect the entropy of the 16×16 patches to be 4 times greater than that of the 8×8 patches. Furthermore, if the subpatches are independent, then we would expect the relative dimensionality for a given proximity to increase by a factor of 4 by doubling the size of the patch (e.g., the *RD* of 16×16 patches at a given proximity would be 4 times the *RD* of 8×8 patches at that same proximity).

Figure 19(a) depicts the proximity distribution functions for patches of size 8×8 (black circles) and 16×16 (white circles) selected from Gaussian white-noise images created with $\sigma=36$ (see Section 2). The proximity distribution function for the 8×8 patches has been offset such that the average log nearest-neighbor distance is 5.0 at $N=1$; accordingly, the proximity distribution function for the 16×16 patches has been offset to maintain the relative vertical displacement between curves. Also shown in Fig. 19(a) (as solid curves) are the predicted proximity distribution functions that would result if the 8×8 subpatches of the 16×16 patches were statistically independent (i.e., requiring 4 times as many samples to achieve the same nearest-neighbor distances as those obtained using the 8×8 patches). Notice that the actual proximity distribution function for the 16×16 Gaussian white-noise patches is very much in agreement with the predicted proximity distribution function, which confirms that the Gaussian white-noise subpatches are indeed independent.

Figure 19(b) depicts the proximity distribution functions for patches of size 8×8 (black circles) and 16×16

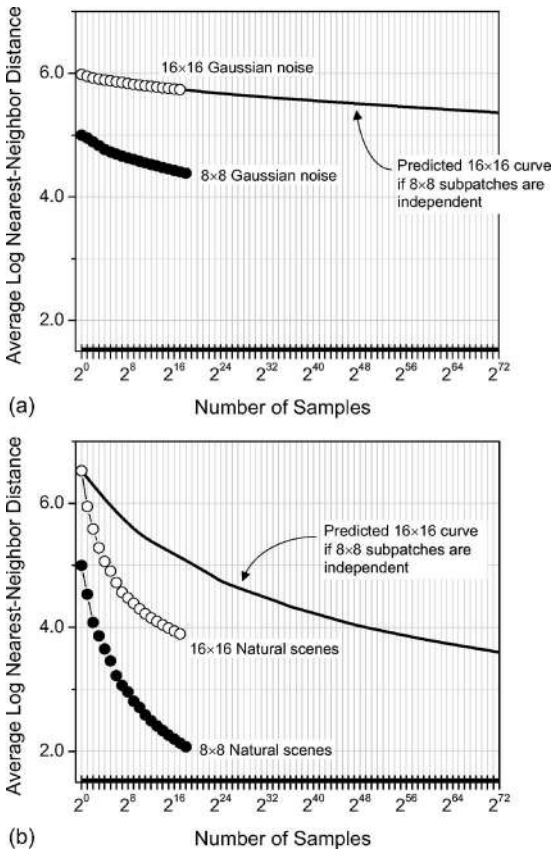

Fig. 19. Proximity distribution functions for patches of size 8×8 (black circles) and 16×16 (white circles). (a) Data for Gaussian white noise; (b) data for natural scenes. The solid black curves in each graph denote the proximity distribution functions that would result if the 8×8 subpatches were statistically independent (thus requiring 4 times the entropy of 8×8 patches to describe a 16×16 patch). Note that the predicted curves have been vertically offset to match their corresponding data.

