QUARTERLY OF APPLIED MATHEMATICS
S 0033-569X(07)01063-2
Article electronically published on December 5, 2007

FROM INFORMATION SCALING OF NATURAL IMAGES TO REGIMES OF STATISTICAL MODELS

By

YING NIAN WU (*Department of Statistics, University of California, Los Angeles, California*),

CHENG-EN GUO (*Acuity Technologies, Menlo Park, California*),

AND

SONG-CHUN ZHU (*Departments of Statistics and Computer Science, University of California, Los Angeles, California*)

Abstract. Vision can be considered a highly specialized data collection and analysis problem. We need to understand the special properties of natural image data in order to construct statistical models and develop statistical methods for representing and recognizing the wide variety of natural image patterns. One fundamental property of natural image data that distinguishes vision from other sensory tasks such as speech recognition is that scale plays a profound role in image formation and interpretation. Specifically, visual objects can appear at a wide range of scales in the images due to the change of viewing distance as well as camera resolution. The same objects appearing at different scales produce different image data with different statistical properties. In particular, we show that the entropy rate of the image data changes over scale. Moreover, the inferential uncertainty changes over scale too. We call these changes information scaling. We then examine both empirically and theoretically two prominent and yet largely isolated classes of image models, namely, wavelet sparse coding models and Markov random field models. Our results indicate that the two classes of models are appropriate for two different entropy regimes: sparse coding targets low entropy regimes, whereas Markov random fields are appropriate for high entropy regimes. Because information scaling connects different entropy regimes, both sparse coding and Markov random fields are necessary for representing natural image data, and information scaling triggers transitions between these two regimes. This motivates us to propose a modeling scheme that embraces both regimes of models in a common framework. The contribution of our work is two-fold. First, the study of information scaling provides a unifying perspective for the rich variety of natural image patterns. Second, the modeling scheme that we develop provides a natural integration of different regimes of image models.

Received January 20, 2007.
2000 *Mathematics Subject Classification.* Primary 62M40.

1. Introduction. Computer vision can be considered to be a highly specialized data collection and analysis problem, where existing concepts and methods in statistical theory and information theory can in principle be used to model and interpret the image data [23, 39]. However, vision also proves to be a highly specialized data collection and analysis task. We must understand the special characteristics of the natural image data in order to develop realistic models and efficient algorithms for representing and recognizing the wide variety of natural image patterns.

(a) (b)

FIG. 1. Image patterns at different scales. (a) Tree leaves at different viewing distances. (b) Tree trunks, branches, and twigs of different sizes and viewing distances.

One fundamental property of natural image data that distinguishes vision from other sensory tasks such as speech recognition is that scale plays a profound role in image formation and interpretation. Specifically, visual objects can appear at a wide range of scales in the images due to the change of viewing distance as well as camera resolution. The same objects appearing at different scales produce different image data with different statistical properties. See Figure (1.a) for an example. It shows tree leaves in four different distance ranges. In region A at near distance, the leaves appear at a relatively large scale, and the individual shapes of the leaves can be recognized. In region B at intermediate distance, the image becomes more complex, and the leaves can no longer be recognized individually. Instead, only a collective foliage impression is formed from this part of the image. In region C at still farther distance, the image looks like noise. In region D at very far distance, the image appears to be a smooth region. Figure (1.b) shows another example, where tree trunks, branches and twigs appear at different scales, producing image data with different appearances. These two examples show that the change of scale causes the change of image properties, which may trigger the change of the modeling scheme for image representation.

In this paper, we study the change of statistical properties, in particular, some information-theoretical properties, of the image data over scale. We show that the entropy rate, defined as entropy per pixel, of the image data changes over scale. Moreover, the inferential uncertainty of the outside scene that generates the image data also changes with scale. We call these changes information scaling.

