This paper was published in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 78-91, 2013, doi 10.1109/TPAMI.2012.78, and is available at http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6175904

I. Marín-Franch and D. H. Foster, "Estimating Information from Image Colors: An Application to Digital Cameras and Natural Scenes," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 78-91, 2013.
DOI 10.1109/TPAMI.2012.78

Estimating information from image colors: an application to digital cameras and natural scenes

Iván Marín-Franch and David H. Foster

***Abstract*—The colors present in an image of a scene provide information about its constituent elements. But the amount of information depends on the imaging conditions and on how information is calculated. This work had two aims. The first was to derive explicitly estimators of the information available and the information retrieved from the color values at each point in images of a scene under different illuminations. The second was to apply these estimators to simulations of images obtained with five sets of sensors used in digital cameras and with the cone photoreceptors of the human eye. Estimates were obtained for 50 hyperspectral images of natural scenes under daylight illuminants with correlated color temperatures 4000 K, 6500 K, and 25000 K. Depending on the sensor set, the mean estimated information available across images with the largest illumination difference varied from 15.5 to 18.0 bits and the mean estimated information retrieved after optimal linear processing varied from 13.2 to 15.5 bits (each about 85% of the corresponding information available). With the best sensor set, 390% more points could be identified per scene than with the worst. Capturing scene information from image colors depends crucially on the choice of camera sensors.**

***Index Terms*—Color vision, color information, digital color cameras, color processing, information theory, natural scenes, kth-nearest-neighbor statistics, color constancy.**

I. Introduction

Color provides information about the reflecting properties of surfaces, thereby allowing regions of a scene to be demarcated and the elements of regions to be distinguished. Yet how much information about the content of a scene is captured by the colors of the reflected light? More specifically, if images of a scene under a particular illumination are obtained with a digital trichromatic camera, to what extent can the elements of the scene be identified by their colors, independent of spatial position?

A priori, it seems unlikely that all the elements in a scene can be characterized in this way. One problem is that the color values at each point in an image depend on the spectrum of the illumination on the scene, so that when the illumination changes, so generally do the color values. This confounding effect of illumination can be largely discounted by correcting color values by so-called von Kries scaling [1], [2], although not completely. Another problem is that different reflectance spectra at different points in the scene under the same illumination can produce the same color values. This is the phenomenon of metamerism [3] and is a consequence of the number of degrees of freedom in natural reflectance spectra being greater than the number of degrees of freedom in color values, namely three with a trichromatic camera.

Nevertheless, there remains a strong dependency between the color values of different images of the same scene under different illuminations. This dependency can be quantified with Shannon's mutual information [4]–[6]. The advantages of this measure over other measures of dependency, such as linear correlation, are well known [7]–[10].

Two kinds of mutual information associated with the images of a scene were used here, namely the information available from the color values at each point in the scene imaged under different illuminations and the information retrieved in the basic task of matching points across those images by their color values. The information available is, by construction [4]–[6], founded on a theoretical camera with an infinite number of pixels. It sets, therefore, an upper (finite) bound on the information actually available from any camera with a finite number of pixels, which can yield only a finite sample of color values. The information available necessarily depends on factors such as the spectral reflectances of the surfaces in the scene and their relative abundances, the spectral radiances of the illuminations on the scene, and the spectral sensitivities of the camera sensors. The information retrieved depends not only on these factors, but also on how the sensor signals are processed and then matched, for example, by von Kries scaling and by nearest-neighbor matching. The information available is also an upper bound on the information retrieved.

In previous analyses, estimates of the information available and information retrieved have been used to reveal both the efficiency and the limits of color processing by the human eye in viewing natural scenes under different illuminations. In one application [11], it was shown that coding at the receptors was highly redundant, as expected given the overlap of the spectral sensitivities of the medium- and long-wavelength-sensitive cone pigments [12]. But with optimal linear postreceptoral processing, redundancy was reduced and efficiency increased so that the information retrieved from color images of natural scenes under different daylight illuminants reached almost 90% of that achievable by an ideal observer. The coefficients of the linear transformations describing this postreceptoral processing are similar to those estimated independently by behavioral methods [13]. Such calculations illustrate the importance of estimating both kinds of mutual information associated with the images of a scene.

One of the two aims of this work was to put on a firmer basis the derivation and verification of the estimators of the information available and the information retrieved to answer the question posed earlier, namely how much information about a scene's content is captured by its colors. The other

School of Electrical and Electronic Engineering, University of Manchester, Sackville Street, Manchester M60 1QD, UK. Iván Marín-Franch is currently with the Indiana University School of Optometry, Bloomington 47405, US, and the Department of Optometry and Visual Science, City University London, London EC1V 0HB, UK. E-mails: imarinfr@indiana.edu; d.h.foster@manchester.ac.uk.

aim of the work was to illustrate an application of these estimators to five sets of sensors used in commercial digital trichromatic cameras. To provide a reference, the estimators were also applied to the cone photoreceptors of the human eye. Conveniently, mutual information can be interpreted as the logarithm of the mean number of distinct elements or points that can be identified without error across images of the scene under different illuminations [11], [14]. The resulting information estimates therefore provided an answer to the more specific question of the extent to which the elements of a scene can be identified by their colors, independent of spatial position.

The organization of this work was as follows. Images of natural scenes were generated from a set of 50 hyperspectral images of rural and urban scenes under each of three daylight illuminants with correlated color temperatures (CCTs) of 4000 K, 6500 K, and 25000 K. Information available and information retrieved were estimated for images of a scene under pairs of these illuminants. As a practical matter, naïve estimates of mutual information based on histograms, here of color values at each point, are known to be susceptible to bias, and so methods were employed that were asymptotically unbiased and reasonably efficient for both kinds of information estimate. As already implied, by the nature of these calculations, information estimates did not incorporate any spatial data. Thus, images were not segmented into uniform regions in any way, except for the trivial limit defined by pixel resolution, and any estimate of the number of identifiable points was assumed to be an upper bound on the number of identifiable regions. No allowance was made for noise in the sensors or in post-sensor processing. In this way, the effects of spectral tuning of the sensors could be most clearly demonstrated.

It was found that for the largest difference in daylight illuminants—with CCTs of 25000 K and 4000 K—the mean estimated information available varied from 15.5 bits to 18.0 bits, depending on the set of sensors. These values are equivalent to 4.7×10^4 and 2.7×10^5 distinct identifiable points per scene, an increase of 470% from the worst to the best sensor set. The corresponding estimated information retrieved was lower, at 13.2 and 15.5 bits, equivalent to 9.5×10^3 and 4.7×10^4 distinct identifiable points per scene, an increase of 390% from the worst to the best sensor set.

For the eye, the mean estimated information available and information retrieved for the same illuminants were 17.1 and 14.7 bits, respectively, equivalent to 1.37×10^5 and 2.7×10^4 distinct identifiable points per scene, similar to the highest values recorded with some camera sensors.

To help set these estimates in context, the pointillistic painting by Georges Seurat "A Sunday Afternoon on the Island of La Grande Jatte" (1884-1886) would require 16.6 bits to code the more than 10^5 "points" on the canvas.

Some implications of present findings, and their limitations, are considered in the Discussion.

II. Sensor signals and image entropy

Consider a scene illuminated by a spatially uniform global illuminant with incident spectral radiance $e(\lambda)$ at wavelength λ. Suppose that at a point (x, y) in the scene the effective spectral reflectance [15] is $\rho(\lambda; x, y)$ so that the reflected spectral radiance is given by $c(\lambda; x, y) = e(\lambda)\rho(\lambda; x, y)$.[1] Suppose that this reflected spectrum is sampled by the long-, medium-, and short-wavelength-sensitive (conventionally, R, G, and B) sensors of a digital camera (or cone photoreceptors of the eye) with spectral sensitivities $s_{\mathrm{R}}(\lambda)$, $s_{\mathrm{G}}(\lambda)$, and $s_{\mathrm{B}}(\lambda)$, respectively. The corresponding triplet of color values (r, g, b) at (x, y) encodes the spectrum $c(\lambda; x, y)$ thus

$$\begin{aligned} r &= \int s_{\mathrm{R}}(\lambda) c(\lambda; x, y)\, \mathrm{d}\lambda \ , \\ g &= \int s_{\mathrm{G}}(\lambda) c(\lambda; x, y)\, \mathrm{d}\lambda \ , \\ b &= \int s_{\mathrm{B}}(\lambda) c(\lambda; x, y)\, \mathrm{d}\lambda \ , \end{aligned} \tag{1}$$

where the integral is evaluated over the visible wavelength range. If the point (x, y) within the scene is chosen randomly, the color values r, g, and b in (1) may be treated as instances of continuous random variables [12], R, G, and B, say. The triplet $a = (r, g, b)$ is an instance of a trivariate continuous random variable $A = (R, G, B)$, whose probability density function (pdf) is f, say. This pdf characterizes the nature of the unpredictability of the color values for a particular scene, illuminant, and set of sensors.

A discretized version of the continuous random variable A can be obtained [6] by partitioning the space in which the (bounded) variables R, G, and B take their values. Suppose that the partition has a finite number of bins, D say, indexed by an integer d with $1 \leq d \leq D$. Suppose that each bin has equal edge lengths $\Delta r = \Delta g = \Delta b$, and let $\Delta a = \Delta r \Delta g \Delta b$. For each d, let a_d be the value of a within the dth bin such that

$$f(a_d)\Delta a = \int_{r_d}^{r_d + \Delta r_d} \int_{g_d}^{g_d + \Delta g_d} \int_{b_d}^{b_d + \Delta b_d} f(r, g, b) \mathrm{d}r \mathrm{d}g \mathrm{d}b \, .$$

Denote by $A^\Delta = (R^\Delta, G^\Delta, B^\Delta)$ the discretized version of A whose probability mass function (pmf) p is given by

$$p(a_d) = P\left\{A^\Delta = a_d\right\} = f(a_d)\Delta a, \ \text{for } d = 1, \ldots, D \, . \tag{2}$$

The entropy $H(A^\Delta)$ of the discrete random variable A^Δ for a particular scene, illuminant, and set of sensors is then defined [4]–[6] by

$$H(A^\Delta) = -\sum_{d=1}^{D} p(a_d) \log p(a_d) \, , \tag{3}$$

where the probabilities $p(a_d)$ are given by (2) and where conventionally $0 \log 0 = 0$. The entropy $H(A^\Delta)$ ranges from zero to $\log D$. If the logarithm is to the base 2, then the entropy is in bits; if it is the natural logarithm, then the entropy is in nats.

[1] With a spatially uniform global illuminant $e(\lambda)$, the effective spectral reflectance $\rho(\lambda; x, y)$ at (x, y) is defined by $c(\lambda; x, y)/e(\lambda)$, given that $e(\lambda) > 0$ for all λ. This representation confounds variations in spectral reflectance with the effects of surface orientation, occlusion, and mutual illumination, but this confound is not critical in this application [15]. The notation in [15] differs slightly from that used in this work.

If all the points in a scene have the same color value, so that the pmf $p(a_d)$ of A^Δ is zero except at one particular value of d, e.g., if the scene is a perfectly homogeneous surface so that all the points have the same effective reflectance spectrum or if the binning is too coarse (i.e., D is too small) to capture the differences in effective spectral reflectance between points, then there is no uncertainty about the color value at any chosen point, and $H(A^\Delta) = 0$. Conversely, if all the points in a scene have different color values, so that the pmf of A^Δ is uniform, i.e., $p(a_d) = 1/D$ for all $d = 1, \ldots, D$, then the uncertainty about the color value of the chosen point is maximum and $H(A^\Delta) = \log D$.

III. Information available and its estimators

In general, two different global illuminants, say $e_1(\lambda)$ and $e_2(\lambda)$, illuminating the same scene, one at a time, give rise to two different continuous random variables, A_1 and A_2, respectively, with pdfs f_1 and f_2 and joint pdf f_{12}, and, likewise, to two different discrete random variables, A_1^Δ and A_2^Δ, with pmfs p_1 and p_2 and joint pmf p_{12}. As noted earlier, the random variables A_1 and A_2 are strongly dependent on each other and so are A_1^Δ and A_2^Δ. The mutual information between A_1^Δ and A_2^Δ can be derived from their entropies, as follows.