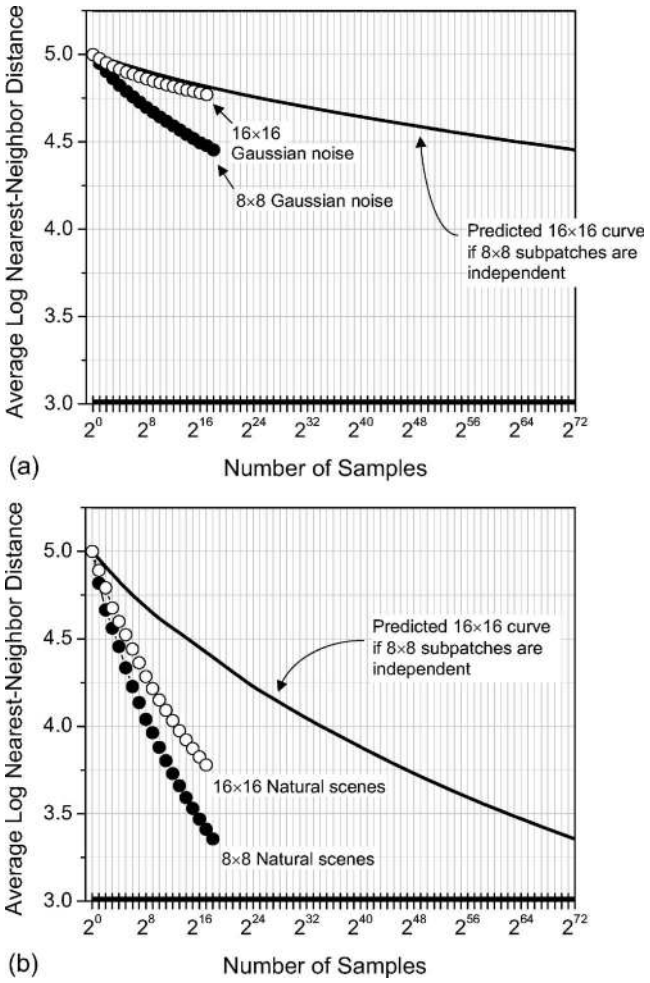

Fig. 20. Proximity distribution functions for patches of size 8×8 (black circles) and 16×16 (white circles) in which each patch was mean and contrast normalized as described in Subsection 3.B. (a) Data for Gaussian white noise; (b) data for natural scenes. The solid black curves in each graph denote the proximity distribution functions that would result if the 8×8 subpatches were statistically independent (thus requiring 4 times the entropy of 8×8 patches to describe a 16×16 patch). Note that the predicted curves have been vertically offset to match their corresponding data.

(white circles) selected from the natural scenes (see Section 2). The proximity distribution function for the 8×8 patches has been offset such that the average log nearest-neighbor distance is 5.0 at $N=1$, and the proximity distribution function for the 16×16 patches have been offset to maintain the relative vertical displacement between curves. Figure 19(b) also shows (as solid curves) the predicted proximity distribution functions that would result if the 8×8 subpatches were independent. Whereas in the Gaussian noise condition where the actual proximity distribution function for the 16×16 patches was similar to the corresponding proximity distribution function predicted assuming independence, here we see that the actual proximity distribution function is substantially lower than the proximity distribution function predicted assuming independence. These data demonstrate that the 8×8 subpatches of the 16×16 patches are not independent; rather, natural scenes demonstrate a marked statistical dependency across space.

Figures 20(a) and 20(b) show corresponding data measured for mean- and contrast-normalized patches (see Subsection 3.B) of size 8×8 and 16×16 pixels. Observe in Fig. 20(a), which depicts the results for Gaussian white-noise patches, that the 8×8 high-contrast subpatches are nearly independent; i.e., the actual proximity distribution function for the high-contrast 16×16 patches (white circles) is similar to the proximity distribution functions predicted assuming independence (solid curves). However, as shown in Fig. 20(b), the data obtained for high-contrast natural-scene patches give rise to a proximity distribution function that is markedly lower than the proximity distribution function predicted assuming independence. These data suggest that the high-contrast *patterns* found in natural scenes demonstrate a statistical dependency across space.

Unfortunately, extrapolation of the proximity distribution for the 16×16 patches is more problematic than for the 8×8 patches because the 16×16 RD curve is very far from converging on the intrinsic dimensionality of 256. However, if we assume that for numbers of samples beyond that measured ($>2^{18}$), the remaining portion of the proximity distribution for the 16×16 patches continues as if the 8×8 subpatches were independent, then we obtain an estimate of 567 bits (2.2 bits/pixel) for the entropy of the 16×16 natural-scene patches. Clearly, obtaining sufficient numbers of samples to extrapolate the proximity distributions for larger patches proves quite difficult. Although we expect further reductions in the entropy rate (bits/pixel) for larger patches, the ultimate entropy one obtains with larger patches (e.g., 256×256) will be a function of both the image content and the noise in the signal.

