Local correlations, information redundancy, and sufficient pixel depth in natural images

Yury Petrov

University Laboratory of Physiology, Oxford University, Oxford, OX1 3PT, UK

L. Zhaoping

Department of Psychology, University College London, London, WC1H 0AP, UK

Received May 7, 2002; revised manuscript received July 19, 2002; accepted August 1, 2002

A mathematical framework that permits the factorization of a joint probability distribution into its localized components for a two-dimensional array of pixels is presented. The factorization was used to estimate the contribution to mutual information of two- (I_2) and three-pixel (I_3) luminance correlations for a large ensemble of natural images analyzed at various spatial scales and pixel depths b. It is shown that both I_2 and I_3 saturate at $b \sim 6$ bits per pixel. Three-pixel correlations are shown to produce only a marginal increase of information redundancy (4%) over two-pixel correlations (50%). Implications for neural representation in visual cortex are discussed. © 2003 Optical Society of America

OCIS codes: 330.4270, 200.3050, 110.2960, 070.6020.

1. INTRODUCTION

Of all human senses vision is the most valuable for constructing our representation of the physical world. A visual signal that conveys information from the environment is so rich that even its two-dimensional gray-scale projection on the eye's retina makes a vivid and coherent impression. Through this projection the cohesive structure and mutual relations of physical objects are represented as complex and multifunctional dependencies among local luminance values detected by the retina's photoreceptors. The gargantuan task of decoding these luminance correlations is carried out by largely unknown algorithms in the visual cortex.

Therefore it is not surprising that studies of the luminance correlations in natural images and their effect on neural representation in the visual cortex are of great interest and importance. Statistical analysis of natural images was pioneered shortly after development of information theory by Shannon.[1] First studies were inspired by television industry applications in which one-pixel probability distribution and two-pixel correlation functions,[2] as well as two- and three-pixel probability distributions,[3] were obtained for television frames. The results were used to estimate the information content of the images. Along with early psychophysical studies of gray-level predictability,[4,5] the estimates showed that natural images are highly redundant, mainly because of intensity correlations among adjacent pixels. Shortly thereafter, it was suggested that the role of early visual processing in the brain is to remove these natural redundancies by recoding the incoming signal into a more efficient form[6,7] such that the recoded signal utilizes the least information transmission capacity, measured in bits per second, while preserving the input information. This hypothesized efficient coding process has three main benefits[8]: reduction of the information bottleneck between the retina and the visual cortex, associative learning,[9] and pattern recognition[10,11] facilitation.

The task of finding the optimal encoding strategy can be separated into two independent parts: reduction of information redundancy in a single-neuron response (intraneuron redundancy), and separation of correlated responses of nearby neurons (interneuron redundancy) into statistically independent responses. The intraneuron redundancy can be easily removed by adjusting the gain control (input–output) function of the neuron to the single-pixel statistics of the natural environment.[12] Interneuronal redundancies, especially the second-order correlations between pixels, can be minimized by the optimal design of the neuron's receptive field. Atick and colleagues[8,13–15] have demonstrated that receptive fields of retinal ganglion cells can be successfully explained as local whitening filters that remove the second-order (two-pixel) correlations between pixels.

However, the decorrelation condition does not define the optimal receptive field uniquely, but rather allows a broad class of solutions related by orthogonal transformations.[16,17] The localized, center–surround retinal receptive fields; bandpass, multiscale, and oriented receptive fields of V1 neurons; and even global Fourier filters are all possible solutions for removal of second-order correlations. Therefore further constraints are required to reproduce specifics of neuronal responses in the retina and visual cortex. A particular set of constraints related to object constancy was explored by Li and Atick.[16] They showed that constraint of the neuronal representation to be self-similar under the translation and scaling of input results in a variety of receptive field types similar to those found in V1.

An alternative way to overcome the second-order

decorrelation ambiguity is to search for solutions which also minimize higher-order correlations. This approach gained momentum after the realization that statistics of the natural world are largely non-Gaussian, i.e., they include significant contribution from higher-than-second-order correlations.[18–20] Because natural probability distributions are non-Gaussian (have high kurtosis) it is possible to find optimal (maximally independent) receptive fields by maximizing the non-Gaussianisity of the neuronal output. This technique, known as independent component analysis,[21–24] and related algorithms[17,25,26] were applied to ensembles of natural images to calculate the optimal neuronal representation. The resulting independent components reproduce well such features of V1 receptive fields as orientation selectivity, bandpass, and scaling properties.

Considering the success of these approaches, one is tempted to deduce that cortical receptive fields are *explained* by the minimization of higher-order correlations carried out in V1. However, to make this claim one must demonstrate that the removal of higher-order dependencies results in a substantial increase in the efficient coding benefits listed in the beginning of this section. Basically, all the benefits are based on the reduction in information redundancy which, besides alleviating the bottleneck problem, decreases the number of possible associations as well as the amount of "vocabulary" necessary for pattern recognition. In other words, it is necessary to check that the redundancy due to the third- and higher-order correlations in natural images is rather large when measured in terms of coding bits, which is what matters for data compression and coding efficiency. This was one of the motivations for the present study.

Previous work on higher-order correlations in natural images was mainly focused on quantities (such as kurtosis) that describe the degree to which higher-order moments in single-pixel intensity distribution differ from those of Gaussian distribution with the equivalent second-order statistics.[20] Recent studies also provide some information about the structure of the fourth-[27,28] and higher-order[29] moments defined for more than one pixel. Kurtosis however is not an appropriate measure of information content. To illustrate this point, consider a probability distribution $P(x) \sim \exp[-(|x|/\sigma)^m]$ in which the value of the parameter m defines the degree of peakedness of $P(x)$, while the value of σ for each m is chosen in such a way that all $P(x)$ of different m's have the same second-order statistics, or variance, which is set to one. The entropy of $P(x)$ given by $H = \log 2\Gamma(1/m)^{3/2}/[m\Gamma(3/m)^{1/2}] + 1/m$ (Γ is the gamma function) is shown in Fig. 1 as a function of the distribution kurtosis $\kappa = \Gamma(1/m)\Gamma(5/m)/\Gamma(3/m)^2 - 3$. The value of parameter m was varied from 0.6 to 10 in this plot. Note that the change in m causes a dramatic change in the shape of $P(x)$, which is shown in insets to Fig. 1. It also greatly changes the distribution kurtosis: from -1.1 for $m = 10$ to 12.6 for $m = 0.6$. The entropy, on the other hand, is almost insensitive to the shape changes. This example shows that kurtosis is more a measure of the distribution's shape than its information content.

Nevertheless there has been no exhaustive study exploring the higher-order redundancy in terms of information coding. An earlier study has measured three-pixel redundancy in terms of coding bits, and suggested that it gives a small contribution to the overall redundancy.[3] However, this study ignored the two-dimensional nature of the images and accounted only for the redundancy arising from three neighboring pixels aligned horizontally. Furthermore, it performed the measurement for an ensemble of only a few small television frames at one level of image intensity quantization (64 shades of gray). It will be shown in Section 5 that information redundancy depends strongly on the gray-level quantization, or equivalently on the number of intensity bits per pixel (pixel depth). Therefore statistical analysis of a wide range of pixel depths is required.