For each $d_1 = 1, \ldots, D$, let $p_{2|1}(a_{2d_2}|a_{1d_1})$ be the conditional probability that $A_2^\Delta = a_{2d_2}$ for each $d_2 = 1, \ldots, D$. The entropy $H(A_2^\Delta|A_1^\Delta = a_{1d_1})$ of A_2^Δ, given $A_1^\Delta = a_{1d_1}$, is defined [6] by

$$H(A_2^\Delta|A_1^\Delta = a_{1d_1}) = - \sum_{d_2=1}^{D} p_{2|1}(a_{2d_2}|a_{1d_1}) \log p_{2|1}(a_{2d_2}|a_{1d_1}) \,.$$

The conditional entropy $H(A_2^\Delta|A_1^\Delta)$ is then defined [4], [5] as the value of $H(A_2^\Delta|A_1^\Delta = a_{1d_1})$ averaged over all possible values of A_1^Δ; that is,

$$H(A_2^\Delta|A_1^\Delta) = \sum_{d_1=1}^{D} p_1(a_{1d_1}) H(A_2^\Delta|A_1^\Delta = a_{1d_1}) \,.$$

The conditional entropy measures the uncertainty about random variable A_2^Δ given that the value of A_1^Δ is known. It is always lower than $H(A_2^\Delta)$, unless the two random variables are independent, in which case the two quantities are equal. The difference $H(A_2^\Delta) - H(A_2^\Delta|A_1^\Delta)$ gives [4], [5] the mutual information $I(A_1^\Delta; A_2^\Delta)$ between A_1^Δ and A_2^Δ. As the conditional entropy $H(A_2^\Delta|A_1^\Delta)$ is simply the difference between the joint entropy $H(A_1^\Delta, A_2^\Delta)$ and $H(A_1^\Delta)$, the mutual information may therefore be written as

$$I(A_1^\Delta; A_2^\Delta) = H(A_1^\Delta) + H(A_2^\Delta) - H(A_{12}^\Delta) \,, \tag{4}$$

where A_{12}^Δ stands for (A_1^Δ, A_2^Δ). Explicitly [6],

$$I(A_1^\Delta; A_2^\Delta) = \sum_{d_1=1}^{D} \sum_{d_2=1}^{D} p_{12}(a_{1d_1}, a_{2d_2}) \log \frac{p_{12}(a_{1d_1}, a_{2d_2})}{p_1(a_{1d_1}) p_2(a_{2d_2})} \,. \tag{5}$$

Shannon's channel-coding theorem [4], [5] gives the mutual information an operational interpretation which was alluded to earlier. That is, since the maximum number of points that can be encoded with I bits is 2^I, if $I = I(A_1^\Delta; A_2^\Delta)$, then 2^I is the maximum number of distinct points that can be identified reliably across two images of a scene under the two global illuminants e_1 and e_2 [11].

As bin size $\Delta r = \Delta g = \Delta b$ tends to zero, the entropies $H(A_1^\Delta)$, $H(A_2^\Delta)$, and $H(A_{12}^\Delta)$ each tend to infinity, but not the mutual information $I(A_1^\Delta; A_2^\Delta)$, which tends to the limit [16, Chapter 4], [6, Chapter 9]

$$I(A_1; A_2) = \int f_{12}(a_1, a_2) \log \frac{f_{12}(a_1, a_2)}{f_1(a_1) f_2(a_2)} \, \mathrm{d}a_2 \mathrm{d}a_1 \,, \tag{6}$$

the continuous analog of (5). The value of (6) is invariant under differentiable invertible transformations of the continuous random variables A_1 and A_2, and decreases otherwise [16, Chapter 4]. The quantity $I = I(A_1; A_2)$ is the least upper bound on the mutual information $I(A_1^\Delta; A_2^\Delta)$ defined over all possible discretizations of the continuous random variables A_1 and A_2 [16, Chapter 4]. This least upper bound I is the information that is available across two images of a scene each under a different illuminant.

Let

$$N_I = 2^I \,. \tag{7}$$

Then N_I is the least upper bound on the number of distinct points that can be identified reliably across two images of the scene. Notice that if $A_1 = A_2$, the mutual information $I(A_1; A_2)$ is infinite.

In practice, estimating the information available, i.e., $I(A_1; A_2)$, is not straightforward. Several methods, including some used in this analysis, make use of the fact that mutual information can be expressed as a combination of differential entropies [4]–[6]. The differential entropy $h(A_1)$ of A_1 is defined [4], [5] by

$$h(A_1) = - \int f_1(a_1) \log f_1(a_1) \, \mathrm{d}a_1 \,, \tag{8}$$

which, unlike the limit of discrete entropy (3) as bin size tends to zero, need not be infinite, although it does depend on the units in which the values of A_1 are expressed. The differential entropy $h(A_2)$ of A_2 and the joint differential entropy $h(A_{12})$ of $A_{12} = (A_1, A_2)$ are defined analogously. The information available is [6]

$$I(A_1; A_2) = h(A_1) + h(A_2) - h(A_{12}) \,, \tag{9}$$

mirroring (4).

One method of estimating $I(A_1; A_2)$ is to estimate the pdfs f_1 and f_2 and the joint pdf f_{12}, and use them to estimate $h(A_1)$, $h(A_2)$, and $h(A_{12})$ from the corresponding definitions, e.g., (8). The estimates of f_1, f_2, and f_{12} are necessarily based on finite samples. Accordingly, suppose that N points $\{(x_i, y_i) \mid i = 1, \ldots, N\}$ are sampled uniformly from a scene and their color values are calculated from (1) for illuminants e_1 and e_2, yielding the sets

$$\begin{aligned} &\{a_{1i} = (r_{1i}, g_{1i}, b_{1i}) \mid i = 1, \ldots, N\} \,, \\ &\{a_{2i} = (r_{2i}, g_{2i}, b_{2i}) \mid i = 1, \ldots, N\} \,, \end{aligned} \tag{10}$$

and

$$\{a_{12i} = (a_{1i}, a_{2i}) \mid i = 1, \ldots, N\} \ . \tag{11}$$

The difficulty is arriving at the estimates of f_1, f_2, and f_{12} from these samples. If the sample size N is large enough, reliable estimates of f_1, f_2, and f_{12} may be obtained by simple histogram-based methods such as binning and adaptive partitioning [17]. This is equivalent to D-bin quantization with pmfs as in (2), although if partitioning is adaptive, variable values of Δr, Δg, and Δb are needed. Even so, D must be very large to obtain accurate estimates; otherwise the pdf is not well approximated by the empirical pmf, leading to bias [18]. In addition, the sample size N has to be much larger than the number of bins with non-zero probability [19]. Some of these disadvantages can be overcome with the use of kernel density estimators [10], [18], but systematic errors in the differential entropy estimates remain [20], [21].

By contrast, methods based on kth-nearest-neighbor statistics [20], [21] avoid estimating pdfs and instead involve calculating distances in the neighborhood of each point in a sample drawn from the spaces spanned by A_1, A_2, and A_{12}, such as the sets in (10) and (11). The advantages of kth-nearest-neighbor estimators have been documented elsewhere [21]–[23].

For completeness, four estimators were used to estimate the information available: (A) a kernel density estimator [10]; (B) a generalized version of a nearest-neighbor estimator due to Kozachenko and Leonenko [20], [24]; (C) a nearest-neighbor estimator due to Kraskov, Stögbauer, and Grassberger [21]; and (D) an experimental offset modification [14] used to improve both estimators (A) and (B).

A. Kernel density estimator

A kernel density estimator provides estimates $\hat{f}_1$, $\hat{f}_2$, and $\hat{f}_{12}$ of the corresponding pdfs f_1, f_2, and f_{12} by smoothing each finite sample of N color values (10) and (11) with a kernel function K_σ, which, in one dimension, is often chosen as a Gaussian density, $K_\sigma(u) = (2\pi\sigma^2)^{-1/2} \exp(-u^2/2\sigma^2)$ [18], where σ is the bandwidth of the smoother. Thus, for any point $a_1 = (r_1, g_1, b_1)$ in the space spanned by A_1, the kernel density estimate $\hat{f}_1(a_1)$ of f_1 at a_1 is defined by

$$\hat{f}_1(a_1) = \frac{1}{N} \sum_{i=1}^{N} K_{\sigma_R}(r_1 - r_{1i}) K_{\sigma_G}(g_1 - g_{1i}) K_{\sigma_B}(b_1 - b_{1i}) \, . \tag{12}$$

If required, the product of the three univariate Gaussian densities can be replaced by a single multivariate Gaussian density. The estimates $\hat{f}_2(a_2)$ and $\hat{f}_{12}(a_{12})$ are defined analogously. An inappropriate choice of bandwidth σ can, however, give misleading results, and two automatic methods are described in Appendix A.

The kernel-density-based estimator $\hat{h}_{\mathrm{KD}}(A_1)$ of the differential entropy $h(A_1)$ is obtained [10] by plugging the estimator $\hat{f}_1(a_1)$ into (8); that is,

$$\hat{h}_{\mathrm{KD}}(A_1) = -\int \hat{f}_1(a_1) \log \hat{f}_1(a_1) \, \mathrm{d}a_1 \, ;$$

and analogously for $\hat{h}_{\mathrm{KD}}(A_2)$ and $\hat{h}_{\mathrm{KD}}(A_{12})$. The kernel-density-based estimator $\hat{I}_{\mathrm{KD}}(A_1; A_2)$ of the information available $I(A_1; A_2)$ follows from (9); that is,

$$\hat{I}_{\mathrm{KD}}(A_1; A_2) = \hat{h}_{\mathrm{KD}}(A_1) + \hat{h}_{\mathrm{KD}}(A_2) - \hat{h}_{\mathrm{KD}}(A_{12}) \, .$$

B. Generalized Kozachenko-Leonenko estimator

The nearest-neighbor estimator of differential entropy proposed by Kozachenko and Leonenko [20] was generalized by Goria et al. [24] to estimators based on kth-nearest neighbors. For a fixed integer k with $0 < k < N$, let μ_{1i} be the Euclidean distance between a_{1i} and its kth-nearest neighbor for the illuminant e_1. If ψ denotes the digamma function and $v = \pi^{m/2}/\Gamma(m/2 + 1)$ is the volume of an m-dimensional open ball of unit radius, then the generalized Kozachenko-Leonenko estimator $\hat{h}_{\mathrm{KL}}(A_1)$ of the differential entropy $h(A_1)$ is defined, in nats, by

$$\hat{h}_{\mathrm{KL}}(A_1) = \frac{m}{N} \sum_{i=1}^{N} \ln \mu_{1i} + \ln(N - 1) - \psi(k) + \ln v \, ,$$

where $m = 3$, the dimension of A_1. The estimators $\hat{h}_{\mathrm{KL}}(A_2)$ and $\hat{h}_{\mathrm{KL}}(A_{12})$ are defined analogously with $m = 3$ and 6, respectively. The Kozachenko-Leonenko estimator $\hat{I}_{\mathrm{KL}}(A_1; A_2)$ of the information available $I(A_1; A_2)$ then follows, as before, from (9); that is,

$$\hat{I}_{\mathrm{KL}}(A_1; A_2) = \hat{h}_{\mathrm{KL}}(A_1) + \hat{h}_{\mathrm{KL}}(A_2) - \hat{h}_{\mathrm{KL}}(A_{12}) \, .$$

C. Kraskov-Stögbauer-Grassberger estimator

Two nearest-neighbor estimators were described by Kraskov et al. [21], but they yielded similar results, and only the one giving the smaller systematic errors was used, namely that denoted by $I^{(2)}$ in [21]. For a fixed integer k with $0 < k < N$, let l_{1i} and l_{2i} be the edge lengths of the smallest rectangle around (a_{1i}, a_{2i}) containing k neighbors. For some norm $||\cdot||$, let n_{1i} and n_{2i} be the numbers of a_{1j} and a_{2j}, $i \neq j$, in the paired sample such that $||a_{1i} - a_{1j}|| \leq l_{1i}/2$ and $||a_{2i} - a_{2j}|| \leq l_{2i}/2$. Denote by $\bar{\psi}$ the average of the values $\psi(n_{1i}) + \psi(n_{2i})$ of the digamma function over all i. The Kraskov-Stögbauer-Grassberger estimator $\hat{I}_{\mathrm{KSG}}(A_1; A_2)$ of the information available $I(A_1; A_2)$ is then defined, in nats, by

$$\hat{I}_{\mathrm{KSG}}(A_1; A_2) = \psi(k) - 1/k - \bar{\psi} + \psi(N) \, .$$

D. Offset estimators

The foregoing estimators were found to converge slowly with Gaussian images (Appendix B). To improve the convergence of the kernel density estimator and Kozachenko-Leonenko estimator, each was decomposed into two components: one, the mutual information between equivalent Gaussian variables with known variance-covariance structure; the other, an offset obtained by applying the estimator to normalized versions of A_1, A_2, and A_{12}. This decomposition was not possible with the Kraskov-Stögbauer-Grassberger estimator $\hat{I}_{\mathrm{KSG}}(A_1; A_2)$, which estimates mutual information directly.