6. DISCUSSION

In this paper, we have used proximity distributions to investigate the entropy and dimensionality of natural scenes. In general, the technique employed here requires far fewer samples than that required for directly estimating the probability distribution and thereby estimating entropy. For example, for 3×3 patches that follow a uniform distribution, at least 2^{72} samples would be required to measure the probability distribution and thereby measure entropy. However, as the dimensionality grows, and even for 8×8 patches, nearest-neighbor-based techniques too require a prohibitively large number of samples. Although we have proposed three methods of extrapolation, verifying and improving the accuracy of the extrapolations is certainly an area that requires further investigation. Still, by comparing the entropy estimates of different image types, we can gain insight into the contributions of various forms of redundancy to the entropy.

It is generally accepted that the intensity values of images drawn from the natural environment possess a degree of statistical redundancy. Several factors contribute to this redundancy: (1) Natural scenes typically demonstrate $1/f^{\alpha}$ power spectra ($1/f^{\alpha/2}$amplitude spectra) where α is typically in the range of 1.4 to 2.8. The dominance of low spatial frequencies in natural scenes implies slow spatial changes in intensity, and thus neighboring intensity values are spatially correlated. (2) The local structure

in natural scenes is non-Gaussian; rather, marginal probability distributions of discrete cosine transform and discrete wavelet transform coefficients are typically well modeled by using a leptokurtotic generalized Gaussian density.[46] (3) The local mean luminance and local luminance contrast in natural scenes follow a non-Gaussian distribution; many of the patches drawn from natural scenes are devoid of significant contrast.

Although these forms of redundancy have been well studied, there remains the question of how much of the redundancy of images is attributable to each form. Accordingly, in addition to natural scenes, we have measured the entropy of patches of Gaussian white noise and patches of spectrum-equalized noise; and we have measured the entropy of mean- and contrast-normalized versions of all image types. This approach of normalizing the images according to different parameters provides insight into how these different forms of redundancy contribute to the entropy.[46]

A. Effects of Spatial Correlations

Much of the redundancy in natural scenes is commonly attributed to correlations described by the power (amplitude) spectra. Clearly, data that are spatially correlated are also redundant. However, the reverse is not true: Data that are redundant need not be correlated; rather, the redundancies can arise from other forms of statistical dependence. Indeed, several investigators have shown that the statistical dependencies in natural scenes arise from more than just the spatial correlations.[1,4,5,7,8]

Here, we have measured the entropy of 3×3 patches of Gaussian noise, 3×3 patches of natural scenes, and 3×3 patches of Gaussian noise with a power spectrum equivalent to that of natural scenes (spectrum-equalized noise). In addition, we have provided an extrapolated estimate of entropy (XEntropy) of 8×8 patches of these three image types. For the spectrum-equalized noise, the real and imaginary components of each DFT coefficient of the spectrum-equalized noise were drawn from a Gaussian distribution with standard deviation equivalent to the standard deviation measured for the corresponding Fourier components of the natural scenes. Thus, the spectrum-equalized noise and natural scenes possess the same power spectrum, although the distributions of local mean, contrast, and frequency components remained unique for each image type.

For 3×3 patches of Gaussian white noise, the entropy was estimated to be 63 bits (7.2 bits/pixel); this entropy is equivalent to that computed directly via Eq. (9). For 3×3 patches of spectrum-equalized noise, the entropy was estimated to be 49 bits (5.5 bits/pixel), which is also equivalent to the entropy computed directly via Eq. (9). For 3×3 patches of the natural scenes used here, the entropy was estimated to be 35 bits (3.9 bits/pixel). These results reveal that for the sample of natural scenes used here, 3×3 natural scenes have approximately 71% the entropy of 3×3 images with the same power spectra but random phase spectra.