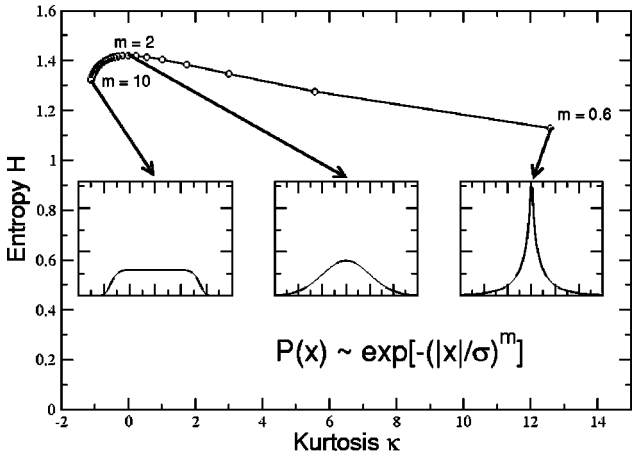

Fig. 1. Entropy as a function of kurtosis for distribution $P(x) \sim \exp[-(|x|/\sigma)^m]$ (see insets) of the same variance or second-order statistics (by adjusting σ with m). Arrows link points on $H(\kappa)$ to corresponding $P(x)$'s.

In Section 2 a mathematical framework accounting for a full set of interpixel redundancies in a two-dimensional array of pixels is presented. The sampling of two- and three-pixel probability distributions based on a large ensemble of natural images is described in Section 3. These probability distributions were used to calculate two-pixel (I_2) and three-pixel (I_3) contributions to mutual information along with corresponding redundancies for three spatial scales and twelve pixel depths. The results and their implications are discussed in Sections 4 and 5.

2. MATHEMATICAL FRAMEWORK

In order to study redundancy that is the result of local correlations one needs to represent the global probability distribution in terms of its local components. First consider a one-dimensional array of enumerated pixels $\{1, 2, 3, \ldots\}$. Denoting the intensity probability distribution for the ith pixel $P(i)$, one has for the whole array

$$\begin{aligned} P(1, 2, 3, \ldots) &= P(1)P(2|1)P(3|2, 1)P(4|3, 2, 1)\ldots = P(1) \\ &\times \left[P(2)\frac{P(2|1)}{P(2)}\right]\left[P(3)\frac{P(3|2)}{P(3)}\frac{P(3|2, 1)}{P(3|2)}\right] \\ &\times \left[P(4)\frac{P(4|3)}{P(4)}\frac{P(4|3, 2)}{P(4|3)}\frac{P(4|3, 2, 1)}{P(4|3,2)}\right]\ldots . \quad (1) \end{aligned}$$

Then for an infinite boundless array $\{\ldots, 1, 2, 3, \ldots\}$, expression (1) becomes

$$P(\ldots, 1, 2, 3, \ldots, i, \ldots) = \prod_{i=-\infty}^{\infty} \tilde{P}(i|i-1, i-2, \ldots), \tag{2}$$

where

$$\tilde{P}(i|i-1, i-2, \ldots) = P(i)\frac{P(i|i-1)}{P(i)} \times \frac{P(i|i-1, i-2)}{P(i|i-1)} \frac{P(i|i-1, i-2, i-3)}{P(i|i-1, i-2)} \ldots \tag{3}$$

corresponds to the ith pixel probability distribution conditioned on the intensity of all the previous pixels. Clearly, Eq. (2) holds for an arbitrary pixel enumeration. Since for natural images, intensity correlations fall quickly with pixel-to-pixel separations, the consecutive enumeration makes the above factorization the most localized in the sense that the multipixel fractions in Eq. (3) quickly converge to unity. For the consecutive enumeration, the nth fraction in each product characterizes the conditioning that two pixels separated by $n-1$ pixels impose on each other in addition to the effect of the $n-1$ intermediate pixels. As such the nth fraction represents a contribution from n-point correlations in addition to lower-order correlations.

By taking the negative mean of the logarithm of expression (2) one obtains the total entropy of the system

$$\begin{aligned} H &= -\mathbf{E}\log_2[P(\ldots, 1, 2, 3, \ldots, i, \ldots)] \\ &= -\sum_{i=-\infty}^{\infty} \mathbf{E}\log_2\left[P(i)\frac{P(i|i-1)}{P(i)}\frac{P(i|i-1, i-2)}{P(i|i-1)}\ldots\right] \\ &= -\sum_{i=-\infty}^{\infty} \left\{\mathbf{E}\log_2[P(i)] + \mathbf{E}\log_2\left[\frac{P(i|i-1)}{P(i)}\right] + \mathbf{E}\log_2\left[\frac{P(i|i-1, i-2)}{P(i|i-1)}\right]\ldots\right\}, \end{aligned} \tag{4}$$

where $\mathbf{E}$ denotes the expectation value (mean) over the intensity values for all pixels in a given term. Assuming that probability distributions do not depend on i (i.e., the image ensemble is stationary), Eq. (4) becomes

$$H = N[H_1 - I_2 - I_3 - I_4 - \ldots], \tag{5}$$

where N stands for the total number of pixels in the array and

$$\begin{aligned} H_1 &= -\mathbf{E}\log_2[P(i)], \\ I_2 &= \mathbf{E}\log_2\left[\frac{P(i|i-1)}{P(i)}\right], \\ I_3 &= \mathbf{E}\log_2\left[\frac{P(i|i-1, i-2)}{P(i|i-1)}\right], \\ &\ldots. \end{aligned} \tag{6}$$

Thus the total entropy of an image ensemble can be decomposed into single-pixel entropy (NH_1) minus two-point (NI_2), three-point (NI_3), etc., contributions to the total redundancy. What is notable, the nth mutual information contribution accounts only for additional mutual information that is not present in the $n-1$ and lower-order terms. Therefore by defining information redundancy that is a result of interpixel correlations as $1 - H/(H_1N) = I_2/H_1 + I_3/H_1 + \ldots$, one can access the redundancy in an orderly manner.

Before extending the above analysis to two-dimensional pixel arrays, it is necessary to make the following observations. First all fractions in expression (2) are invariant to swapping the end pixels, i.e.,

$$\begin{aligned} P(a|b, c|d) &\equiv \frac{P(a|b, c, d)}{P(a|b, c)} \equiv \frac{P(a, b, c, d)P(b, c)}{P(a, b, c)P(b, c, d)} \\ &= P(d|b, c|a). \end{aligned} \tag{7}$$

Second, swapping the end pixel with one of the intermediate pixels can be carried out using the relationship

$$P(a|b, c|d) = P(b|a, c|d)\frac{P(a|c|d)}{P(b|c|d)}, \tag{8}$$

where a and b pixels were swapped. This formula extends to any number of the intermediate pixels in a trivial way. Now we proceed to the analysis of two-dimensional images.

All the formulas above are applicable once the array of pixels is enumerated. However there is an important difference between enumeration in one dimension and enumeration in two dimensions. In a two-dimensional array some neighboring pixels are assigned numbers which are very far apart no matter how the enumeration is carried out. For example if one enumerates pixels within rows first, then pixels next to each other within a column will have numbers separated by the number of pixels in a row. Therefore factorization (2) will no longer be well localized and some refactorization is necessary. The goal of this section is to describe such refactorization for three-pixel factors.