In more detail, if A is a random variable and t an invertible linear transformation, then in general

$$h(A) = h(tA) - \log|t|\,, \tag{13}$$

where $|t|$ is the absolute value of the determinant of the matrix representing t. Set $t_1 = (\mathrm{Var}\,A_1)^{-1/2}$, $t_2 = (\mathrm{Var}\,A_2)^{-1/2}$, and $t_{12} = \mathrm{Var}\,(A_{12})^{-1/2}$. Let $I_{\mathrm{EG}}(A_1; A_2)$ be the mutual information [6] of the equivalent Gaussian variables, i.e. having the same variance-covariance structure as A_1 and A_2, so that

$$I_{\mathrm{EG}}(A_1; A_2) = \frac{1}{2}\log\left(\frac{|\mathrm{Var}\,A_1||\mathrm{Var}\,A_2|}{|\mathrm{Var}\,A_{12}|}\right). \tag{14}$$

Accordingly, from (13) and (9), the mutual information between A_1 and A_2 can be written

$$I(A_1; A_2) = I_{\mathrm{EG}}(A_1; A_2) + h(t_1A_1) + h(t_2A_2) - h\,(t_{12}A_{12})\,. \tag{15}$$

An estimator $\hat{I}_{\mathrm{EG}}$ of I_{EG} is obtained by replacing the variances in (14) by the sample variances. The offset versions of the kernel-density estimator and Kozachenko-Leonenko estimator are then obtained by applying each to the three differential entropies $h(t_1A_1)$, $h(t_2A_2)$, and $h(t_{12}A_{12})$ in (15).

In Appendix C, the convergence of the kernel density estimator and the Kozachenko-Leonenko estimator is compared with that of their offset versions by applying each of them to Gaussian images.

IV. Information retrieved and its estimators

As already noted in Section III, the quantity N_I defined by (7) is the least upper bound on the number of distinct points that can be identified reliably across two images of a scene under illuminants e_1 and e_2. This identification does, however, assume that matching is by maximum likelihood [25] or its equivalent. That is, for a sample of N points (10), a particular point (x_j, y_j) with color value a_{2j} is matched to the point (x_i, y_i) with color value a_{1i} that maximizes the probability of a_{1i} given a_{2j}.

For maximum-likelihood matching to be applied, the conditional pmfs $p_{1|2}$ must be known or estimated reliably, which is not generally feasible. Instead, a nearest-neighbor criterion may be used, which may not be optimal [25], but may approach optimality with a judiciously chosen metric. The information retrieved is then the logarithm of the maximum number of distinct points that can be reliably identified by nearest-neighbor matching across two images of a scene under different illuminants. It is always lower than or equal to the information available. An equivalent definition of information retrieved in the more general case is μ-capacity; see e.g. [25]. By contrast with the information available, invertible transformations of the sample values (10) can increase the information retrieved. With a nearest-neighbor criterion defined in accordance with a measure μ, typically the Euclidean or Mahalanobis distance, the transformations t_1 and t_2 that make the transformed values

$$\begin{aligned}&\{t_1a_{1i} \mid i = 1, \ldots, N\}\,,\\ &\{t_2a_{2i} \mid i = 1, \ldots, N\}\,,\end{aligned} \tag{16}$$

as close to each other as possible maximize the information retrieved. These optimal transformations depend on the sample values (10); they were here constrained to be linear.

In practice, estimating the information retrieved is more difficult than estimating the information available [25]. Approximations, upper bounds (tighter than the trivial one given by the information available) and lower bounds have been proposed for μ-capacity [25]–[27] and specifically for color-dependent identification [28], [29]. The approximations in [28], [29] were based, as in [27], on additive Gaussian noise channels, for which the nearest-neighbor criterion with a Mahalanobis distance coincides with the maximum-likelihood criterion [6].

In more detail, suppose that the means of t_1A_1 and t_2A_2 coincide, so that, for some zero-mean random variable W,

$$t_2A_2 = t_1A_1 + W\,. \tag{17}$$

If the random variables in (17) are Gaussian and t_1A_1 and W independent of each other, then the mutual information takes a particularly simple form, which may be used to approximate the information retrieved. This Gaussian approximation $I_{\mathrm{GA}}(t_1A_1; t_2A_2)$ is given by

$$I_{\mathrm{GA}}(t_1A_1; t_2A_2) = \frac{1}{2}\log\left(\frac{|\mathrm{Var}(t_2A_2)|}{|\mathrm{Var}(t_2A_2 - t_1A_1)|}\right). \tag{18}$$

An estimator $\hat{I}_{\mathrm{GA}}$ of I_{GA} is obtained by replacing the variances in (18) by the sample variances. A slightly different version in which the variance of the numerator of (18) is assumed to be $|\mathrm{Var}(t_1A_1) + \mathrm{Var}(t_2A_2 - t_1A_1)|$ was used in [28], [29]. The results obtained with the two versions were similar.

If, instead, the random variables in (17) are not necessarily Gaussian but t_1A_1 and W are still independent of each other, then the mutual information can be expressed as a difference between two differential entropies, which may also be used to approximate the information retrieved. The additive approximation $I_{\mathrm{AA}}(t_1A_1; t_2A_2)$ is given by

$$I_{\mathrm{AA}}(t_1A_1; t_2A_2) = h(t_2A_2) - h(t_2A_2 - t_1A_1)\,. \tag{19}$$

An estimator $\hat{I}_{\mathrm{AA}}$ of I_{AA} is obtained by replacing the differential entropies h in (19) by the offset Kozachenko-Leonenko estimator $\hat{h}_{\mathrm{KLo}}$. These approximations (18) and (19) are rough but useful and were used (Section VIII) to explore optimal post-sensor processing.

A very different approach to estimating the information retrieved, outlined in [28], [29], is to quantify the entropy of point matching. Some results have been reported in [11], [14].

A. Nearest-neighbor errors and entropy of point matching

The estimator of the information retrieved that was developed in [11], [28], [29] was based on the entropy of the error of a theoretical observer making nearest-neighbor matches across two images of a scene under different illuminants. A slightly different interpretation of that estimator can be derived from the relationship between the minimum number of bits needed to encode a random variable and the entropy of that variable [6, Chapter 5].

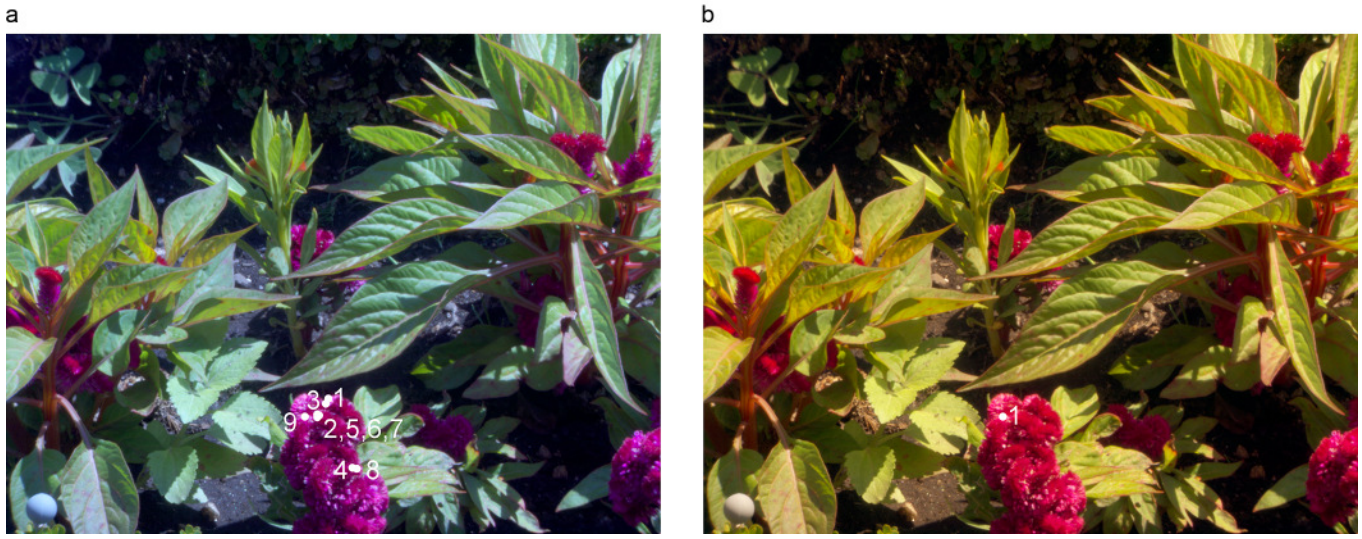

Fig. 1. Identification errors across images of a scene under daylight illuminants with correlated color temperatures of (a) 25000 K and (b) 4000 K. After scaling of sensor responses to the spatial mean (see Section VII), the color values of points in image a marked 1, 2, ..., 8 (not all are distinguishable) are all closer to those of point 1 in image b than those of point 9, the correct match. For the purposes of illustration, the images themselves have not been scaled to the mean.

Without any prior information, the number of bits needed to encode a sample of N points from an image of a scene with color values $t_1 a_{1i}$ under illuminant e_1 is $\log N$, i.e., the entropy of a random variable with a discrete uniform distribution. But with prior information, namely the color values $t_2 a_{2j}$ of the N points of the same scene under illuminant e_2, the number of bits is reduced. Thus, for a point (x_j, y_j) with color value $t_2 a_{2j}$ under illuminant e_2, there is only a subset of points (x_i, y_i) whose color values $t_1 a_{1i}$ under illuminant e_1 are sufficiently close to $t_2 a_{2j}$ to be confused with (x_j, y_j), with respect to some nearest-neighbor criterion μ. The largest such subset excluding (x_j, y_j) is given by

$$\{(x_i, y_i) \mid \mu(t_1 a_{1i}, t_2 a_{2j}) < \mu(t_1 a_{1j}, t_2 a_{2j})\} . \tag{20}$$

The number m of points in (20) (where m should not to be confused with the dimensional variable m of Section III-B) is an instance of a random variable with pmf p_m specifying the number of mismatches. The entropy of that random variable,

$$H(M) = -\sum_{m=0}^{N-1} p_m \log p_m ,$$

is the entropy of point matching [11]. It yields the number of bits needed to encode the N surfaces in an image of a scene under illuminant e_1 given an image of the same scene under illuminant e_2. If matching is perfect, so there are no incorrect matches for any point, then $H(M) = 0$. Conversely, if matching is uniformly random, then $H(M) = \log N$. Figure 1 shows an example of the actual errors in point matching.

The difference $\log N - H(M)$ is the reduction in number of bits needed to encode the N points in an image of a scene under illuminant e_1 given an image of the same scene under illuminant e_2. An estimator of information retrieved by nearest-neighbor matching $I_{\mathrm{NN}}(t_1 A_1; t_2 A_2)$ is defined precisely by that difference. That is,

$$I_{\mathrm{NN}}(t_1 A_1; t_2 A_2) = \log N - H(M) .$$

The dependence of this estimator on the distributions of color values in the images and on the nearest-neighbor criterion is evident in eq. (20).

B. Grassberger estimator

The naïve estimator of the entropy $H(M)$ in Section IV-A is usually biased when the number of non-zero probabilities p_m is close to the sample size N, and a bias-corrected estimator due to Grassberger [19] was therefore used in this analysis. If ψ denotes the digamma function, the Grassberger estimator $\hat{H}_{\mathrm{G}}(M)$ of $H(M)$ is defined, in nats, by

$$\hat{H}_{\mathrm{G}}(M) = \ln N - \sum_{m=0}^{N-1} p_m \psi(N p_m) .$$

The Grassberger estimator $\hat{I}_{\mathrm{NNG}}(t_1 A_1; t_2 A_2)$ of the information retrieved is accordingly

$$\hat{I}_{\mathrm{NNG}}(t_1 A_1; t_2 A_2) = \log N - \hat{H}_{\mathrm{G}}(M) .$$

In Appendix D, the naïve estimator and the Grassberger estimator are compared by applying each of them to Gaussian images.