For 8×8 patches of Gaussian white noise, our XEntropy C estimate was 449 bits (7.0 bits/pixel); the actual entropy computed via Eq. (9) was 462 bits (7.2 bits/pixel). The XEntropy C estimate for 8×8 patches of spectrum-equalized noise was 324 bits (5.1 bits/pixel), which is very close to the actual entropy of 328 bits (5.1 bits/pixel) computed via Eq. (9). The XEntropy C estimate for 8×8 patches of natural scenes was 184 bits (2.9 bits/pixel). Although there are certainly limitations to these extrapolated measures, these results suggest that for the sample of natural scenes used here, 8×8 natural scenes have approximately 57% the entropy of 8×8 images with the same power spectra but random phase spectra.

B. Effects of Local Mean and Contrast

In addition to the characteristic power spectrum, natural scenes also exhibit non-Gaussian distributions of local mean luminance and local luminance contrast. As we noted, patches drawn from natural scenes are often devoid of significant contrast. These factors also contribute to the statistical redundancy (reduced entropy) of natural scenes. Accordingly, in Subsection 3.B, we also examined the nearest-neighbor-distance behavior of mean- and contrast-normalized 8×8 patches to investigate the entropy of the underlying *patterns* found in natural scenes without regard to the absolute luminance or RMS contrast.

By normalizing for RMS contrast, the absolute entropy depends on the contrast (variance) to which the data are normalized. Accordingly, here we report entropy estimates relative to the entropy of the mean- and contrast-normalized Gaussian white noise. By applying the XEntropy C extrapolation to the mean- and contrast-normalized proximity distribution functions, we find that for the sample of natural scenes used here, 8×8 high-contrast patches of natural scenes have approximately 57% of the entropy of 8×8 high-contrast patches of Gaussian white noise, and 8×8 high-contrast patches with the same power spectrum as that of 8×8 natural scenes have approximately 87% of the entropy of 8×8 high-contrast patches of Gaussian white noise. Furthermore, 8×8 high-contrast patches of natural scenes have approximately 77% of the entropy of 8×8 high-contrast patches with the same power spectra but random phase spectra.

C. Relative Dimensionality

As mentioned in Section 1, there exists a wide body of research geared toward measuring intrinsic dimensionality[28–34] (see Ref. 35 for a review). Here, we have emphasized the RD of the data as a function of the sampling density. Our main assumption is that, given a sufficiently large number of samples, the RD converges on the intrinsic dimensionality. We have measured the RD of 3×3 and 8×8 patches of Gaussian white noise, spectrum-equalized noise, and natural scenes, as well as the RD of 8×8 patches of $1/f$ and $1/f^2$ noise.

For 3×3 patches (Subsection 2.D), the RD curves for Gaussian white noise, spectrum-equalized noise, and natural scenes all converge on the same (intrinsic) dimensionality of 9, but the curves converge at different rates. Specifically, for samples sizes $<2^{17}$, natural scenes appear lower dimensional than both Gaussian white noise and spectrum-equalized noise (at corresponding sample sizes),

and spectrum-equalized noise appears lower dimensional than Gaussian white noise (at corresponding sample sizes).

For 8×8 patches (Subsection 3.A), our extrapolations are derived from the assumption that the RD curves converge on a value of 64 given a sufficiently large number of samples. For samples sizes $\leqslant 2^{18}$, $1/f^2$ noise appears to be lower dimensional than natural scenes, natural scenes appear lower dimensional than spectrum-equalized noise, spectrum-equalized noise appears lower dimensional than $1/f$ noise, and $1/f$ noise shows lower dimensionality than Gaussian white noise. These ranks are approximately maintained for mean- and contrast-normalized 8×8 patches (Subsection 3.B) with the exception that given $N\geqslant 2^{10}$ samples, natural scenes appear lower dimensional than $1/f^2$ noise.