Since the strongest three-pixel correlations are between the closest neighbors, we limit the correlation neighborhood of a pixel by two interpixel distances, as shown in Fig. 2(a) by dashed curves. In terms of decomposition (2) this means that all nodes which are more than two pixels away from either of the end pixels can be omitted. In Fig. 2(a) these nodes are shown by open circles. Consider the enumeration order shown. Using expression (3) for a given pixel $i = 7$ marked by the concentric circles

$$\begin{aligned} \tilde{P}(7|6, 5, \ldots) &= P(7)P(7\|6)P(7|6|5)P(7|6, 5|4)P(7|6, 5, 4|3) \\ &\quad \times P(7|6, 5, 4, 3|2)P(7|6, 5, 4, 3, 2|1) \\ &= P(7)P(7\|6)P(7|6|5)P(7|6, 5|4)P(7|6, 4|3) \\ &\quad \times P(7|4, 3|2)P(7|4, 3, 2|1), \end{aligned} \tag{9}$$

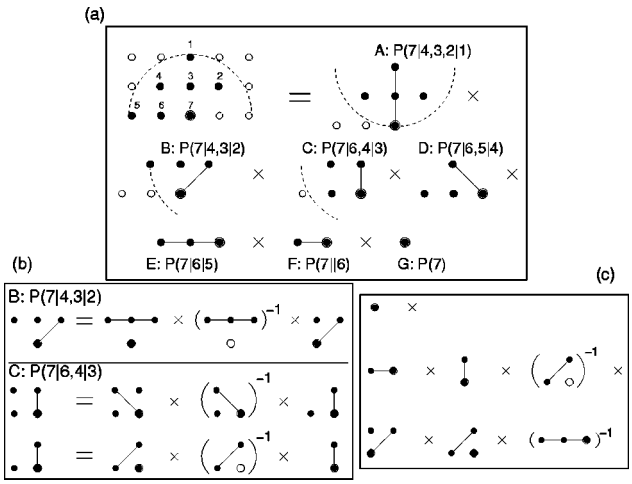

Fig. 2. (a) Joint probability factorization for a square grid. Dashed curves indicate the maximum correlation radius used for three-pixel factors. Open circles mark nodes in the factorization beyond the correlation radius. The set of diagrams illustrates factors in Eq. (9) (shown in reversed order here); gray lines connect the end pixels for each factor. (b) Correlation factors B and C [see Fig. 2(a)] transformed using Eqs. (7) and (8) to the effect of factoring out two- and three-pixel correlations. The second line in C shows transformation of the last factor from the previous line. (c) One-, two-, and three-pixel correlation factors resulting from the local probability factorization.

where nodes outside the two-pixel correlation radius were omitted in the second equality. As defined by Eq. (7), $P(7\|6)$ notation stands for the $P(7|6)/P(7)$ factor. This factorization can be conveniently represented by a set of diagrams, each representing a factor in Eq. (9), as shown in Fig. 2(a). Gray lines indicate conditioning between end pixels described by the corresponding factor. In this graphical representation the need for partial refactorization becomes clear. In particular, diagram C in Fig. 2(a) represents the $P(7|6, 4|3)$ factor in Eq. (9), which describes mutual conditioning between pixels 3 and 7 given certain values for pixels 4 and 6. Since this conditioning clearly has a major direct $P(7\|3)$ contribution, the $P(7\|3)$ correlation term must be factored out of the remaining multipixel correlations. In this way Eq. (9) can be localized by minimizing the contribution from multipixel correlation factors. The required minimization is achieved by applying relations (7) and (8) to each diagram to maximize the two- and three-pixel correlation terms, which are factored out in the process. Figure 2(b) represents this refactorization process. In essence, the distance between the end pixels is maximized for all resulting factors.

Refactorization is not necessary for the first diagram and the last three diagrams in Fig. 2(a). For the remaining diagrams (B, C, and D) the transformation steps are given by the formulas below and are illustrated in Fig. 2(b). For diagram B [Fig. 2(b)] and D, respectively,

$$P(7|4, 3|2) = P(4|7, 3|2)\frac{P(7|3|2)}{P(4|3|2)}, \tag{10}$$

$$P(7|6, 5|4) = P(7|6, 4|5)\frac{P(7|6|4)}{P(7|6|5)}. \tag{11}$$

For diagram C two steps are required to minimize multipixel contributions for all resulting factors [Fig. 2(b)]. These are

$$P(7|6, 4|3) = P(7|6, 3|4)\frac{P(7|6|3)}{P(7|6|4)},$$

$$P(7|6|3) = P(6|7|3)\frac{P(7\|3)}{P(6\|3)}. \tag{12}$$

Collecting one-, two-, and three-pixel factors from Eqs. (9), (10), (11), and (12), one obtains

$$\tilde{P}(7|6, 5, \dots) \simeq P(7)\frac{P(7\|6)P(7\|3)}{P(6\|3)}\frac{P(7|3|2)P(6|7|3)}{P(4|3|2)}, \tag{13}$$

as is illustrated in Fig. 2(c). After taking the negative mean of the logarithm of relation (13) one obtains for the I_2 and I_3 terms

$$I_2 = \mathbf{E}\{\log_2[P(7\|6)] + \log_2[P(7\|3)] - \log_2[P(6\|3)]\}, \tag{14}$$

$$I_3 = \mathbf{E}\{\log_2[P(7|3|2) + \log_2(P(6|7|3)] - \log_2[P(4|3|2)]\}. \tag{15}$$

Note that the resulting information decomposition is not unique. Specifically the orientation of the diagonal in the diagonal and corner factors in Fig. 2(c), as well as the orientation (vertical or horizontal) of the three-pixel row factor depends on the enumeration chosen in Fig. 2(a). Analysis of a large ensemble of images described in Sections 3 and 4 shows that the choice of orientation in the diagonal and corner factors is immaterial since the ensemble information is symmetric relative to the left $\leftrightarrow$ right reflection. Choice of the three-pixel row configuration orientation makes a slight difference since the vertical orientation contributes slightly more mutual information. Since the sign of the row term is negative in relation (15), to maximize I_3 we chose the horizontal orientation as shown in Fig. 2(c).

3. PROBABILITY DISTRIBUTION SAMPLING

An ensemble of 4212 linearized gray-scale images of 1024×1536 taken from the van Hateren and van der Schaaf database[23] was used to sample the pixel-intensity probability distribution at three spatial scales ($1\times$, $2\times$, and $3\times$). The larger-spatial-scale images were obtained by block averaging the originals over 2×2 and 3×3 pixel blocks. Examples of the images can be viewed at http://hlab.phys.rug.nl/archive.html. First, three-pixel-intensity joint probability distributions were accumulated, then two-pixel and single-pixel distributions were obtained by consecutively summing the three-pixel matrices over one of the dimensions. Images were used as provided in the database, i.e., at an optimal exposure setting for each picture without taking the corresponding intensity normalization factor into consideration. In other words, statistics of natural images were collected as projected on the retina with a camera aperture functioning as a pupil adjusting to the overall luminance level, rather

than in the outside world *per se*. Also no log-transform mimicking the first stages of visual processing was applied to the raw pixel intensities. It will be shown in Section 5 that any point transformation such as the intensity log-transform does not change mutual information nor, consequently, the bulk of our results. In order to study pixel depth dependence, lower-bit distributions were obtained from the distributions originally sampled at 9 bits per pixel (12 bits per pixel for one- and two-pixel distributions) by consecutive block summing over neighboring probability matrix elements, reducing each matrix dimension twofold at every step [see Eqs. (22) and (24)].

Besides the original database pixel depth (~13 bits per pixel), sampling noise due to the limited number of images in the ensemble influences statistical analysis. Given about 6.6×10^9 pixels in the ensemble one obtains an average of 50 counts per element of a $512 \times 512 \times 512$ (9 bits per pixel) three-pixel joint probability matrix. In addition the probability distribution is highly peaked along its main diagonal, and more than half of the probability matrix elements are equal to zero. This makes three-pixel probability matrices at large pixel depths rather noisy.