V. Experimental simulations

As mentioned earlier, calculations were based on a set of 50 hyperspectral images of rural and urban scenes [15], [30], illuminated by simulated daylights with CCTs of 4000 K, 6500 K, and 25000 K. The five camera sensor spectral sensitivities are shown in Fig. 2 for (a) an Agilent CMOS sensor array from a Concord EyeQ digital camera, data digitized from Fig. 8(C) in [31]; (b) a Foveon X3 sensor array from a Sigma SD9 digital camera, data digitized from Fig. 7 in [31]; (c) a Kodak frame-transfer CCD sensor array from a Kodak DCS-460 digital camera, data digitized from Fig. 8(A) in [31]; (d) a CCD sensor array from a Nikon D1 digital camera, data

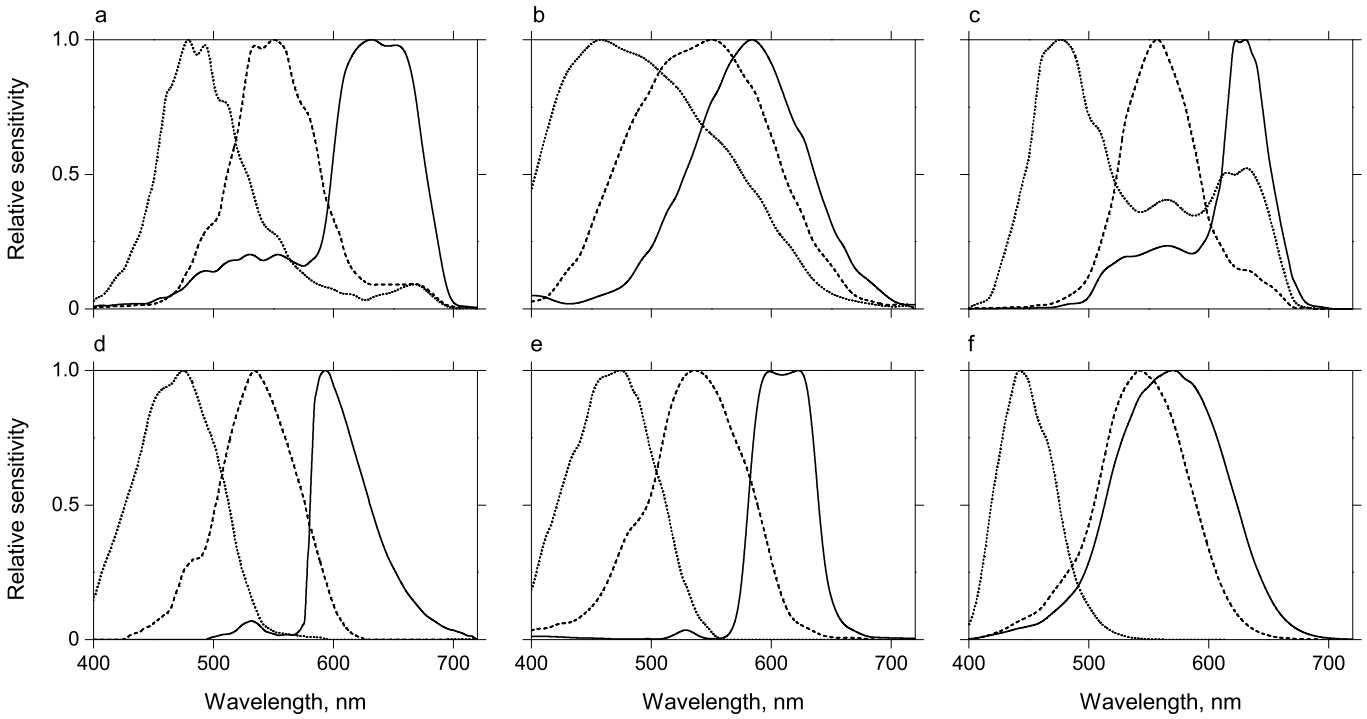

Fig. 2. Normalized spectral sensitivities for (a) an Agilent CMOS sensor array from a Concord EyeQ digital camera [31], (b) a Foveon X3 sensor array from a Sigma SD9 digital camera [31], (c) a Kodak CCD sensor array from a Kodak DCS-460 digital camera [31], (d) a CCD sensor array from a Nikon D1 digital camera [32], (e) a Sony CCD sensor array from a Hewlett-Packard digital camera [31], and (f) the cone photoreceptors of the human eye [33].

Fig. 3. Color images of a sample of eight scenes from the 50 scenes analyzed in this work. The upper row is from mainly vegetated scenes and the lower row from mainly nonvegetated scenes.

digitized from Fig. 9 (top) in [32]; and (e) a Sony interline CCD sensor array from a Hewlett Packard digital camera, data digitized from Fig. 8(B) in [31]. The spectral sensitivities of the cone photoreceptors of the eye in (f) are from the Stockman and Sharpe fundamentals [33].

The 50 hyperspectral images were divided into two groups of 29 mainly vegetated scenes and 21 mainly nonvegetated scenes [15]. Example images are shown in Fig. 3. Each hyperspectral image had spatial dimensions $\leq 1344 \times 1024$ pixels and spectral range 400–720 nm sampled at 10-nm intervals. At each pixel (x, y), the effective spectral reflectance $r(\lambda; x, y)$ was therefore defined at 33 values of λ (Section III and Footnote 1), which is sufficiently dense for the present purposes [34], [35]. Further details about the hyperspectral images and effective global illuminants and reflectances can be found in [15]. To reduce computation time, and to accommodate the approximately 1.3 pixel line spread function of the camera system [15], images were spatially subsampled,

with only alternate pixels being used, so that the subsampled images had spatial dimensions $\leq 672 \times 512$. Results obtained with these subsampled images were closely similar to those obtained with full-sized images.

For convenience, the three daylight illuminants were reconstructed from daylight basis functions [36], although the use of these functions has no particular significance in this analysis. The two pairs with CCTs of 4000 K and 6500 K and of 6500 K and 25000 K had similar chromaticity differences and the remaining pair with CCTs of 25000 K and 4000 K had a much larger chromaticity difference.

The kernel density estimator (Section III-A) was implemented with the Kernel Density Estimation Toolbox, KDE 2003, for MATLAB (The MathWorks, Inc., Natick, MA), at http://www.ics.uci.edu/~ihler/code/. The Kozachenko-Leonenko estimator (Section III-B) was implemented with the Approximate Nearest Neighbor Searching Library, ANN, ver 1.1.1, at http://www.cs.umd.edu/~mount/ANN/ [37], which contained an efficient C++ routine for exact nearest-neighbor search. The Kraskov-Stögbauer-Grassberger estimator (Section III-C) was implemented with MILCA 2004 for MATLAB at http://www.klab.caltech.edu/~kraskov/MILCA/ [21], [23]. The naïve estimator and the Grassberger estimator of information retrieved (Sections IV-A and IV-B) were implemented with the ANN, ver 1.1.1, library.

VI. Estimates of information available

Figure 4 shows the convergence of the estimates of the information available with increasing size N of random samples from Scene d of Fig. 3 under daylight illuminants with CCTs of 25000 K and 4000 K. The sensors were from the Foveon X3 sensor array (Fig. 2 b). The estimators were the equivalent Gaussian estimator $\hat{I}_{\mathrm{EG}}$ (dash-dotted curve), Section III-D; the offset kernel density estimators $\hat{I}_{\mathrm{KDo}}$ each with a different automatic bandwidth-selection method, namely rule-of-thumb (dashed curve), Sections III-A and III-D, Appendix A, and likelihood cross-validation (dotted curve), Sections III-A and III-D, Appendix A; and the offset Kozachenko-Leonenko estimator $\hat{I}_{\mathrm{KLo}}$ (solid curve), Sections III-B and III-D. Sample size N ranged from 2^3 to 2^{18} for all estimates except for the kernel density estimator with likelihood cross-validation for which the maximum sample size was limited to 2^{16} because of the lengthy computation time required with larger samples.

The kernel density estimator $\hat{I}_{\mathrm{KDo}}$ with automatic bandwidth selection by rule-of-thumb and the offset Kozachenko-Leonenko estimator $\hat{I}_{\mathrm{KLo}}$ converged to similar values, whereas the equivalent Gaussian estimator $\hat{I}_{\mathrm{EG}}$ appears biased upwards with respect to these estimators. The kernel density estimator $\hat{I}_{\mathrm{KDo}}$ with automatic bandwidth selection by likelihood cross-validation did not have an obvious asymptote, even as N approached the maximum sample size.

The rate of convergence depends on the linear correlation between the images of the scene. With this particular scene, illuminants, and the Foveon X3 sensor set, the correlation between the two images was very strong, with average correlation coefficient[2] $\rho = 0.9994$.

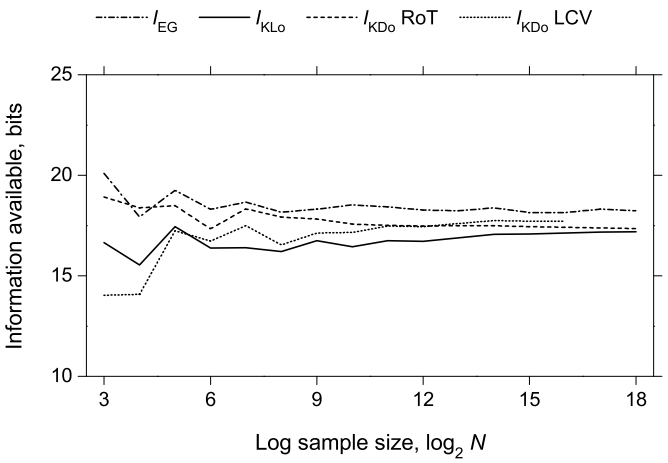

Fig. 4. Sample-size dependence of estimates of the information available across images of Scene d of Fig. 3 under daylight illuminants with CCTs of 25000 K and 4000 K. Information available is plotted against sample size N for the equivalent Gaussian estimator $\hat{I}_{\mathrm{EG}}$ (dash-dotted curve), offset kernel density estimator $\hat{I}_{\mathrm{KDo}}$ with rule-of-thumb (RoT, dashed curve) and likelihood cross-validation (LCV, dotted curve), and offset Kozachenko-Leonenko estimator $\hat{I}_{\mathrm{KLo}}$ (solid curve). The sensors were from the Foveon X3 sensor array (Fig. 2 b).

Appendices B and C set out a systematic comparison of these and other estimators with both strongly and weakly correlated synthetic Gaussian images, the outcome of which suggests that the offset Kozachenko-Leonenko estimator $\hat{I}_{\mathrm{KLo}}$ was the best estimator of those tested. In the following, estimates of the information available are reported only for $\hat{I}_{\mathrm{KLo}}$.

Figure 5 shows a dotplot of the mean estimated information available with the estimator $\hat{I}_{\mathrm{KLo}}$ and each of the sets of sensors of Fig. 2. Different symbols show results from scenes under daylight illuminants with large chromaticity differences (circles) and small chromaticity differences (inverted and upright triangles). Mean estimated information available for illuminants with CCTs of 25000 K and 4000 K varied over the sets of sensors from 15.5 to 18.0 bits. From (7), these values correspond to 4.7×10^4 and 2.7×10^5 distinct identifiable points per scene. Mean estimated information available for illuminants with CCTs of 25000 K and 6500 K and of 4000 K and 6500 K were, as expected, larger, and varied over the sets of sensors from 18.5 to 20.6 bits, which correspond to 3.6×10^5 and 1.54×10^6 distinct identifiable points per scene.

The estimated information available was usually less with mainly vegetated scenes than with mainly nonvegetated scenes, by 0.7–1.5 bits, depending on the set of sensors (cf. [15]). The estimates varied little over scenes, with SDs of 1.1–1.4 bits, a result that extends an earlier finding [28] with the Gaussian approximation (18) of information retrieved described in Section IV.

[2]The average correlation coefficient was defined as $(\mathrm{cor}\,(R_1, R_2) + \mathrm{cor}\,(G_1, G_2) + \mathrm{cor}\,(B_1, B_2))/3$, where $\mathrm{cor}\,(X, Y) = \mathrm{cov}\,(X, Y)/(\mathrm{var}\,(X)\mathrm{var}\,(Y))^{1/2}$.

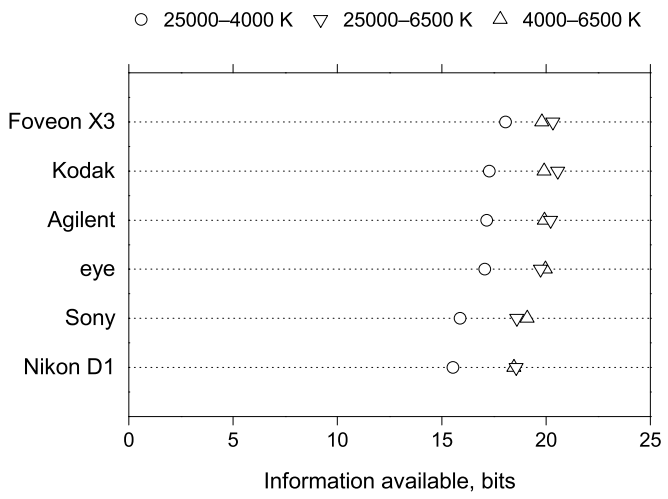

Fig. 5. Mean estimated information available across images of scenes under pairs of daylight illuminants with correlated color temperatures of 25000 K and 4000 K (circles), 25000 K and 6500 K (inverted triangles), and 4000 K and 6500 K (upright triangles). Estimates from the offset Kozachenko-Leonenko estimator $\hat{I}_{\mathrm{KLo}}$ (Sections III-B and III-D) are for each of the sensors of Fig. 2. Means were taken over 50 scenes. SDs were 1.1–1.4 bits. The horizontal scale has been extended to allow comparison with other information plots.