In contrast to dimensionality-reduction techniques such as principal components analysis or more recently developed nonlinear techniques,[33,34] RD does not specify a particular technique for representing the data given a fixed number of dimensions (e.g., unrolling the Swiss roll), nor does it provide information regarding what the dimensions represent. Instead, the RD of a data set specifies only the dimensionality the data appear to have given the particular sampling density (i.e., the number of samples). Clearly, this RD depends on the technique used to explore the data space; e.g., RD is linked to the sampling method and to the metric used to measure the distance between samples. Here, we have measured RD by using what is arguably one of the simplest approaches: measuring the average (log) distance to the single nearest neighbor for samples drawn randomly from the space. Other techniques, such as using the k nearest neighbors ($k>1$) or using a more uniform sampling technique, may very well lead to different *RD*s.

However, regardless of the approach used to measure RD, the primary utility of the RD curve is its ability to specify the maximum number of samples required to reconstruct the geometry of the data space. Namely, when the RD curve of a data set has converged to the intrinsic dimensionality of the data, there are a sufficient number of samples to uniquely specify the space. Performing the actual reconstruction of the space from those samples is a task suitable for other algorithms.[33,34]

D. Other Estimates of the Entropy of Natural Scenes

Previous researchers have applied different approaches to investigate the entropy of natural images. Parks[47] employed a variant of Shannon's classical guessing game in which human subjects were used as optimal predictors to estimate the entropy of half-tone (binary) images; the entropy of these binary images was estimated to be approximately 0.3 bits/pixel. Tzannes *et al.*[48] used a similar technique to estimate the entropy of 3-bit images; the entropy in Ref. 48 was estimated to be 1.6 bits/pixel. These psychophysical-based approaches were later extended by Kersten[3] to estimate the entropy rate of 4-bit images; Kersten estimated lower and upper bounds on entropy rates of approximately 0.8 and 1.5 bits/pixel, respectively.

Other computational approaches have also been used to investigate the information content in natural scenes. Via a Voronoi tessellation of the space of zero-mean contrast-normalized 3×3 patches, Lee *et al.*[49] have reported that both natural scenes and range images occupy only a small fraction of the surface area of the 7-sphere. More recently, Costa and Hero[50] have developed a measure of Renyi entropy that was used to estimate the entropy of images from the Yale Face Database.

Here, we have used a nearest-neighbor-based technique and an extrapolation (XEntropy C) to estimate an entropy of 184 bits for 8×8 patches of natural scenes. Although differences in patch size and luminance resolution make it difficult to perform a direct comparison of our results with previous estimates, maximum-quality JPEG compression (which is a block-based strategy that operates on 8×8 blocks) provides an average bit rate of 4.1 bits/pixel (263 bits per 8×8 patch) for the natural scenes used in this study, a value which is 42% greater than our estimate of entropy.

Of course, knowledge of the entropy of 8×8 patches does not immediately reveal the entropy of larger-sized images (e.g., images of size 512×512 pixels) unless the 8×8 patches within the larger-sized images are statistically independent. Still, one can use the entropy of 8×8 patches to bound the entropy of larger-sized images. Namely, if $h(\mathbf{X}_{8\times8})$ is the entropy of 8×8 patches of some image class, then the entropy of an $N\times N$ image of that class ($N\geqslant8$) is given by $h(\mathbf{X}_{N\times N})\leqslant(N/8)^2h(\mathbf{X}_{8\times8})$, with equality if the 8×8 patches are independent.

E. Other Applications

The application of nearest-neighbor-based techniques to estimating entropy and dimensionality is not limited to natural scenes. Victor[19] has applied the technique to estimate the entropy of neural spike trains. Kraskov *et al.*[20] has applied a related technique to estimate the mutual information in both gene expression data and ECG signals. Kybic[41] has proposed a related estimate of mutual information for image registration applications. We do, however, wish to note that the estimated entropy is only one component of the analysis in this paper. We believe the full proximity distributions described here provide important insights into the data that go beyond the one number described by the estimated entropy.

We are currently developing extensions of the techniques presented here to investigate the amount of additional information provided by color images (in comparison with luminance-only images); the amount of information provided by the phase spectrum, including measurements of the mutual information between the power and the phase spectra of natural scenes; and the amount of information in natural paintings. In addition, we are investigating the application of the techniques presented here to other types of signals, including natural sounds and video.