It is straightforward to show that sampling noise causes an underestimation of distribution entropy and an overestimation of mutual information. Let N_i^{noisy} be the number of counts accumulated for a given pixel intensity i, and let it be Poisson distributed with the true mean N_i. Since $P_i^{\text{noisy}} \equiv N_i^{\text{noisy}}/N$, where N is the total number of sampled pixels, then P_i^{noisy} can be written as $P_i + \nu$, where $P_i = N_i/N$ is the true value of the probability and the noise term ν has the mean $\mathbf{E}\nu = 0$ and the dispersion $\mathbf{E}\nu_i^2 = N_i/N^2$. The effect of noise on the distribution entropy is then given by

$$\begin{aligned} H^{\text{noisy}} - H &= \sum_{i=1}^{n} P_i \log_2 P_i - \sum_{i=1}^{n} (P_i + \nu_i)\log_2(P_i + \nu_i) \\ &= -\sum_{i=1}^{n} \nu_i \log_2 P_i - \sum_{i=1}^{n} P_i\left(1 + \frac{\nu_i}{P_i}\right)\log_2\left(1 + \frac{\nu_i}{P_i}\right). \end{aligned} \tag{16}$$

We divide the second summation in (16) into two parts. The first part consists of m terms for which $|\nu_i/P_i| \leqslant 1$, so that the logarithm in (16) can be expanded in a power series as

$$\sum_{i=1}^{m} P_i\left(1 + \frac{\nu_i}{P_i}\right)\log_2\left(1 + \frac{\nu_i}{P_i}\right) = \frac{1}{\log 2}\sum_{i=1}^{m} P_i\left(1 + \frac{\nu_i}{P_i}\right)\sum_{k=1}^{\infty} \frac{-1}{k}\left(-\frac{\nu_i}{P_i}\right)^k. \tag{17}$$

The second part consists of $n - m$ terms for which $\nu_i/P_i > 1$:

$$\sum_{i=m+1}^{n} (P_i + \nu_i)\log_2\left(1 + \frac{\nu_i}{P_i}\right) \equiv R > 0 \tag{18}$$

These terms correspond to pixel intensities for which noise ν_i is larger than P_i. The probability of this happening for the Poisson-distributed N_i^{noisy} decreases with the number of accumulated counts (it is 8% for $N_i = 1$, 5% for $N_i = 2$, 3% for $N_i = 3$, etc.).

Because $(1 + x)\log(1 + x) > x + x^2/2 - x^3/6$, we want to keep the first three terms in Eq. (17). Then

$$H^{\text{noisy}} - H < -\sum_{i=1}^{n} \nu_i \log_2 P_i - \frac{1}{\log 2}\sum_{i=1}^{m} \nu_i - \frac{1}{2\log 2}\sum_{i=1}^{m}\left(\frac{\nu_i^2}{P_i} - \frac{1}{3}\frac{\nu_i^3}{P_i^2}\right). \tag{19}$$

The third moment of ν_i is $\mathbf{E}(N_i^{\text{noisy}} - N_i)^3/N_i^3 = N_i/N^3$; therefore, after taking the expectation value of relation (19),

$$\mathbf{E}H^{\text{noisy}} < H - \frac{1}{N2\log 2}\sum_{i=1}^{m}\left(1 - \frac{1}{3N_i}\right), \tag{20}$$

which proves that sampling noise reduces the entropy.

For the two-pixel mutual information $I^{1,2}$,

$$I_{1,2}^{\text{noisy}} - I_{1,2} \equiv (H_1^{\text{noisy}} - H_1) + (H_2^{\text{noisy}} - H_2) - (H_{1,2}^{\text{noisy}} - H_{1,2}).$$

The power series in Eq. (17) is well approximated by the first three terms. Omitting fourth- and higher-order noise terms,

$$\mathbf{E}I_{1,2}^{\text{noisy}} - I_{1,2} \simeq \frac{1}{N2\log 2}\left[\sum_{i,j=1}^{m'}\left(1 - \frac{1}{3N_{ij}}\right) - 2\sum_{i}^{m}\left(1 - \frac{1}{3N_i}\right)\right] + R_{1,2} - R_1 - R_2. \tag{21}$$

It is easy to show that this expression is positive. The first summation involves double summation, which makes it $\sim n^2$, while the second summation is only $\sim n$. Also, because $N_i = \Sigma_{j=1}^{n} N_{ij}$, there are n undersampled elements in P_{ij} for one such element in P_i. Therefore, $R_{1,2} > R_1 + R_2$ because the input to R terms is mainly through the strongly undersampled elements. Thus the overall effect of sampling noise is to cause a fast (quadratic) growth of mutual information as a function of n, or equivalently, an exponential growth as a function of pixel depth b (since $n = 2^b$). In simple terms, the sampling noise increases mutual information because P_{ij} is more undersampled than P_i, and the resulting decrease of $H_{1,2}$ is larger than that for H_1 and H_2.

4. RESULTS

The joint probability distributions obtained were transformed into local correlation factors $P(i|j|k)$ as defined by

relation (7) and then into local information distributions $P(i, j, k)\log_2 P(i|j|k)$. The resulting matrices are shown in Fig. 3.

Altogether, four two-pixel factors $P(i, j)\log_2 P(i\|k)$ (corresponding to vertical, horizontal and diagonal pixel pair geometries) were calculated [Fig. 3(a)]. Positive and negative matrix elements shown by shades of white and black indicate, respectively, where the two-pixel joint probability is higher or lower than the probability if the two pixels were uncorrelated. Six three-pixel factors $P(i, j, k)\log_2 P(i|j|k)$—vertical row, horizontal row and four corner patterns ('⌈' '⌉' '⌊' '⌋')—were also calculated. The three-dimensional matrices corresponding to the vertical row [Fig. 3(b)] and one corner configuration [Fig. 3(c)] are shown by means of their four planar cross sections. The underlying three-pixel correlations are best illustrated by the last cross section in Fig. 3(b) and Fig. 3(c), where the intensity value of pixel 2 is fixed at 1/4 of the intensity range. Nonzero matrix elements show where the joint distribution for pixels 1 and 3 deviates from the uncorrelated distribution.

It is instructive to compare images in the last column of Fig. 3. The localized bright area along the main diagonal in the two-pixel information distribution matrix (the top image) indicates that the intensity of pixels 1 and 3 is positively correlated. The three-pixel matrices, on the other hand, indicate that pixels 1 and 3 are not only correlated through pixel 2 but are directly correlated with each other. The dark cross structure with a bright dot in the center indicates that if the intensity value of either pixel 1 or pixel 3 is equal to that of pixel 2, the probability distribution for the remaining pixel is sharply peaked around the same value. Otherwise, corner and row configuration distributions differ markedly in location of the rest of the positive correlations. For the corner configuration they are distributed along the main diagonal with very weak lobes extending along the other matrix diagonal. This indicates that the intensity of pixels 1 and 3 tends to be either brighter or darker than that of pixel 2. For the row configuration, on the other hand, the positive lobes are strongest along the orthogonal diagonal. Therefore if pixel 1 is brighter than 2, then 3 is darker, and vice versa, if 1 is darker, then 3 is brighter. In other words luminance changes gradually from 1 to 2 to 3. The corner configuration correlations can be understood in the same way: as an oblique luminance gradient from pixels 1 and 3 to pixel 2. Thus three-pixel dependencies encode the probability distribution for the local luminance gradient. A weaker positive lobe along the main diagonal of the row matrix describes the probability distribution for relatively few luminance extrema in the image ensemble.

By summing over elements of the matrices one calculates the contribution to I_2 and I_3 from the different pixel configurations as given by relations (14) and (15). The results for 1× scale are given in Table 1. The terms obtained are added to the overall I_2 and I_3 mutual information contributions. The resulting H_1, I_2, and I_3 for three spatial scales are given in Table 2 and illustrated in Fig. 4 as a function of pixel depth b. Statistical errors due to sampling noise were estimated by adding the Poisson noise to matrix elements and calculating the resulting standard deviation for each information term. Because in all cases the errors were less than 0.001 bits, they are not indicated in Fig. 4.