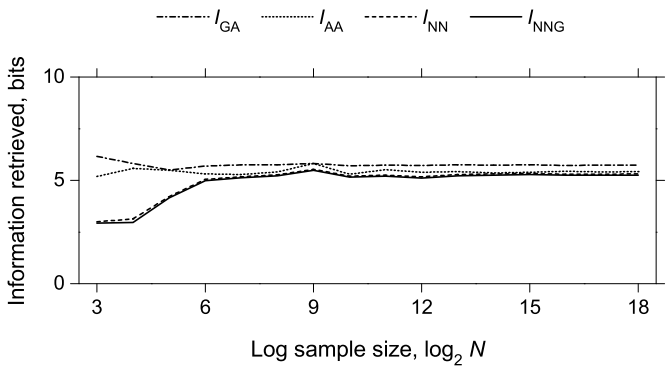

Fig. 6. Sample-size dependence of estimates of the information retrieved across images of Scene d of Fig. 3 under daylight illuminants with CCTs of 25000 K and 4000 K. Information is plotted against sample size N for Gaussian approximation $\hat{I}_{\mathrm{GA}}$ (dash-dotted curve), additive approximation $\hat{I}_{\mathrm{AA}}$ (dotted curve), naïve estimator of the information retrieved $\hat{I}_{\mathrm{NN}}$ dashed curve), and the Grassberger estimator $\hat{I}_{\mathrm{NNG}}$ (solid curve).

VII. Estimates of information retrieved

The information retrieved with a nearest-neighbor criterion depends critically on the transformations t_1 and t_2 of the sample values (16). Typically, t_1 and t_2 represent scaling by von Kries' rule. As originally conceived by von Kries [1], [2], the eponymous scaling assumes that the spectral effects of the prevailing light on the sensitivity of long-, medium-, and short-wavelength-sensitive cone photoreceptors of the eye are contingent only on the response of each photoreceptor class and in a linear way. But von Kries' rule leaves unspecified precisely how the prevailing light determines the coefficients that describe the adjustment of each photoreceptor sensitivity (see e.g. [38], [39]).

Many machine models of color constancy, including Land's Retinex models [40], [41], assumed that von Kries scaling applies also to lights reflected from surfaces. Subsequent analysis showed that it does indeed give a good description of the effects of illuminant changes with artificial scenes [42], [43] and natural scenes [30]. Departures from von Kries' rule have been addressed by relaxing the scaling so that it is dependent on the signals from all three sensor classes [44]–[46] or by making it nonlinear [47].

For von Kries scaling proper, the transformations t_1 and t_2 (16) can each be expressed as a diagonal matrix transformation [48]. The coefficients of the transformations depend on the spectral sensitivities of the sensors, the scene being imaged, and the illuminants. A common procedure for determining the coefficients is by the so-called gray-world assumption [41], [49]; that is, taking the inverse of the spatial average of the N sample color values a_{1i}, $i = 1, \ldots, N$, for t_1, and $a_{2i}, i = 1, \ldots, N$, for t_2. In deciding on the coefficients, however, it is important to distinguish between the problem of estimating the spectrum e_1 or e_2 [50], [51] and the problem of finding an approximately correct one-to-one correspondence between triplets given the illuminants [52]. Only the second problem is relevant here, and it does not matter whether the mean reflectance of the scene is neutral.

Figure 6 shows the convergence of the estimates of the information retrieved with increasing size N of random samples from Scene d of Fig. 3 under daylight illuminants with CCTs of 25000 K and 4000 K. The sensors were from the Foveon X3 sensor array (Fig. 2 b). The estimators were the Gaussian approximation $\hat{I}_{\mathrm{GA}}$ (dash-dotted curve), Section IV; the additive approximation $\hat{I}_{\mathrm{AA}}$ (dotted curve), Section IV; the naïve estimator of information retrieved $\hat{I}_{\mathrm{NN}}$ (dashed curve), Section IV-A; and the Grassberger estimator $\hat{I}_{\mathrm{NNG}}$ (solid curve), Section IV-B. The sample size N ranged from 2^3 to 2^{18} for all estimators.

The differences between the naïve estimator $\hat{I}_{\mathrm{NN}}$ and the Grassberger estimator $\hat{I}_{\mathrm{NNG}}$ were very small in this example. The additive approximation $\hat{I}_{\mathrm{AA}}$ was slightly biased upwards with respect to these two estimators, and the Gaussian approximation $\hat{I}_{\mathrm{GA}}$ rather more so. Small differences between $\hat{I}_{\mathrm{NN}}$ and $\hat{I}_{\mathrm{NNG}}$ also emerged with strongly correlated synthetic Gaussian images (see Appendix D). In the following, estimates of the information retrieved are reported only for $\hat{I}_{\mathrm{NNG}}$.

Figure 7 shows a dotplot of the mean estimated information retrieved with the Grassberger estimator $\hat{I}_{\mathrm{NNG}}$ and each of the sets of sensors of Fig. 2. Mean estimated information retrieved from scenes under daylight illuminants with CCTs of 25000 K and 4000 K varied over the sets of sensors from 5.9 to 9.2 bits, which correspond, respectively, to 62 and 592 distinct identifiable points per scene. As with information available, information retrieved was higher for illuminants with smaller chromaticity differences. Thus, mean estimated information retrieved from scenes under illuminants with CCTs 25000 K and 6500 K and of 4000 K and 6500 K varied over the sets of sensors from 8.0 to 11.6 bits, corresponding to 260 and 3.2×10^3 distinct identifiable points per scene. The mean estimated information retrieved was again less with mainly vegetated scenes than with mainly nonvegetated scenes, by 0.2–1.0 bits, depending on the set of sensors. The estimates

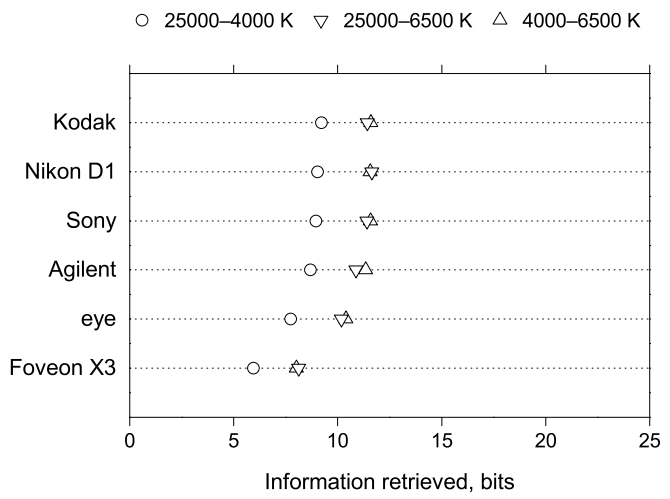

Fig. 7. Mean estimated information retrieved with the Grassberger estimator $\hat{I}_{\mathrm{NNG}}$ and von Kries scaling of the sensor signals. SDs were 1.0–1.3 bits. Other details as for Fig. 5.

varied little over scenes, with SDs of 1.0–1.3 bits.

The information retrieved is much less than the information available, as is evident from a direct comparison of Fig. 7 with Fig. 5. For all except the Foveon X3 sensor set, the mean estimated information available was 50–62% of the information available; for the Foveon X3, it was only 38%. The ordering of the sets of sensors by information retrieved is also quite different from that by information available (Fig. 5).

VIII. Optimized sensor transformations

Fortunately, more information can be retrieved if spectral sensitivities are first transformed effectively by sharpening [43]–[45], before von Kries scaling, so that nearest-neighbor matching approaches more closely maximum-likelihood matching. The required sharpening transformation can be represented as a linear combination of the signals from the sensors of the camera (or from the cone photoreceptors of the eye, for which these transformations have been justified on both behavioral and physiological grounds, although their effects extend beyond narrowing spectral sensitivities [53], [54]; see also [11], [13]). When sharpening is combined with von Kries scaling, the transformations t_1 and t_2 of the random variables A_1 and A_2 (16) can be written as $t_1 = t'_1 t_0$ and $t_2 = t'_2 t_0$, where t_0 represents the sharpening transformation (the same for the two images) and t'_1 and t'_2 the diagonal matrix transformations representing von Kries scaling (different for the two images). The task is to find the coefficients of t_0 that maximize the information retrieved between $t'_1 t_0 A_1$ and $t'_2 t_0 A_2$.

Given two images of a scene under two different illuminants, one way to estimate the coefficients of t_0 from (17) is to represent the random variable A_2 as a linear transform of A_1 and a noise term, that is,

$$A_2 = t_0^{-1} \left(t_2'^{-1} t'_1 \right) t_0 A_1 + t_0^{-1} t_2'^{-1} W \,,$$

and then find a linear mapping $t = t_0^{-1}(t_2'^{-1} t'_1) t_0$ that maximizes the information retrieved between tA_1 and A_2 (alternatively, it is possible to find a linear mapping t that minimizes the sum of the squares of the differences between tA_1 and A_2, as in [44]). A unique solution for t_0 is obtained from the eigenvectors of t (see e.g. [44]) by setting the diagonal elements of t_0 to unity and preserving the ordering of the spectral locations of the modified spectral sensitivities. The information retrieved may be maximized with one of the approximations or estimators introduced in Sections IV, IV-A, and IV-B. For the Gaussian approximation $\hat{I}_{\mathrm{GA}}$ (Section IV) the solution is given by the linear mapping t for which tA_1 and $A_2 - tA_1$ are uncorrelated. If $\mathrm{Cov}\,(A_1, A_2)$ is the matrix of covariances between the elements of A_1 and A_2, that is,

$$\mathrm{Cov}\,(A_1, A_2) = \begin{pmatrix} \mathrm{cov}\,(R_1, R_2) & \mathrm{cov}\,(R_1, G_2) & \mathrm{cov}\,(R_1, B_2) \\ \mathrm{cov}\,(G_1, R_2) & \mathrm{cov}\,(G_1, G_2) & \mathrm{cov}\,(G_1, B_2) \\ \mathrm{cov}\,(B_1, R_2) & \mathrm{cov}\,(B_1, G_2) & \mathrm{cov}\,(B_1, B_2) \end{pmatrix} ,$$

then the optimal linear mapping is given by

$$t = \mathrm{Cov}\,(A_2, A_1)\,(\mathrm{Var} A_1)^{-1} \quad . \tag{21}$$

For the additive approximation $\hat{I}_{\mathrm{AA}}(tA_1; A_2)$ (Section IV) and the Grassberger estimator $\hat{I}_{\mathrm{NNG}}(tA_1; A_2)$ (Section IV-B), there is no analytic solution.

A simplex optimization algorithm (see e.g. [55]) was used to find the coefficients of the sharpening transformation t_0 for which information retrieved with the Grassberger estimator $\hat{I}_{\mathrm{NNG}}$ was maximum. Random samples of size 2^{12} were used to compute the value of $\hat{I}_{\mathrm{NNG}}$ at each iteration of the simplex algorithm. The algorithm was initialized with the solution (21) maximizing the Gaussian approximation $\hat{I}_{\mathrm{GA}}$ to the information retrieved, although similar results, not reported here, were obtained with initial values from the additive approximation $\hat{I}_{\mathrm{AA}}$ and by least squares in exploratory simulations with the spectral sensitivities of the photoreceptors of the eye.

The six off-diagonal coefficients of the optimal sharpening transformation varied markedly from camera to camera owing to their different sensor spectral sensitivities. Nevertheless, with the exception of the Foveon X3 sensor set (Fig. 2 b), the optimal coefficients varied little over the 50 scenes and the three illuminant pairs, with SDs from 3.7×10^{-3} to 7.9×10^{-2} bits. For the Foveon X3, the SDs of the optimal coefficients were considerably larger, from 0.1 to 0.6 bits.

The optimal sharpening transformation can be averaged over scenes and illuminant pairs, yielding a unique, fixed transformation for each set of sensors. Figure 8 shows a dotplot of the mean estimated information retrieved with the Grassberger estimator $\hat{I}_{\mathrm{NNG}}$ and the sensors of Fig. 2 with a fixed sharpening transformation. The mean estimated information retrieved as a percentage of the information available was 63–81%. This is larger than for estimates without sensor sharpening and with von Kries scaling by 18%–25% (see Section VII). As with the previous estimates without sensor interactions (Fig. 7), the estimates of the information retrieved varied little over scenes, with SDs lower than 1.5 bits, except again for the Foveon X3, with SDs as high as 2.3 bits.

Rather than keeping the sharpening transformation t_0 fixed for each set of sensors, it can be allowed to vary over scenes and illuminants. Figure 9 shows a dotplot of the

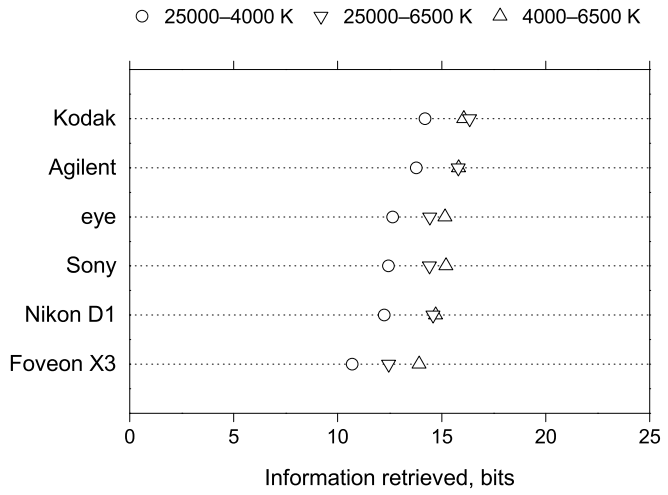

Fig. 8. Mean estimated information retrieved with the Grassberger estimator $\hat{I}_{\mathrm{NNG}}$ and optimal sensor sharpening fixed for each camera and the eye. SDs were 1.1–2.3 bits. Other details as for Fig. 5.