Nearest-neighbor-based techniques also have a long history in the field of pattern classification (see Refs. 51 and 52). Indeed, the entropy of a data set is clearly related to the difficulty of classifying data from the set. We believe the use of proximity distributions for natural scenes will prove useful for understanding scene classification and can provide insights into the differences be-

tween two images classes. We are currently investigating the use of proximity distributions for classification purposes.

In theory, the techniques described here can be applied to a wide range of data types. However, it must be stressed that the techniques rely on the assumption that given a sufficient number of samples, the proximity distribution converges to a linear function of the (log) number of samples; i.e., the quantity $-d\log_2(N)/dE\{\log_2 D_N^*\}$ is equivalent to the intrinsic dimensionality of the data for sufficiently large N. Clearly, there exist forms of data for which this assumption does not hold or for which the notion of distance is difficult to quantify (e.g., language), and thus a goal of future research might involve modifications of nearest-neighbor-based techniques and/or the development of proper distance metrics for these types of data.

7. CONCLUSIONS

This paper presented the results of three experiments performed to investigate the entropy and dimensionality of natural scenes. Nearest-neighbor distances were measured for a large collection of samples drawn from various types of images, and the resulting proximity distributions were used to examine the entropies and RDs of the image types.

Our current results indicate that 8×8 natural-scene patches have less than half the entropy of 8×8 Gaussian white-noise patches. This reduction in entropy cannot be attributed solely to the power spectrum, nor can it be attributed to the prevalence of low-contrast patches. Furthermore, the ratio of entropy to patch size decreases with increasing size, suggesting that natural scenes demonstrate a statistical dependency across space. In addition, given $N=2^{18}$ samples, 8×8 natural-scene patches exhibit a RD that is substantially less than the RD of 8×8 Gaussian white-noise patches.

The techniques presented here require far fewer samples than that required to estimate the entropy by first estimating the full probability distribution; however, the presented techniques still possess several limitations. In particular, for the images tested here, even 3×3 patches required roughly 2^{17} samples to obtain accurate estimates of entropy. Although 2^{17} samples is not computationally prohibitive, often one does not have access to this many samples. Furthermore, for 8×8 patches, extrapolations of the data were required; these extrapolations relied on the fact that the RD curves eventually converged on the intrinsic dimensionality of the data, and therefore the extrapolations require knowledge of this intrinsic dimensionality. However, even in those cases where the data have not converged and extrapolations are tentative, the proximity distribution provides important insights into the underlying forms of redundancy. By comparing these distributions for different signal classes (e.g., those with the same power spectra), we can gain insights into the relative contribution of different forms of redundancy.

We certainly do not want to imply that this technique provides a definitive answer to the entropy question. Future research in this area will certainly lead to improved methods of extrapolation and consequently lead to improved estimates of entropy. However, we believe the result provides a new approach to estimate entropy and dimensionality in complex data sets. Our results have so far been limited to relatively small patches, but we believe that with some basic assumptions, we can estimate rational bounds on the entropy for much larger data sets. Overall, we hope this approach will provide insights into both the fundamental limits of compression as well as the question of how different statistical properties relate to the total redundancy that exists in complex data sets.

ACKNOWLEDGMENTS

This work was supported by National Geospatial Intelligence Agency contract HM 1582-05-C-0007 to D. J. Field. We thank Jonathan Victor for his helpful comments on this work, and we thank Richard Darlington for his assistance in deriving Eq. (10).

Corresponding author D. J. Field can be reached by e-mail at djf3@cornell.edu.