The pixel depth dependence is such that while the single-pixel entropy grows proportionally to the pixel depth, both I_2 and I_3 terms appear to saturate around 2.5 and 0.2 bits, respectively. Horizontal row statistics for

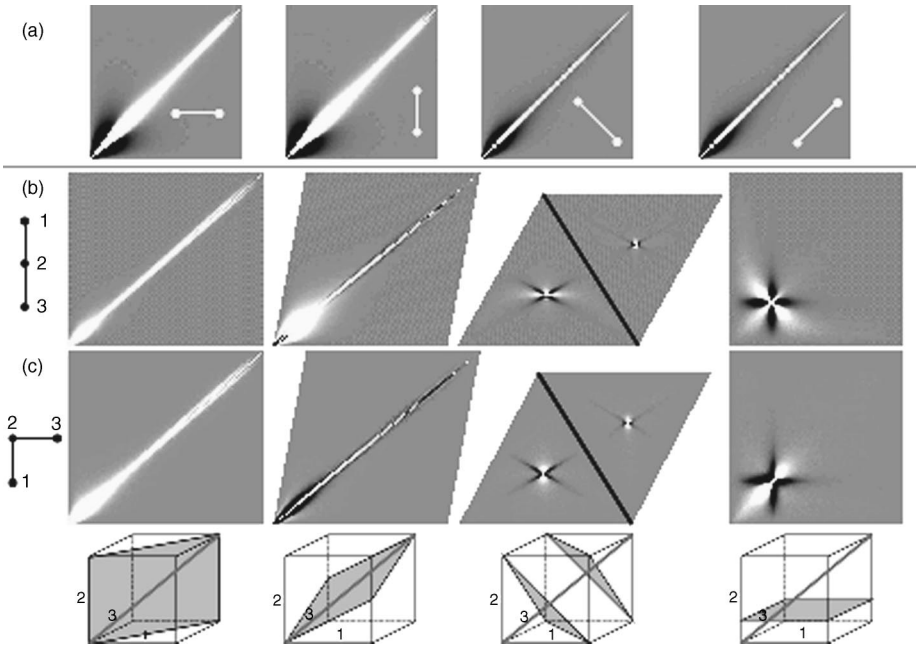

Fig. 3. Information distribution matrices for $b = 7$. Lower left and upper right corners in each matrix correspond to (0, 0) and (127, 127) pixel pair intensity, respectively, while positive and negative elements are shown by light and dark shades. (a) Two-pixel matrices with corresponding pixel geometry shown in the insets. (b) Vertical row three-pixel matrix and (c) one of the four corner, three-pixel matrices, visualized by means of four planar cross sections. Corresponding three-dimensional cuts are shown below each cross section; thick line marks the main diagonal.

Table 1. Individual Contributions from Two-Pixel and Three-Pixel Configurations to Mutual Information at Scale 1× and 12 Pixel Depths

N_{bits}	1	2	3	4	5	6	7	8	9	10	11	12
hor. pair	0.19	0.48	0.94	1.45	1.83	2.07	2.21	2.27	2.29	2.30	2.30	2.31
vert. pair	0.19	0.47	0.93	1.41	1.77	2.00	2.12	2.18	2.20	2.21	2.22	2.22
diagonal	0.18	0.44	0.85	1.29	1.63	1.83	1.94	1.99	2.01	2.02	2.02	2.03
hor. row	0.005	0.015	0.034	0.060	0.087	0.113	0.131	0.142	0.154	—	—	—
vert. row	0.005	0.017	0.039	0.067	0.095	0.119	0.135	0.148	0.162	—	—	—
'[' corner	0.009	0.029	0.065	0.101	0.130	0.151	0.163	0.173	0.186	—	—	—
'⌊' corner	0.008	0.028	0.063	0.100	0.129	0.148	0.160	0.170	0.183	—	—	—

Table 2. One-, Two-, and Three-Pixel Contributions to Information Content of Natural Images for 3 Spatial Scales and 12 Pixel Depths[a]

N_{bits}	1	2	3	4	5	6	7	8	9	10	11	12
H_1	0.24	0.70	1.51	2.48	3.45	4.43	5.42	6.41	7.41	8.41	9.41	10.41
	0.23	0.69	1.51	2.48	3.45	4.43	5.41	6.41	7.41	8.40	9.40	10.40
	0.22	0.68	1.50	2.47	3.44	4.42	5.40	6.40	7.40	8.40	9.40	10.40
I_2	0.20	0.51	1.02	1.56	1.98	2.24	2.39	2.46	2.48	2.49	2.50	2.50
	0.19	0.49	0.98	1.51	1.93	2.21	2.37	2.46	2.49	2.51	2.51	2.52
	0.18	0.47	0.94	1.46	1.86	2.12	2.27	2.35	2.39	2.40	2.41	2.42
I_3	0.013	0.044	0.095	0.144	0.173	0.188	0.195	0.203	0.217	—	—	—
	0.011	0.039	0.085	0.124	0.144	0.153	0.156	0.160	0.182	—	—	—
	0.011	0.037	0.082	0.121	0.142	0.152	0.157	0.167	0.202	—	—	—

[a] Results for scales 1×, 2×, and 3× are given in first, second, and third row for each group.

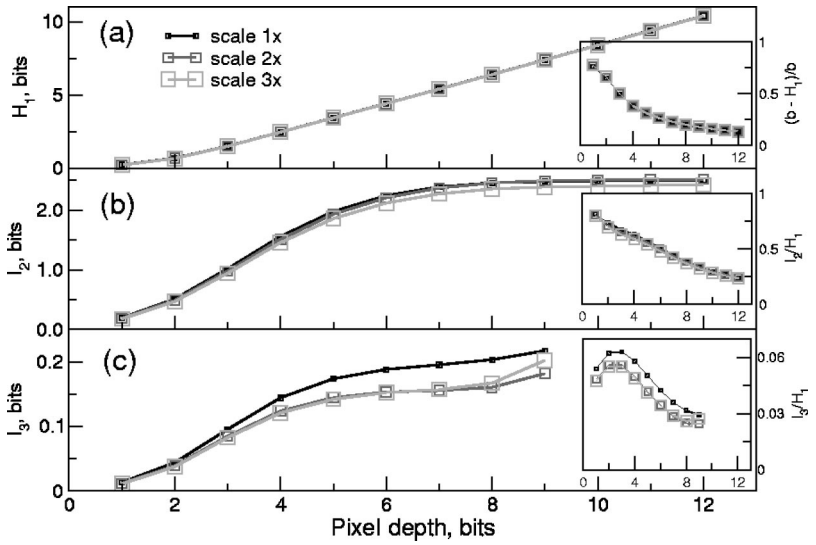

Fig. 4. Local information as a function of pixel depth and spatial scale. Small, medium, and large squares represent scales 1×, 2×, and 3×, respectively. (a) Single-pixel entropy. (b) Two-pixel mutual information as given by relation (14). (c) Three-pixel contribution to mutual information as given by relation (15). Redundancy factors are shown as insets: (a) intrapixel redundancy factor; (b), (c) two- and three-pixel contributions, respectively, to the interpixel correlation redundancy factor.