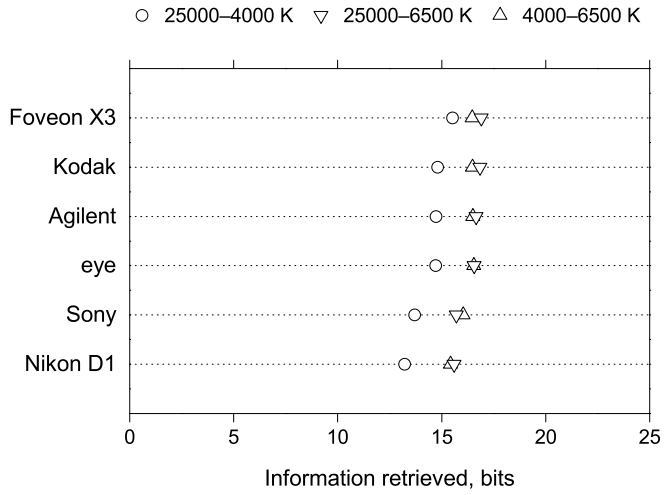

Fig. 9. Mean estimated information retrieved with the Grassberger estimator $\hat{I}_{\mathrm{NNG}}$ and optimal sensor sharpening varying over cameras and the eye, scenes, and illuminant pairs. SDs were 0.8–1.2 bits. Other details as for Fig. 5.

mean estimated information retrieved with the Grassberger estimator $\hat{I}_{\mathrm{NNG}}$ and the sets of sensors in Fig. 2 with variable sensor sharpening. The mean estimated information retrieved as a percentage of the information available was 82–86%. This is larger than for estimates with a fixed sharpening transformation by 3–10%, except for the Foveon X3, for which the increase was considerably larger, by 21%. Such an increase is consistent with the larger variabilities in the coefficients of the optimal sharpening transformation noted earlier for the Foveon X3. The maximum SD over scenes declined from 1.5 to 1.2 bits for all the sets of sensors, except for the Foveon X3, for which the decline in SD was more substantial, from 2.3 to 0.9 bits.

The extent of the mean information retrieved as a proportion of the information available with a variable sharpening transformation may be a slight underestimate. An analysis of the information retrieved with the cone photoreceptors of the eye [11] showed that with a variable sharpening transformation, the convergence of the Grassberger estimator $\hat{I}_{\mathrm{NNG}}$ failed to asymptote at the maximum sample size N available, i.e. 2^{18}.

IX. Discussion

Capturing scene information from image colors depends crucially on the choice of camera sensors. Although not all of the information available can be retrieved with any particular set of sensors, providing that the sensor spectral sensitivities are optimally modified with a sharpening transformation, the information retrieved can approach the information available, depending of course on the scene and illumination. As shown in this work, estimating the continuous and discrete informational quantities involved and comparing them over different sets of camera sensors is not straightforward, but clear differences between sensor sets did emerge over a range of natural scenes and daylight illuminants. Most notably, with the best sensor set about 390% more points could be identified per scene than with the worst. In the following subsections, some of the factors contributing to these differences in performance are examined in more detail.

A. Estimators and estimates

For the information available, which requires an estimator for trivariate continuous random variables (Section III), the offset Kozachenko-Leonenko estimator proved the best of the several estimators tested: in addition to its fast convergence, it has the important property of asymptotic unbiasedness [20], [24]. For the information retrieved, which requires an estimator for discrete random variables (Section IV), the Grassberger estimator [19], which also has the property of asymptotic unbiasedness, sufficed. Both the offset Kozachenko-Leonenko estimator and the Grassberger estimator yielded good estimates with Gaussian images (Appendices C and D).

Importantly, despite the different nature of these two estimators, one for continuous variables, the other for discrete, the estimated information retrieved from Gaussian images converged to the estimated information available as sample size increased. This convergence provided an essential control, for with Gaussian images a nearest-neighbor criterion based on a Mahalanobis distance coincides with a maximum-likelihood criterion.

With the 50 natural scenes considered here, illuminated by daylights with the largest chromaticity difference, i.e. with CCTs of 25000 K and 4000 K, the mean estimated information available across each pair of images varied from 15.5 bits to 18.0 bits, depending on the set of sensors. These values correspond to 4.7×10^4 and 2.7×10^5 distinct identifiable elements or points per scene, i.e., a ratio of 570% between the best and worst set of sensors. The mean estimated information retrieved with the same daylight illuminants, and with a sharpening transformation optimized for scenes, illuminants, and sensor sets, varied from 13.2 to 15.5 bits, respectively, corresponding to 9.5×10^3 and 4.7×10^4 distinct identifiable

points per scene, i.e., a ratio of 490% between the best and worst set of sensors.

For the human eye, the mean estimated information available for the same daylight illuminants was 17.1 bits, corresponding to 1.37×10^5 distinct identifiable points per scene, and the mean estimated information retrieved was 14.7 bits, corresponding to 2.7×10^4 distinct identifiable points per scene.

For daylight illuminants with smaller chromaticity differences, both the information available and the information retrieved were fittingly larger. For illuminants with CCTs of 25000 K and 4000 K and of 4000 K and 6500 K, the mean estimated information available varied over the sets of sensors from 18.5 to 20.6 bits, corresponding to 3.6×10^5 and 1.54×10^6 distinct identifiable points per scene. The mean estimated information retrieved varied from 15.4 to 16.9 bits, corresponding to 4.4×10^4 and 1.22×10^5 distinct identifiable points per scene.

Both the estimates of the information available and the information retrieved were fairly stable over scenes, with SDs of about 1.2 bits for each. Estimates were larger for nonvegetated scenes than for vegetated ones, by about 1.0 and 0.5 bits, respectively, consistent with the lower frequency of metamerism found in nonvegetated scenes [15]. Although expressed as points per scene, these estimates of the numbers of distinct identifiable points refer effectively to the number of surfaces or surface elements in a scene with distinct spectral reflectances. Such numbers therefore express precisely how well the elements of a scene can be identified by their color values, independent of spatial context or their spatial position. Incorporating processing noise into the estimates would inevitably lower these values, as has been demonstrated elsewhere by including the probabilistic nature of human judgments in estimates of the number of perceptually distinct surface colors in natural scenes [56].

B. Camera sensors

Of the five sets of camera sensors tested, the Foveon X3 [31] yielded the greatest information available (Fig. 5) and with a sharpening transformation optimized for both scenes and daylight illuminants, it also yielded the greatest information retrieved (Fig. 9), namely 15.5 bits, corresponding to 4.7×10^4 distinct point per scene, although it was not so successful with a fixed sharpening transformation (Fig. 8). More generally, the ranking of camera sensors by information available (Fig. 5) coincided with the ranking by information retrieved with a variable sharpening transformation (Fig. 9), unlike that with von Kries scaling alone (Fig. 7) or with a fixed sharpening transformation (Fig. 8).

The advantage of a fixed over a variable sharpening transformation for each camera is in the simplification of the estimation problem, i.e. determining three coefficients instead of nine [44], [46]. But the simplification is at a cost: for the Agilent, Kodak, Nikon D1, and Sony sets of sensors, the reduction in information retrieved with a fixed rather than variable sharpening transformation was only about 3–6% of the information available. For the cone photoreceptors of the eye, it was about 10% and for the Foveon X3 set of sensors, 21%.

In spite of the marked differences between some sensor sets, caution should be exercised in extrapolating these results to camera performance in practice. This analysis took no account of design features such as the spatial resolution of the camera, its color depth, and the level of internal noise, all of which can influence the identifiability of reflected spectra.

C. Sampling limits

In general, with a particular set of scenes and daylight illuminants, the factors that determine the information available are primarily the spectral positions of the sensors, i.e. the wavelengths at which sensitivity is maximum, and the spectral widths of the sensors. Determining the optimum spectral positions of a set of sensors is a sampling problem, complicated by the variation with wavelength of reflected spectra under changes in illuminant. Determining the optimum spectral width is also a sampling problem, albeit constrained by the trade-off between the von Kries invariance provided by an infinitesimal spectral width and the spectral coverage provided by a spectral width that extends over the visible spectrum. Both factors can be modulated by the sharpening transformation discussed in Section VIII. For the sets of sensors considered here, the mean spacing of the peaks actually accounted for little of the variance in the information available. The area under the spectral sensitivity curves accounted for somewhat more, i.e. 20–65%. Ultimately, optimization is an empirical issue.

Although most of the information available in a scene can be retrieved by von Kries scaling and sensor sharpening, i.e. 82–86% depending on the set of sensors and the daylight illuminants, still more information can be retrieved if nonlinear sensor transformations or probabilistic methods are allowed. As indicated in Section IV, the ideal approach to matching would be by maximum likelihood, but this would require the estimation of the conditional probability density functions of sensor signals contingent on scene and illuminants, which, in turn, require very large image samples. If that were achievable, then the information retrieved would tend to the information available. Given their simplicity, however, linear models performed remarkably well in retrieving the information available.

Even so, it is important to recognize the real limits on the recovery of scene information from image colors. With the largest difference in daylight illuminants, a retrieval of 82–86% of the information available when interpreted as numbers of distinct identifiable points per scene represents just 8–22% of the points available. The fact that for the cone photoreceptors of the eye the number of distinct identifiable points per scene falls in the upper part of this range suggests that even with the reduction in performance associated with noise in the photoreceptors and in postreceptoral processing, the spectral positioning of cone photopigments may be close to optimal with natural scenes.

ACKNOWLEDGMENT

We thank M. S. Mould, R. Petersen, K. Żychaluk, G. Feng, and R. Senatore for useful comments and for critically reading

the manuscript; J. Worthey for advice on sources of spectral data for digital trichromatic cameras; and S. M. C. Nascimento and K. Amano for use of a set of hyperspectral images of natural scenes. This work was supported by the EPSRC (grant nos. EP/B000257/1 and EP/E056512/1).

REFERENCES

[1] J. von Kries, "Theoretische Studien über die Umstimmung des Sehorgans," in *Festschrift der Albrecht-Ludwigs-Universität in Freiburg zum fünfzigjährigen Regierungs-Jubiläum Seiner Königlichen Hoheit des Grossherzogs Friedrich*. C. A. Wagner's Universitäts-Buchdruckerei, Freiburg i. Br., 1902, pp. 145–158, [Translation: D. L. MacAdam, *Sources of Color Science*, MIT Press, Cambridge, 1970].

[2] ——, "Die Gesichtsempfindungen," in *Handbuch der Physiologie des Menschen*, W. Nagel, Ed. Braunschweig: Vieweg und Sohn, 1905, vol. 3 Physiologie der Sinne, pp. 109–282, [Translation: D. L. MacAdam, *Sources of Color Science*, MIT Press, Cambridge, 1970].

[3] G. Wyszecki and W. Stiles, *Color Science: concepts and methods, quantitative data and formulae*, 2nd ed. New York: John Wiley, 1982.

[4] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[5] ——, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 623–656, 1948.

[6] T. Cover and J. Thomas, *Elements of Information Theory*, ser. Wiley series in telecommunications. New York: John Wiley & Sons, Inc, 1991.

[7] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, pp. 287–314, 1994.

[8] A. Hyvärinen, "Survey on independent component analysis," *Neural Computing Surveys*, vol. 2, pp. 94–128, 1999.

[9] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, pp. 411–430, 2000.

[10] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, "The mutual information: Detecting and evaluating dependencies between variables," *Bioinformatics*, vol. 18, pp. S231–S240, 2002.

[11] D. H. Foster, I. Marín-Franch, K. Amano, and S. M. C. Nascimento, "Approaching ideal observer efficiency in using color to retrieve information from natural scenes," *Journal of the Optical Society of America A*, vol. 26, no. 11, pp. B14–B24, 2009.

[12] G. Buchsbaum and A. Gottschalk, "Trichromacy, opponent colours coding and optimum colour information transmission in the retina," *Proceedings of the Royal Society of London. Series B, Biological Sciences*, vol. 220, no. 1218, pp. 89–113, 1983.

[13] D. H. Foster, *Chromatic Function of the Cones*. Oxford: Academic Press, 2010, pp. 266–274.

[14] I. Marín-Franch, "Information-theoretic analysis of trichromatic images of natural scenes under different phases of daylight," Ph.D. dissertation, School of Electrical and Electronic Engineering, University of Manchester, Manchester M60 1QD, UK., May 2009.