REFERENCES AND NOTES

1. D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," J. Opt. Soc. Am. A **4**, 2379–2394 (1987).
2. J. J. Atick, "Could information theory provide an ecological theory of sensory processing?" Network **3**, 213–251 (1992).
3. D. Kersten, "Predictability and redundancy of natural images," J. Opt. Soc. Am. A **4**, 2395–2400 (1987).
4. O. Schwartz and E. P. Simoncelli, "Natural signal statistics and sensory gain control," Nat. Neurosci. **4**, 819–825 (2001).
5. E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," Annu. Rev. Neurosci. **24**, 1193–1216 (2001).
6. W. S. Geisler, J. S. Perry, B. J. Super, and D. P. Gallogly, "Edge co-occurence in natural images predicts contour grouping performance," Vision Res. **41**, 711–724 (2001).
7. D. J. Field, "Scale-invariance and self-similar 'Wavelet' transforms: an analysis of natural scenes and mammalian visual systems," in *Wavelets, Fractals and Fourier Transforms: New Developments and New Applications* (Oxford U. Press, 1993), pp. 151–193.
8. Y. Petrov and L. Zhoaping, "Local correlations, information redundancy, and the sufficient pixel depth in natural images," J. Opt. Soc. Am. A **20**, 56–66 (2003).
9. T. S. Lee, D. Mumford, R. Romero, and V. A. F. Lamme, "The role of the primary visual cortex in higher level vision," Vision Res. **38**, 2429–2454 (1998).
10. B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: a strategy employed by *V*1?" Vision Res. **37**, 3311–3325 (1996).
11. A. J. Bell and T. J. Sejnowski, "The independent components of natural scenes are edge filters," Vision Res. **37**, 3327–3338 (1997).
12. W. B. Pennebaker and J. L. Mitchell, *The JPEG Still Image Data Compression Standard* (Van Nostrand Reinhold, 1993).
13. International Organization for Standardization, "Information technology—JPEG 2000 image coding system: core coding system," Tech. Rep. ISO/IEC FDIS15444-1:2000 (International Organization for Standardization, 2000).
14. N. G. Deriugin, "The power spectrum and the correlation function of the television signal," Telecommun. **1**, 1–12 (1957).
15. D. L. Ruderman and W. Bialek, "Statistics of natural