$b = 6$ bits per pixel were previously computed from three horizontally adjacent pixels in a television raster scan by Schreiber.[3] The resulting I_2 value averaged between two images was 2.4 bits, which is slightly higher than our value (~2.07 bits for $b = 6$). The three-pixel information I_3 that can be calculated from the three- and two-pixel entropies reported by Schreiber[3] is 0.43 bits, which is significantly higher than our result (0.11 bits). Considering the combination of sampling noise (two small images were used by Schreiber) and video signal noise, the high mutual information values are not surprising, particularly for I_3. Accordingly the sharp increase of I_3 in our data between $b = 7$ and 9 bits per pixel shown in Fig. 4 is attributed to the sampling noise effects described in Section 3. Indeed one can see that the increase becomes progressively more pronounced for the 2× and 3× spatial scales, for which, respectively, 1/4 and 1/9 of the original 6.6×10^9 pixels were sampled. Otherwise the spatial

scale appears to produce only a minor effect on all information terms, the largest difference being observed between scales 1× and 2× for the three-pixel information. This agrees with approximate scale invariance in natural images.[28,30–32]

5. DISCUSSION

The observed dependence of information terms on pixel depth is explained below. Consider moving from a certain pixel depth $b+1$ to b. This corresponds to changing the total number of intensity values from $2n$ to n, where $n=2^b$, by summing the $2n$ probability distribution over pairs of neighboring elements as

$$P_n(i/2) = P_{2n}(i) + P_{2n}(i+1). \tag{22}$$

Denoting

$$P_1 = P + \delta P = P_{2n}(i),$$
$$P_2 = P - \delta P = P_{2n}(i+1),$$

where $P=(P_1+P_2)/2$, the ith pair input to the single-pixel entropy change is given by

$$\begin{aligned} H_i^1(2n) - H_i^1(n) &= (P_1+P_2)\log_2(P_1+P_2) - P_1\log_2 P_1 - P_2\log_2 P_2 \\ &= 2P - P\log_2\left[1-\left(\frac{\delta P}{P}\right)^2\right] + \delta P \\ &\quad\times\left[\log_2\left(1-\frac{\delta P}{P}\right) - \log_2\left(1+\frac{\delta P}{P}\right)\right]. \end{aligned}$$

Because $|\delta P/P|<1$ we can expand all logarithms as power series of $\delta P/P$. After some regrouping of terms in the power series one obtains

$$H_i^1(2n) - H_i^1(n) = 2P - \frac{1}{\log 2}\sum_{k=1}^{\infty}\frac{1}{k(2k-1)}\frac{(\delta P)^{2k}}{P^{2k-1}}.$$

The single-pixel entropy change is then given by summing the input from all pairs as

$$\begin{aligned} H^1(2n) - H^1(n) &= \sum_{i=1}^{n} 2P_i - \frac{1}{\log 2}\sum_{i=1}^{n}\sum_{k=1}^{\infty}\frac{1}{k(2k-1)}\frac{(\delta P_i)^{2k}}{P_i^{2k-1}} \\ &= 1 - \frac{1}{\log 2}\sum_{i=1}^{n}\sum_{k=1}^{\infty}\frac{1}{k(2k-1)}\frac{(\delta P_i)^{2k}}{P_i^{2k-1}}, \end{aligned} \tag{23}$$

where we used $\Sigma_{i=1}^{n} 2P_i = \Sigma_{i=1}^{n}(P_{1i}+P_{2i}) = 1$. Thus the single-pixel entropy change is equal to one bit minus a small positive measure proportional to local variability of the probability distribution. The k summation in Eq. (23) can be well approximated by the leading term alone. Given continuous probability distribution $P(x)$ of pixel intensity x and intensity binning $\delta x \sim 1/n$, the discrete distribution $P_i \sim P(x_i)\delta x \sim 1/n$. Therefore $\delta P_i = [P'(x_i)\delta x]\delta x \sim (\delta x)^2$ if we assume a smooth (nonfractal) $P(x)$ such that $P(x)$ is differentiable [i.e., $P'(x)$ does not depend on n]. Consequently the leading term in the variability measure in Eq. (23) is proportional to $1/n^2$, or equivalently to $\exp(-2b)$. Summing Eq. (23) over b, one can see that up to a constant, the single-pixel entropy $H^1(b)$ is proportional to pixel depth b minus a positive value that decreases exponentially as b increases. This is indeed the functional form of H^1 shown in Fig. 4(a).

To determine how mutual information terms depend on pixel depth, it is necessary to repeat the above argument for the two-pixel entropy $H^{1,2}$. In this case the probability distribution matrix P is two-dimensional. Therefore by changing the pixel depth from $b+1$ to b, four neighboring matrix elements are added as

$$\begin{aligned} P_n(i/2, j/2) = {} & P_{2n}(i,j) + P_{2n}(i,j+1) \\ & + P_{2n}(i+1,j) + P_{2n}(i+1,j+1). \end{aligned} \tag{24}$$

Denoting

$$P_1 = P + \delta P_1 = P_{2n}(i,j),$$
$$P_2 = P + \delta P_2 = P_{2n}(i,j+1),$$
$$P_3 = P + \delta P_3 = P_{2n}(i+1,j),$$
$$P_4 = P + \delta P_4 = P_{2n}(i+1,j+1),$$

where $P=(P_1+P_2+P_3+P_4)/4$ and $\delta P_1+\delta P_2+\delta P_3+\delta P_4=0$, and repeating the same steps as for H^1 for the leading term in the logarithm expansion, one obtains

$$\begin{aligned} H_{ij}^{1,2}(2n) - H_{ij}^{1,2}(n) &= (P_1+P_2+P_3+P_4)\log_2(P_1+P_2+P_3+P_4) \\ &\quad - P_1\log_2 P_1 - P_2\log_2 P_2 - P_3\log_2 P_3 - P_4\log_2 P_4 \\ &\simeq 8P - \frac{(\delta P_1)^2+(\delta P_2)^2+(\delta P_3)^2+(\delta P_4)^2}{2P\log 2}. \end{aligned}$$

Then for the total change of the two-pixel entropy,

$$\begin{aligned} H^{1,2}(2n) - H^{1,2}(n) &= \sum_{i,j=1}^{n} 8P_{ij} - \frac{1}{2\log 2}\sum_{i,j=1}^{n}\frac{1}{P_{ij}}\sum_{a=1}^{4}(\delta P_{ija})^2 \\ &= 2 - \frac{1}{2\log 2}\sum_{i,j=1}^{n}\frac{1}{P_{ij}}\sum_{a=1}^{4}(\delta P_{ija})^2. \end{aligned} \tag{25}$$

As was true for the single-pixel entropy, the variability sum is proportional to $1/n^2$, since $P_{ij}\sim 1/n^2$ and $\delta P_{ija}\sim 1/n^3$. Finally, using Eqs. (23) and (25) one obtains for mutual information I_2

$$I_2(2n) - I_2(n) = \frac{1}{2\log 2}\left[\sum_{i,j=1}^{n}\frac{1}{P_{ij}}\sum_{a=1}^{4}(\delta P_{ija})^2 - 4\sum_{i=1}^{n}\frac{(\delta P_i)^2}{P_i}\right]. \tag{26}$$

Since each sum in Eq. (26) is proportional to $1/n^2$, or equivalently to $\exp(-2b)$, the two-pixel mutual information $I_2(b)$ exponentially converges to a constant as pixel depth b increases [see Fig. 4(b)]. The saturation constant is positive because the P_{ij} distribution always has a

"spikier" shape (a larger local variation) compared to $\tilde{P}_{ij} = P_i P_j$. The saturation value becomes zero when pixels are totally independent and infinite when all pixels are perfectly correlated (i.e., for an image of uniform intensity). By continuing the argument above for the three-pixel information term I_3 one arrives at the same result as for I_2. Indeed apart from the sampling noise effect for large b values, the I_3 functional form shown in Fig. 4(c) is equivalent to that of I_2.

It is worth noting that electronic noise in a camera's CCD that can arise in higher pixel bits would also contribute to the mutual information saturation. However according to van Hateren and van der Schaaf,[23] noise in the ensemble's images is negligible. Besides, block averaging carried out over four and nine nearby pixels (in order to get images of scales 2× and 3×, respectively) decreases the standard deviation of the electronic noise two- and threefold. If it had been the electronic noise defining the saturation threshold, block averaging would have increased the threshold by 1 and 1.5 bits correspondingly. This is clearly not the case, as one can see in Figs. 4(b) and (c).