[15] D. H. Foster, K. Amano, S. M. C. Nascimento, and M. J. Foster, "Frequency of metamerism in natural scenes," *Journal of the Optical Society of America A*, vol. 23, no. 10, pp. 2359–2372, 2006.

[16] D. B. Osteyee and I. J. Good, *Information, Weight of Evidence, The Singularity Between Probability Measures and Signal Detection*, ser. Lecture Notes in Mathematics, A. Dold, Heidelberg, and B. Eckmann, Eds. Springer-Verlag, 1974, vol. 376.

[17] G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1315–1321, 1999.

[18] B. Silverman, *Density Estimation for Statistics and Data Analysis*, ser. Monographs on Statistics and Applied Probability. New York: Chapman & Hall, 1986.

[19] P. Grassberger, "Entropy estimates from insufficient samplings," *arXiv:physics/0307138v2*, 2008.

[20] L. F. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," *Problems of Information Transmision*, vol. 23, no. 2, pp. 95–101, 1987, translated from Problemy Peredachi Informatsii, 23(2):9–16, 1987.

[21] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, 2004.

[22] J. D. Victor, "Binless strategies for estimation of information from neural data," *Physical Review E*, vol. 66, 2002.

[23] H. Stögbauer, A. Kraskov, S. A. Astakhov, and P. Grassberger, "Least-dependent-component analysis based on mutual information," *Physical Review E*, vol. 70, 2004.

[24] M. N. Goria, N. N. Leonenko, V. V. Mergel, and P. L. Novi Inverardi, "A new class of random vector entropy estimators and its applications in testing statistical hypotheses," *Journal of Nonparametic Statistics*, vol. 17, no. 3, pp. 277–297, 2005.

[25] I. Csiszár, "Channel capacity for a given decoding metric," *IEEE Transactions on Information Theory*, vol. 41, no. 1, pp. 35–43, 1995.

[26] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai, "On information rates for mismatched decoders," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 1953–1967, 1994.

[27] A. Lapidoth, "Nearest neighbor decoding for additive non-gaussian noise channels," *IEEE Transactions on Information Theory*, vol. 42, no. 5, pp. 1520–1529, 1996.

[28] D. H. Foster, S. M. C. Nascimento, and K. Amano, "Information limits on neural identification of colored surfaces in natural scenes," *Visual Neuroscience*, vol. 21, pp. 331–336, 2004.

[29] ——, "Information limits on identification of natural surfaces by apparent colour," *Perception*, vol. 34, pp. 1001–1006, 2005.

[30] S. M. C. Nascimento, F. P. Ferreira, and D. H. Foster, "Statistics of spatial cone-excitation ratios in natural scenes," *Journal of the Optical Society of America A*, vol. 19, no. 8, pp. 1484–1490, 2002.

[31] R. F. Lyon and P. M. Hubel, "Eyeing the camera: into the next century," in *Tenth Color Imaging Conference: Color Science and Engineering Systems, Technologies, Applications*, Scottsdale, Arizona, USA, 2002, pp. 349–355.

[32] J. M. DiCarlo, E. Montgomery, and S. W. Trovinger, "Emissive chart for imager calibration," in *Twelfth Color Imaging Conference: Color Science and Engineering Systems, Technologies, Applications*, Scottsdale, Arizona, USA, 2004, pp. 295–301.

[33] A. Stockman and L. T. Sharpe, "The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype," *Vision Research*, vol. 40, pp. 1711–1737, 2000.

[34] S. M. C. Nascimento, D. H. Foster, and K. Amano, "Psychophysical estimates of the number of spectral-reflectance basis functions needed to reproduce natural scenes," *Journal of the Optical Society of America A*, vol. 22, no. 6, pp. 1017–1022, 2005.

[35] E. K. Oxtoby and D. H. Foster, "Perceptual limits on low-dimensional models of munsell reflectance spectra," *Perception*, vol. 34, pp. 961–966, 2005.

[36] D. B. Judd, D. L. MacAdam, and G. Wyszecki, "Spectral distribution of typical daylight as a function of correlated color temperature," *Journal of the Optical Society of America*, vol. 54, no. 8, pp. 1031–1040, 1964.

[37] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Wu, "An optimal algorithm for approximate nearest neighbor searching," *Journal of the ACM*, vol. 45, no. 6, pp. 891–923, 1998.

[38] J. A. Worthey, "Limitations of color constancy," *Journal of the Optical Society of America A*, vol. 2, no. 7, pp. 1014–1026, 1985.

[39] J. M. Troost, L. Wei, and C. M. M. de Weert, "Binocular measurements of chromatic adaptation," *Vision Research*, vol. 32, no. 10, pp. 1987–1997, 1992.

[40] E. H. Land and J. J. McCann, "Lightness and retinex theory," *Journal of the Optical Society of America*, vol. 61, no. 1, pp. 1–11, 1971.

[41] E. H. Land, "Recent advances in retinex theory," *Vision Research*, vol. 26, no. 1, pp. 7–21, 1986.

[42] D. H. Foster and S. M. C. Nascimento, "Relational colour constancy from invariant cone-excitation ratios," *Proceedings of the Royal Society B*, vol. 257, pp. 115–121, 1994.

[43] B. V. Funt and G. D. Finlayson, "Color constant color indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 5, pp. 522–529, 1995.

[44] G. D. Finlayson, M. S. Drew, and B. V. Funt, "Spectral sharpening: sensor transformations for improved color constancy," *Journal of the Optical Society of America A*, vol. 11, no. 5, pp. 1553–1563, 1994.

[45] ——, "Color constancy: generalized diagonal transforms suffice," *Journal of the Optical Society of America A*, vol. 11, no. 11, pp. 3011–3019, 1994.

[46] B. Funt and H. Jiang, "Non-von-kries 3-parameter color prediction," in *Proceedings SPIE*, vol. 5007, 2003, pp. 182–189.

[47] D. H. Foster and K. Żychaluk, "Is there a better non-parametric alternative to is there a better non-parametric alternative to von kries scaling?" in *CGIV 2008, 4rd European Conference on Colour in Graphics, Imaging and Vision*, Terrassa, Spain, 2008, pp. 41–44.

[48] H. Terstiege, "Chromatic adaptation: a state-of-the-art report," *Journal of Color Appearance*, vol. 1, no. 4, pp. 19–23, 1972.

[49] G. Buchsbaum, "A spatial processor model for object color perception," *Journal of the Franklin Institute*, vol. 310, no. 1, pp. 1–26, 1980.

[50] M. D'Zmura and G. Iverson, "Color constancy. III. General linear recovery of spectral descriptions for lights and surfaces," *Journal of the Optical Society of America A*, vol. 11, no. 9, pp. 2389–2400, 1994.
[51] G. D. Finlayson, S. D. Hordley, and P. M. Hubel, "Color by correlation: A simple, unifying framework for color constancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1209–1221, 2001.
[52] D. H. Foster, "Does colour constancy exist?" *Trends in Cognitive Sciences*, vol. 7, no. 10, pp. 439–443, 2003.
[53] H. G. Sperling and R. S. Harwerth, "Red-green cone interactions in the increment-threshold spectral sensitivity of primates," *Science*, vol. 172, no. 3979, pp. 108–184, 1971.
[54] R. L. De Valois, N. P. Cottaris, S. D. Elfar, L. E. Mahon, and J. A. Wilson, "Some transformations of color information from lateral geniculate nucleus to striate cortex," *Proceedings of the National Academy of Science of the United States of America*, vol. 97, no. 9, pp. 4997–5002, 2000.
[55] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence properties of the Nelder-Mead simplex method in low dimensions," *SIAM Journal on Optimization*, vol. 9, no. 1, pp. 112–147, 1998.
[56] I. Marín-Franch and D. H. Foster, "Number of perceptually distinct surface colors in natural scenes," *Journal of Vision*, vol. 10, no. 9, pp. 1–7, 2010.
[57] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization*, ser. Wiley series in probability and mathematical statistics, J. Wiley, Ed. New York: John Wiley & Sons, Inc, 1992.
[58] R. P. W. Duin, "On the choice of the smoothing parameters for Parzen estimators of probability density functions," *IEEE Transactions on Computers*, vol. C-25, no. 11, pp. 1175–1179, 1976.

Iván Marín-Franch received his B.Sc. in statistics in 1998 and M.Sc. in statistical science and techniques in 2000 from the University of Granada. He received his Ph.D. in 2009 from the University of Manchester, where he was also a research assistant. He then held a research associateship at the University of Turin until 2010. He is currently a research fellow in the Indiana University School of Optometry and the Department of Optometry and Visual Science at City University London.

David H. Foster received his B.Sc. in physics in 1966 and his Ph.D. in 1970, both from Imperial College London. He received his D.Sc. in 1982 from London University. He was appointed lecturer at Imperial College in 1970 and subsequently has held professorships at Keele University, Aston University, UMIST, and the University of Manchester. He is a Fellow of the Institute of Physics, the Institute of Mathematics and its Applications, and the Optical Society of America.

APPENDIX A
AUTOMATIC BANDWIDTH-SELECTION METHODS

Any kernel density estimator (Section III-A) requires an estimate to be made of its bandwidth. For the estimated differential entropy $\hat{h}_{\mathrm{KD}}(A_1)$ of the random variable $A_1 = (R_1, G_1, B_1)$ with pdf f_1, two automatic methods were used, namely rule of thumb and likelihood cross-validation [18], [57]. These two were chosen for their different assumptions and different dependencies on data.

The rule-of-thumb method uses a bandwidth that is optimal for Gaussian images. The bandwidths σ_R, σ_G, and σ_B selected for R_1, G_1, and B_1 were the sample SDs of R_1, G_1, and B_1 multiplied by the constant $N^{-1/(m+4)}$, where N is the sample size, as in (10), and $m = 3$, the dimension of A_1. Bandwidths for $\hat{h}_{\mathrm{KD}}(A_2)$ and $\hat{h}_{\mathrm{KD}}(A_{12})$, where $A_{12} = (A_1, A_2)$, were selected analogously.

The likelihood cross-validation method uses a maximum-likelihood criterion to determine the adequacy of the fit [18]. In its leave-one-out form [58], the bandwidths σ_R, σ_G, and σ_B were selected to maximize the score function

$$\sum_{i=1}^{N} \log \hat{f}_1^{(-i)}\left(a_{1i}; \sigma_R, \sigma_G, \sigma_B\right) ,$$

where $\hat{f}_1^{(-i)}$ is a kernel density estimator for f_1 at that point based on the $N-1$ data points excluding a_{1i} [58]. Analogous expressions hold for f_2 and f_{12}.

APPENDIX B
ESTIMATES OF INFORMATION AVAILABLE WITH GAUSSIAN IMAGES

To test the estimators of the information available (Sections III-A to III-D), pairs of Gaussian images were created with defined variance-covariance structures. The exact information available was then calculated and compared with the estimates from the kernel density estimator $\hat{I}_{\mathrm{KD}}$, the Kozachenko-Leonenko estimator $\hat{I}_{\mathrm{KL}}$, and the Kraskov-Stögbauer-Grassberger estimator $\hat{I}_{\mathrm{KSG}}$ as a function of sample size and different bandwidth-selection and neighborhood criteria. In more detail, the procedure was as follows.

Gaussian random variables A_1 and A_2 were sampled from trivariate Gaussian distributions with constant diagonal correlation matrices, so that $\mathrm{cor}\,(R_1, R_2) = \mathrm{cor}\,(G_1, G_2) = \mathrm{cor}\,(B_1, B_2) = \rho$, and zero off-diagonal elements. The mutual information I is given exactly [6] by

$$I = -\frac{3}{2}\log\left(1-\rho^2\right) . \tag{22}$$

The correlation coefficient was assigned two values, $\rho = 0.9$ for strongly correlated images and $\rho = 0.1$ for weakly correlated images, which yielded values for I of 3.59 bits and 0.02 bits, respectively.

Figure 10 shows as a function of random sample size N estimates of the information available given by $\hat{I}_{\mathrm{KD}}$ (a and d) averaged over 10 iterations of the procedure, by $\hat{I}_{\mathrm{KL}}$ (b and e), and by $\hat{I}_{\mathrm{KSG}}$ (c and f) averaged over 100 iterations. Sample size N ranged from 2^3 to 2^{18}. Only 10 iterations were used for $\hat{I}_{\mathrm{KD}}$ because of its lengthy computation time. The value of N was limited to 2^{16} with the likelihood cross-validation method for the same reason. The upper panels a, b, and c are for strongly correlated images, $\rho = 0.9$, and the lower panels d, e, and f for weakly correlated images, $\rho = 0.1$. The true information available I is indicated by the horizontal gray lines. For clarity, SDs are not shown. They decreased as sample size N increased, and for the maximum sample size $N = 2^{18}$, they were < 0.008 bits. For $\hat{I}_{\mathrm{KD}}$ with the likelihood cross-validation method, for which the maximum sample size $N = 2^{16}$, the SDs were < 0.018 bits.