images: scaling in the woods," Phys. Rev. Lett. **73**, 814–817 (1994).
16. J. Minguillon and J. Pujol, "Uniform quantization error for Laplacian sources with applications to JPEG standard," in *Mathematics of Data/Image Coding, Compression, and Encryption*, M. S. Schmalz, ed., Proc. SPIE **3456**, 77–88 (1998).
17. M. Wainwright, E. P. Simoncelli, and A. Willsky, "Random cascades on wavelet trees and their use in modeling and analyzing natural imagery," Appl. Comput. Harmon. Anal. **11**, 89–123 (2001).
18. L. F. Kozachenko and N. N. Leonenko, "A statistical estimate for the entropy of a random vector," Probl. Inf. Transm. **23**, 9–16 (1987).
19. J. D. Victor, "Binless strategies for estimation of information from neural data," Phys. Rev. E **66**, 051903 (2002).
20. A. Kraskov, H. Stgbauer, and P. Grassberger, "Estimating mutual information," Phys. Rev. E **69**, 066138 (2004).
21. C. E. Shannon, "A mathematical theory of communication," Bell Syst. Tech. J. **27**, 623–656 (1948).
22. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Series in Telecommunications (Wiley, 1991).
23. S. Verdu and T. Han, "The role of the asymptotic equipartition property in noiseless source coding," IEEE Trans. Inf. Theory **43**, 847–857 (1997).
24. E. W. Weisstein, "Birthday problem," from MathWorld—A Wolfram Web Resource, http://mathworld.wolfram.com/BirthdayProblem.html.
25. I. Nemenman, W. Bialek, and R. de Ruyter van Steveninck, "Entropy and information in neural spike trains: progress on the sampling problem," Phys. Rev. E **69**, 056111 (2004).
26. Z. E. Schnabel, "The estimation of total fish population of a lake," Am. Math. Monthly **45**, 348–352 (1938).
27. I. Nemenman, F. Shafee, and W. Bialek, "Entropy and inference, revisited," in *Advances in Neural Information Processing Systems, Vol. 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, eds. (MIT Press, 2002).
28. J. Guckenheimer and G. Buzyna, "Dimension measurements for geostrophic turbulence," Phys. Rev. Lett. **51**, 1438–1441 (1983).
29. R. Badii and A. Politi, "Statistical description of chaotic attractors: the dimension function," J. Stat. Phys. **40**, 725–750 (1985).
30. P. Grassberger, "Generalizations of the Hausdorff dimension of fractal measures," Phys. Lett. A **107**, 101–105 (1985).
31. K. Pettis, T. Bailey, A. K. Jain, and R. Dubes, "An intrinsic dimensionality estimator from near-neighbor information," IEEE Trans. Pattern Anal. Mach. Intell. **1**, 25–36 (1979).
32. W. van de Water and P. Schram, "Generalized dimensions from near-neighbor information," Phys. Rev. A **37**, 3118–3125 (1988).
33. J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," Science **290**, 2319–2323 (2000).
34. S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science **290**, 2323–2326 (2000).
35. J. Theiler, "Estimating fractal dimension," J. Opt. Soc. Am. A **7**, 1055–1073 (1990).
36. K. L. Clarkson, "Nearest neighbor searching and metric space dimensions," in *Nearest-Neighbor Methods for Learning and Vision: Theory and Practice*, G. Shakhnarovich, T. Darrell, and P. Indyk, eds. (MIT Press, 2006), Chap. 2, pp. 15–59.
37. In this paper, we estimate differential entropy assuming that the original images are drawn from an underlying continuous distribution. Under high-rate quantization, the discrete entropy H is related to the differential entropy h by $H \approx h + \log \Delta$, where Δ is the quantization step size (here $\Delta = 1$, $\log \Delta = 0$); see Refs. 19 and 22.
38. J. H. van Hateren and A. van der Schaaf, "Independent component filters of natural images compared with simple cells in primary visual cortex," Proc. R. Soc. London, Ser. B **265**, 359–366 (1998).
39. R. M. Gray and D. L. Neuhoff, "Quantization," IEEE Trans. Inf. Theory **44**, 2325–2384 (1998).
40. see http://redwood.psych.cornell.edu/proximity/.
41. J. Kybic, "High-dimensional mutual information estimation for image registration," in *Proceedings of IEEE International Conference on Image Processing* (IEEE, 2004), pp. 1779–1782.
42. As long as the means are identical, the distance between two patches does not depend on the (common) mean of the underlying Gaussian from which the pixels are drawn.
43. E. W. Weisstein, "Chi-squared distribution," from MathWorld—A Wolfram Web Resource, http://mathworld.wolfram.com/Chi-SquaredDistribution.html.
44. Although the pixel values of the spectrum-equalized noise patches were correlated (and were therefore statistically dependent), the real and imaginary components of the DFT coefficients of each block were independent. Accordingly, the entropy of the spectrum-equalized noise was computed by summing the individual entropies of the real and imaginary part of each DFT coefficient; the individual entropies were computed via Eq. (9).
45. J. A. Nelder and R. Mead, "A simplex method for function minimization," J. Comput. **7**, 308–313 (1965).
46. D. J. Field, "What is the goal of sensory coding?" Neural Comput. **6**, 559–601 (1994).
47. J. R. Parks, "Prediction and entropy of half-tone pictures," Behav. Sci. **10**, 436–445 (1965).
48. N. S. Tzannes, R. V. Spencer, and A. Kaplan, "On estimating the entropy of random fields," Inf. Control. **16**, 1–6 (1970).
49. A. B. Lee, K. S. Pedersen, and D. Mumford, "The nonlinear statistics of high-contrast patches in natural images," Int. J. Comput. Vis. **54**, 83–103 (2003).
50. J. Costa and A. O. Hero, "Geodesic entropic graphs for dimension and entropy estimation in manifold learning," IEEE Trans. Signal Process. **52**, 2210–2221 (2004).
51. B. V. Dasarathy, *Nearest Neighbour (NN) Norms: NN Pattern Classification Techniques* (IEEE, 1973).
52. G. Shakhnarovich, T. Darrell, and P. Indyk, *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice* (MIT Press, 2006).