It is interesting that saturation occurs for $b \sim 5$—6 bits. This corresponds to approximately 50 shades of gray, which appears to be the threshold for human gray-level discrimination.[33] Hardly a coincidence, it follows the general strategy of computing-cost minimization adopted by the visual system. It is through I_2, I_3, and higher-order dependencies that the underlying structure of the world is encoded in visual images; therefore it makes sense for the visual system to process incoming data at the quantization level at which information saturation is reached. Considering that there has been no study of perceptual threshold for gray-level quantization in natural scenes to compare with our data, it would be an interesting psychophysical experiment to perform.

It is customary to log-transform raw image intensities before collecting ensemble statistics. Images used in the present work were not log-transformed because we were interested in statistics of natural images as projected on the retina, rather than in the postretinal representation. This is because the retinal output signal is not only log-transformed, but filtered by the receptive fields of the ganglion cells, which strongly changes statistics of the signal. However our results involving mutual information are the same for log-transformed images. This follows from the property of mutual information to be invariant under arbitrary point transformations. Consider n pixels with the given intensity $\mathbf{x}$: $\mathbf{x} = \{x_1, \ldots, x_n\}$. Let

$$\mathbf{y} = \{y_1, \ldots, y_n\} = \{f_1(x_1), f_2(x_2), \ldots, f_n(x_n)\} \quad (27)$$

be a one-to-one transformation of $\mathbf{x}$ in which each f_i depends on x_i only (point transformation) and has a continuous first derivative. Because probability density volume is conserved for an arbitrary transformation, $P(\mathbf{y}) = P(\mathbf{x})/|\mathbf{J}(\mathbf{x})|$, where $\mathbf{J}(\mathbf{x}) = \det[\partial f_i/\partial x_j]$ is the Jacobian of the transformation. Consequently, for n-pixel entropy,

$$H^n(\mathbf{y}) = H^n(\mathbf{x}) + \mathbf{E}\log_2|\mathbf{J}(\mathbf{x})|, \quad (28)$$

where the logarithm of the Jacobian is averaged over intensity values of n pixels.[1] For a point transformation (27) the Jacobian matrix is diagonal with the ith diagonal element given by $\mathrm{d}f_i/\mathrm{d}x_i$. The entropy [Eq. (28)] therefore becomes

$$H^n(\mathbf{y}) = H^n(\mathbf{x}) + \mathbf{E}\log_2\prod_{i=1}^{n}\frac{\mathrm{d}f_i}{\mathrm{d}x_i} = H^n(\mathbf{x}) + \sum_{i=1}^{n}\mathbf{E}\log_2\frac{\mathrm{d}f_i}{\mathrm{d}x_i}, \quad (29)$$

which shows that each pixel's entropy change is represented by an independent term. Because mutual information is defined as the difference between n-pixel entropy and n single-pixel entropies, the summation in Eq. (29) cancels with n single-pixel terms. The same is true for ordered contributions to mutual information. For I_3, for example,

$$\begin{aligned} I_3(y_1, y_2, y_3) &= H(y_1, y_2, y_3) - H(y_1, y_2) - H(y_2, y_3) + H(y_2) \\ &= I_3(x_1, x_2, x_3) + \sum_{i=1}^{3}\mathbf{E}\log_2\frac{\mathrm{d}f_i}{\mathrm{d}x_i} - \sum_{i=1}^{2}\mathbf{E}\log_2\frac{\mathrm{d}f_i}{\mathrm{d}x_i} \\ &\quad - \sum_{i=2}^{3}\mathbf{E}\log_2\frac{\mathrm{d}f_i}{\mathrm{d}x_i} + \mathbf{E}\log_2\frac{\mathrm{d}f_2}{\mathrm{d}x_2} \\ &= I_3(x_1, x_2, x_3). \end{aligned}$$

Because the probability distributions in our study were sampled for discrete rather than continuous pixel intensities, Eq. (28) is valid only as a linear approximation.

To make sure that the effect of intensity quantization is small, we calculated I_2 and I_3 for log-transformed images from the ensemble. The results show that for $b \leqslant 4$ mutual information somewhat increased, but for larger pixel depths the effect of the transformation was negligible.

The I_2/H_1 and I_3/H_1 ratios shown in the insets to Fig. 4 indicate that the information redundancy due to interpixel correlations quickly decreases with increasing pixel depth ($\sim 1/b$). Assuming that the human gray-level discrimination threshold is somewhere around 6 bits per pixel, the two-pixel contribution to the interpixel redundancy is about 50%. This is higher than the intrapixel redundancy defined by $(b - H_1)/b$ [shown in the inset to Fig. 4(a)], which amounts to about 30% for $b = 6$ bits. The three-pixel redundancy, on the other hand, inputs only about 4%, which is an order of magnitude smaller than either the intrapixel redundancy or the two-pixel redundancy. Even if I_n terms decreased by a factor of three for each consecutive n (a rather conservative assumption), the contribution from the $n > 2$ terms combined would still not exceed 6%.

It is important to point out that I_n as defined by relations (6) is not equivalent to mutual information contribution from nth order correlations in the sense they are usually understood. Consider three pixels with intensity $\mathbf{x} = \{x_1, x_2, x_3\}$ given by the Gaussian probability distribution

$$G_n(\mathbf{x}) \equiv \frac{\exp(-\frac{1}{2}\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x})}{(2\pi)^{n/2}[\det(\mathbf{R})]^{1/2}}, \quad (30)$$

where $\mathbf{R} = \langle \mathbf{x}\mathbf{x}^T \rangle$ is the covariance matrix. It is easy to show that even though third- and higher-order correlations are zero for the Gaussian distribution, $I_3 \neq 0$ unless $\mathbf{R}^{-1}(1, 3) = 0$, or equivalently, $\langle x_1 x_3 \rangle = \langle x_1 x_2 \rangle \times \langle x_2 x_3 \rangle / \langle x_2^2 \rangle$. This is because I_3 includes second-order mutual information from the direct correlation between pixels 1 and 3. This direct correlation excludes an indirect contribution by way of pixel 2, and can be defined as $\langle x_1 \leftrightarrow x_3 \rangle \equiv \langle x_1 x_3 \rangle - \langle x_1 x_2 \rangle \langle x_2 x_3 \rangle / \langle x_2^2 \rangle$. We call it nonlocal in the sense that the correlation is between two pixels that are not nearest neighbors. It is easy to check that given three pixels described by Eq. (30), the nonlocal contribution to I_3 is positive: $\det(\mathbf{R})$ is maximized for $\langle x_1 \leftrightarrow x_3 \rangle = 0$, while Eq. (31) (defined below) shows that this maximizes the Gaussian entropy $H(G_3)$ and therefore minimizes I_3 [given by $I_3 = 2H(G_2) - H(G_1) - H(G_3)$]. Thus I_3 sums two contributions to mutual information. One is due to local (nearest neighbor) higher-than-second-order correlations between three adjacent pixels (1, 2, and 3), such as $\langle x_1 x_2 x_3 \rangle$, and the other is due to nonlocal second-and-higher-order correlation between two next-nearest pixels (1 and 3). Note, though, that to describe the two-pixel, nonlocal correlations, three-pixel probability factors are required.