For both strongly and weakly correlated images, the convergence of $\hat{I}_{\mathrm{KL}}$ (b and e) and $\hat{I}_{\mathrm{KSG}}$ (c and f) to the true information available was faster than that of $\hat{I}_{\mathrm{KD}}$ (a and d). Moreover, information available was always underestimated by $\hat{I}_{\mathrm{KL}}$ and $\hat{I}_{\mathrm{KSG}}$, whereas with $\hat{I}_{\mathrm{KD}}$ the bias could be either downwards or upwards.

With strongly correlated images, the performance of $\hat{I}_{\mathrm{KL}}$ (b) and $\hat{I}_{\mathrm{KSG}}$ (c) was similar, but with weakly correlated images, the convergence of $\hat{I}_{\mathrm{KL}}$ (e) was slower than that of $\hat{I}_{\mathrm{KSG}}$ (f). As noted in Section VI, real images are strongly correlated under different daylight illuminants, and it is clear from panels b and c that convergence to the true value of the information available was fastest with a nearest-neighbor criterion, that is, with $k = 1$.

Although $\hat{I}_{\mathrm{KL}}$ and $\hat{I}_{\mathrm{KSG}}$ behaved similarly with strongly correlated Gaussian images, the convergence of both was more biased for all sample sizes as the correlation between the images increased further (for $\hat{I}_{\mathrm{KL}}$, compare Fig. 10 b with Fig. 11 of Appendix C). Improving the speed of convergence by the offset method, described in Appendix C, was possible only with $\hat{I}_{\mathrm{KD}}$ and $\hat{I}_{\mathrm{KL}}$, since $\hat{I}_{\mathrm{KSG}}$ estimates mutual information directly rather than from differential entropy.

APPENDIX C
OFFSET ESTIMATORS WITH GAUSSIAN IMAGES

As explained in Section III-D, in the offset method, the estimators $\hat{I}_{\mathrm{KD}}(A_1; A_2)$ and $\hat{I}_{\mathrm{KL}}(A_1; A_2)$ of the mutual information between A_1 and A_2 were each decomposed into two components. One component was the mutual information between equivalent Gaussian variables with known variance-covariance structure; the other component was an offset that was obtained by applying the estimator to variance-scaled versions of A_1, A_2, and $A_{12} = (A_1, A_2)$.

Very strongly correlated Gaussian images were generated as in Appendix B but with correlation coefficient $\rho = 0.9999$ corresponding to an information available of 18.4 bits. This particular correlation coefficient was chosen because the true information available was then of the same order as that with images of natural scenes obtained with the Foveon X3 sensor set (Fig. 5, Section VI). Figure 11 shows, as a function of the size N of the random sample, estimates of the information available given by $\hat{I}_{\mathrm{KD}}$ with the likelihood cross-validation bandwidth-selection method (dash-dotted curve) and $\hat{I}_{\mathrm{KL}}$ (dotted curve) and their corresponding offset versions $\hat{I}_{\mathrm{KDo}}$ (dashed curve) and $\hat{I}_{\mathrm{KLo}}$ (solid curve). The number of iterations over which estimates were averaged and the range of samples sizes N were the same as in Appendix B. The true

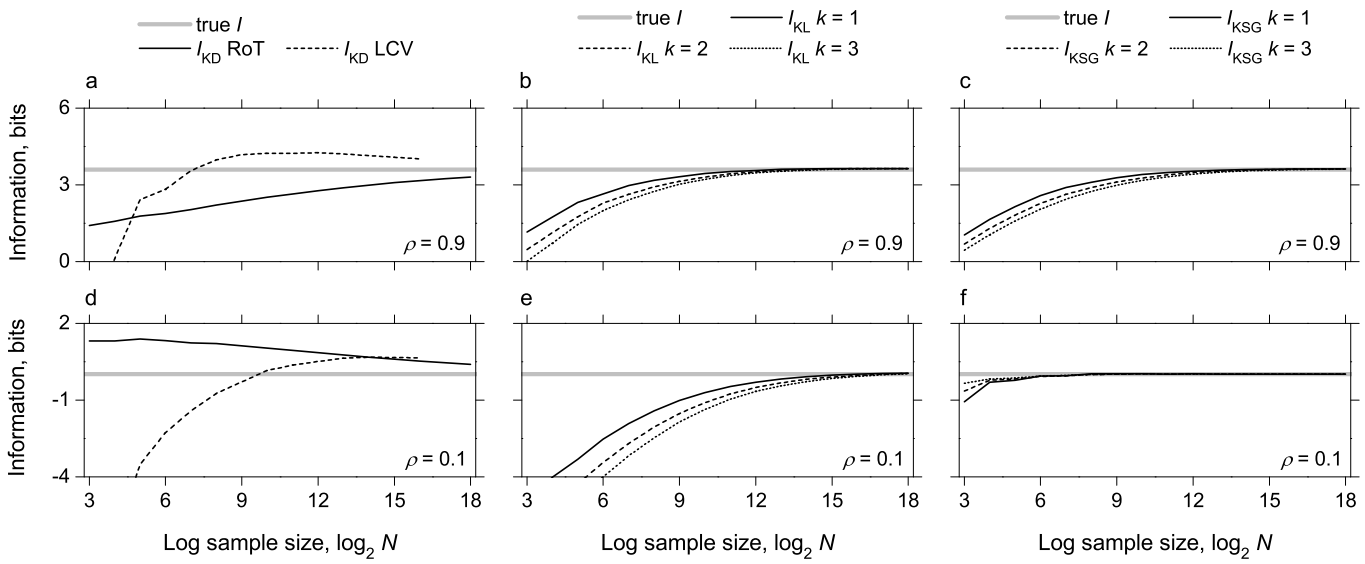

Fig. 10. Estimates of the information available from Gaussian images as a function of sample size N for (a, d) the kernel density estimator $\hat{I}_{\mathrm{KD}}$ with bandwidth selection by rule of thumb RoT (solid curves) and likelihood cross-validation LCV (dashed curves); (b and e) the Kozachenko-Leonenko estimator $\hat{I}_{\mathrm{KL}}$ with neighborhood criteria k = 1, 2, and 3 (solid, dashed, and dotted curves, respectively); and (c and f) the Kraskov-Stögbauer-Grassberger estimator $\hat{I}_{\mathrm{KSG}}$ with neighborhood criteria k = 1, 2, and 3 (solid, dashed, and dotted curves, respectively). The upper panels a, b, and c are for strongly correlated images, $\rho = 0.9$, and the lower panels d, e, and f for weakly correlated images, $\rho = 0.1$. The true information available is indicated by the horizontal gray lines. Means were taken over 10 iterations with the kernel density estimators and over 100 iterations with the Kozachenko-Leonenko and the Kraskov-Stögbauer-Grassberger estimators.

information available I is indicated by the horizontal gray line. Estimates given by $\hat{I}_{\mathrm{KD}}$ with rule-of-thumb bandwidth selection were omitted as the likelihood method always performed better. Comparison estimates from $\hat{I}_{\mathrm{KSG}}$ were also omitted as they were closely similar to those given by $\hat{I}_{\mathrm{KL}}$. Standard deviations were < 0.009 bits at the maximum sample size $N = 2^{18}$ for all except the kernel density estimator; for the latter, the SD was approximately 0.030 bits at $N = 2^{16}$.

The convergence to the information available of the estimates obtained with the offset method was evidently much faster than with the original estimators. Among all the estimators tested and over all the simulations, the offset Kozachenko-Leonenko estimator $\hat{I}_{\mathrm{KLo}}$ always converged the fastest.

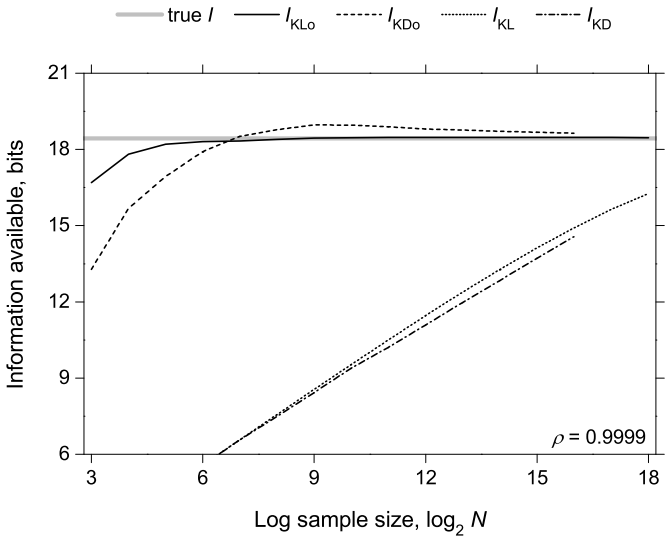

Fig. 11. Estimates of the information available from Gaussian images as a function of sample size N with the kernel density estimator $\hat{I}_{\mathrm{KD}}$ with the likelihood cross-validation bandwidth-selection method (dash-dotted curve), its offset version (dashed curve), the Kozachenko-Leonenko estimator $\hat{I}_{\mathrm{KL}}$ (dotted curve), and its offset version (solid curve). The true information available is indicated by the horizontal gray line. The images were very strongly correlated, $\rho = 0.9999$. Other details as for Fig. 10.

Appendix D
Estimates of information retrieved from Gaussian images

To test the estimators of the information retrieved (Sections IV-A and IV-B), pairs of Gaussian images were created with defined variance-covariance structures as in Appendix B. As noted in Section IX, for pairs of Gaussian images, the nearest-neighbor criterion based on the Mahalanobis distance for point matching is equivalent to the maximum-likelihood criterion and, therefore, information retrieved coincides with information available. Again as in Appendix B, the exact information available was calculated and compared with the estimates from the naïve estimator $\hat{I}_{\mathrm{NN}}$ and the Grassberger estimator $\hat{I}_{\mathrm{NNG}}$ as a function of sample size.

Figure 12 shows, as a function of the size N of the random sample, the mean estimated information retrieved with the naïve estimator $\hat{I}_{\mathrm{NN}}$ (dashed curves) and with the Grassberger estimator $\hat{I}_{\mathrm{NNG}}$ (solid curves), averaged over 100 iterations. Sample size N ranged from 2^3 to 2^{18}. Panel a is for strongly correlated images, $\rho = 0.9$, and panel b for weakly correlated images, $\rho = 0.1$. The true information retrieved I is indicated by the horizontal gray lines. The SDs decreased with N, falling to < 0.01 bits at $N = 2^{18}$. For weakly correlated images, the Grassberger estimator $\hat{I}_{\mathrm{NNG}}$ was closer to the true information retrieved. For strongly correlated images, $\hat{I}_{\mathrm{NN}}$ and

$\hat{I}_{\mathrm{NNG}}$ behaved similarly, although $\hat{I}_{\mathrm{NNG}}$ was systematically slightly lower. With a larger correlation coefficient $\rho = 0.99$ (corresponding to a true information retrieved of 8.48 bits), the difference in estimates fell to 0.02 bits. This particular correlation coefficient was chosen because the true information retrieved was then of the same order as that with images of natural scenes obtained with the Foveon X3 sensor set (see Fig. 7, Section VII). The difference in the two estimators was even smaller with still larger values of ρ.

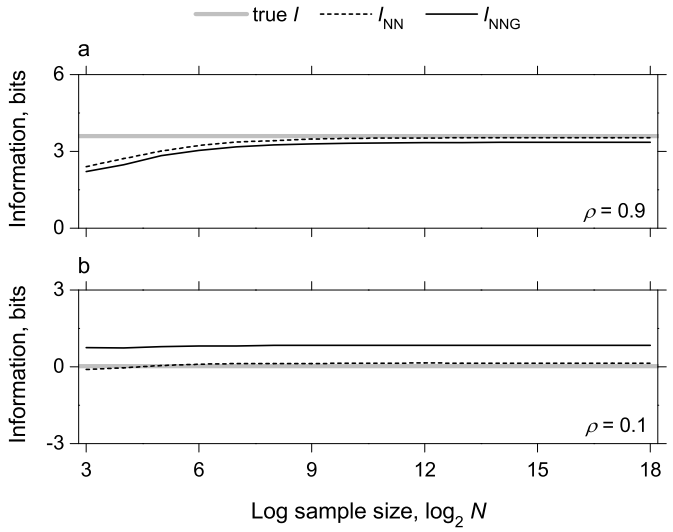

Fig. 12. Estimates of the information retrieved from Gaussian images as a function of sample size N with the naïve estimator $\hat{I}_{\mathrm{NN}}$ (dashed curves) and the Grassberger estimator $\hat{I}_{\mathrm{NNG}}$ (solid curves). Panel a is for strongly correlated images, $\rho = 0.9$, and panel b for weakly correlated images, $\rho = 0.1$. The true information retrieved, which coincides here with the information available, is indicated by the horizontal gray lines. Means were taken over 100 iterations.