To express the information redundancy in terms of correlation orders rather than number of pixels n, we note further that I_2 includes contributions from two-pixel correlations of all orders. It can be written as $I_2 = I_2^{(2)} + \tilde{I}_2$, where $I_2^{(2)}$ accounts for the second-order $\langle x_1 x_2 \rangle$ intensity correlations while the higher-order contributions given by $\tilde{I}_2$ correspond to correlations such as $\langle x_1^2 x_2 \rangle$, $\langle x_1^2 x_2^2 \rangle$, etc. To determine to what extent minimization of higher-order correlations will reduce the interpixel redundancy it is necessary to calculate $I_2^{(2)}$ and compare it with the rest of the mutual information terms.

Since third- and higher-order correlations are zero for the Gaussian distribution, $I_2^{(2)}$ can be estimated by approximating the actual two-pixel probability distributions with Gaussian distributions of the same mean and covariance. It is straightforward to show that entropy of the n-dimensional Gaussian probability distribution of relation (30) is given by[1]

$$H(G_n) = \log_2\{(2\pi e)^{n/2}[\det(\mathbf{R})]^{1/2}\} \tag{31}$$

Then if intensities of two pixels $\{x_1, x_2\}$ are distributed according to relation (30), the mutual information is

$$I(G_2) = 2H(G_1) - H(G_2) = -\log_2\left(1 - \frac{\langle x_1 x_2 \rangle^2}{\langle x_1^2 \rangle^2}\right)^{1/2}. \tag{32}$$

This gives $I_2^{(2)}$ if $\langle x_1 x_2 \rangle$ and $\langle x_1^2 \rangle$ are calculated using the actual probability distributions. For $b = 6$ the $I_2^{(2)}$ calculated using Eq. (32) was 2.02 bits, which is 0.22 bits smaller than the total two-pixel information I_2. Therefore the overall contribution of higher-order correlations estimated as the sum of I_3 and $\tilde{I}_2$ terms is on the order of 0.4 bits, which is five times smaller than the second-order contribution $I_2^{(2)}$. As discussed earlier, along with higher-order correlations, I_3 accounts for nonlocal second-order correlations as well. This makes the higher-order contribution even smaller.

Consequently, removing higher-order correlations would only insignificantly increase coding efficiency. Hence it seems unlikely that the higher-order redundancy minimization is the main constraint shaping the cortical receptive fields, as previously suggested. This is not to say that the higher-order redundancy is not important for vision beyond coding efficiency. After all, it is through correlations between pixels that we extract information from images, and this is manifested by the correspondence of our luminance discrimination threshold with pixel depth where the mutual information (I_2 and I_3) between pixels saturates. Rather, self-similarity of neuronal responses under translation and scaling transformations,[16] or some other kind of perceptually (rather than statistically) important constraints, might be at work instead.

To summarize, we have presented a statistical analysis of a large ensemble of natural images that is based on local factorization of the ensemble's gray-scale joint probability distribution. It is shown that the mutual information resulting from the local two-pixel and three-pixel luminance correlations increases with the pixel depth, saturating for pixel depths larger than 6 bits. We suggest that this accounts for the luminance quantization threshold in humans. The total contribution from the three-pixel correlations to the ensemble's information redundancy was on the order of 4%. It is a small fraction of both the two-pixel redundancy (50%) and the intrapixel encoding redundancy (30%). This suggests that the reduction of higher-than-second-order redundancies is not the main cause of receptive field properties of neurons in V1.

ACKNOWLEDGMENTS

This research was supported by a grant from the Gatsby Foundation. We thank Hans van Hateren for his help with the image database, and anonymous reviewers for useful comments on earlier versions of this paper.

Y. Petrov may be reached by e-mail at yp@physiol.ox.ac.uk.

REFERENCES

1. C. E. Shannon, "A mathematical theory of communication," Bell Syst. Tech. J. **27**, 379–423 (1948).
2. E. R. Kretzmer, "Statistics of television signals," Bell Syst. Tech. J. **31**, 751–763 (1952).
3. W. F. Schreiber, "The measurement of third-order probability distributions of television signals," IRE Trans. Inf. Theory **IT-2**, 94–105 (1956).
4. J. R. Parks, "Prediction and entropy of half-tone pictures," Behav. Sci. **10**, 436–445 (1965).
5. N. S. Tzannes, R. V. Spencer, and A. J. Kaplan, "On estimating the entropy of random fields," Inf. Control. **16**, 1–6 (1970).
6. F. Attneave, "Some information aspects of visual perception," Psychol. Rev. **61**, 183–193 (1954).
7. H. B. Barlow, "Possible principles underlying the transformation of sensory messages," in *Sensory Communication*, W. A. Rosenblith, ed. (MIT, Cambridge, Mass., 1961).
8. J. J. Atick, "Could information theory provide an ecological theory of sensory processing?" Network **3**, 213–251 (1992).

9. H. B. Barlow and P. Foldiak, *The Computing Neuron* (Addison-Wesley, Reading, Mass., 1989).
10. S. Watanabe, "Pattern recognition as a quest for minimum entropy," Pattern Recognition **13**, 381–387 (1981).
11. S. Watanabe, *Pattern Recognition: Human and Mechanical* (Wiley, New York, 1985).
12. S. B. Laughlin, "A simple coding procedure enhances a neuron's information capacity," Z. Naturforsch. **36**, 910–912 (1981).
13. J. J. Atick and A. N. Redlich, "Towards a theory of early visual processing," Neural Comput. **2**, 308–320 (1990).
14. J. J. Atick, Z. Li, and A. N. Redlich, "Understanding retinal color coding from first principles," Neural Comput. **4**, 559–572 (1992).
15. J. J. Atick and A. N. Redlich, "What does the retina know about the natural scenes?" Neural Comput. **4**, 196–210 (1992).
16. Zh. Li and J. J. Atick, "Towards a theory of the striate cortex," Neural Comput. **6**, 127–146 (1994).
17. A. J. Bell and T. J. Sejnowski, "The 'independent components' of natural scenes are edge filters," Vision Res. **37**, 3327–3338 (1997).
18. 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).
19. J. G. Daugman, "Entropy reduction and decorrelation in visual coding by oriented neural receptive fields," IEEE Trans. Biomed. Eng. **36**, 107–114 (1989).
20. D. J. Field, "What is the goal of sensory coding?" Neural Comput. **6**, 559–601 (1994).
21. C. Jutten and J. Herault, "Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture," Signal Process. **24**, 1–10 (1991).
22. A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," Neural Networks **13**, 411–430 (2000).
23. 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).
24. A. Hyvärinen and P. O. Hoyer, "Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces," Neural Comput. **12**, 1705–1720 (2000).
25. B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," Nature **381**, 607–609 (1996).
26. B. A Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: a strategy employed by V1?" Vision Res. **37**, 3311–3325 (1997).
27. M. G. A. Thomson, "Higher-order structure in natural scenes," J. Opt. Soc. Am. A **16**, 1549–1553 (1999).
28. M. G. A. Thomson, "Beats, kurtosis and visual coding," Network Comput. Neural Syst. **12**, 271–287 (2001).
29. A. B. Lee, K. S. Pedersen, and D. Mumford, "The complex statistics of high-contrast patches in natural images," presented at the IEEE Workshop on Statistical and Computational Theories of Vision, Vancouver, Canada, 2001; http://www.cis.ohio-state.edu/szhu/SCTV2001.html
30. N. G. Deriugin, "The power spectrum and the correlation function of the television signal," Telecommunications **1**, 1–12 (1957).
31. D. J. Tolhurst, Y. Tadmor, and T. Chao, "Amplitude spectra of natural images," Ophthalmic Physiol. Opt. **12**, 229–232 (1992).
32. D. L. Ruderman, "The statistics of natural images," Network **5**, 517–548 (1994).
33. D. Kersten, "Predictability and redundancy in natural images," J. Opt. Soc. Am. A **4**, 2395–2400 (1987).