We then examine both empirically and theoretically two prominent and yet largely isolated research themes in image modeling and representation, namely, wavelet sparse coding [33, 42, 8] and Markov random fields [6, 20, 21]. Wavelets originated from harmonic analysis. The key principle is sparsity, where the goal is to find a dictionary of linear bases so that any typical natural image can be represented or approximated by a small number of linear bases selected from the dictionary. Markov random fields originated from statistical physics. Instead of coding the image data deterministically with a small number of linear bases, Markov random fields characterize the image data by pooling some spatial statistics over the image domain.

Our results indicate that sparse coding and Markov random fields are appropriate for two different entropy regimes: sparse coding targets low entropy regimes, whereas Markov random fields are appropriate for high entropy regimes. Because information scaling connects different entropy regimes, both classes of models are necessary for representing and interpreting image data in the whole entropy range, and information scaling triggers transitions between the two regimes of models. This motivates us to propose a modeling scheme that embraces sparse coding and Markov random fields in a common framework.

The contribution of our work is as follows. First, the change of image data over scale has been well understood in the literature [53, 31, 40]. However, the change of statistical properties of the image data over scale, i.e., information scaling, has not been thoroughly studied. Our study of information scaling provides a unifying perspective that connects the rich variety of natural image patterns of widely different statistical properties. Our work is different from previous results on the statistics of natural images [47, 17, 49, 50]. Existing results are concerned with the marginal statistics while integrating over scale. Our work, however, is concerned with the conditional statistics given the scale, especially how such conditional statistics change with the scale.

Second, the two important regimes of image models, i.e., sparse coding and Markov random fields, have largely been isolated from each other, even though both have been used extensively in image modeling and processing. Information scaling provides a unique perspective to bridge the two regimes of models, and our modeling scheme provides a natural integration of different regimes of models.

The plan of the paper is as follows. Section 2 describes an empirical study of a simple model treated by Lee, Mumford and Huang (2001) [32] to illustrate information scaling. Section 3 presents some theoretical results on information scaling. Section 4 examines wavelet sparse coding and Markov random fields. Section 5 proposes a modeling scheme that integrates different regimes of models. Section 6 concludes with a brief discussion.

2. Information scaling of the dead leaves model.

To give the reader some concrete ideas, we first study information scaling empirically by experimenting with the so-called dead leaves model.

2.1. *Model and assumptions.* The dead leaves model [37] was used by Lee, et al. [32] in their investigation of image statistics of natural scenes. The model was also previously used to model natural images [2]. For our purpose, we may consider that the model describes an ivy wall covered by a large number of leaves of similar sizes. See Figure (3)

for some examples. We assume that the leaves are of squared shape and are uniformly colored. Each leaf is represented by:

(1) Its length or width r, which follows a distribution $f(r) \propto 1/r^3$ over a finite range $[r_{\min}, r_{\max}]$.

(2) Its color or shade a, which follows a uniform distribution over $[a_{\min}, a_{\max}]$.

(3) Its position (x, y, z), where the wall serves as the (x, y)-plane, and $z \in [0, z_{\max}]$ is the distance between the leaf and the wall. We assume that $z_{\max}$ is very small, so that z matters only for deciding the occlusions among the leaves.

For the collection of leaves $\{(r_k, a_k, x_k, y_k, z_k)\}$, we assume that the r_k are independent of each other, and so are the a_k. (x_k, y_k, z_k) follow a Poisson process in $\mathbf{R}^2 \times [0, z_{\max}]$. We assume that the intensity of the Poisson process λ is large enough so that the leaves completely cover the wall. As noted by Lee et al. (2001), $\{(r_k, a_k, x_k, y_k, z_k)\}$ is a Poisson process in the joint domain $[r_{\min}, r_{\max}] \times [a_{\min}, a_{\max}] \times \mathbf{R}^2 \times [0, z_{\max}]$ with respect to the measure $f(r)dr da \lambda dx dy dz$.

Lee et al. (2001) showed that this Poisson process is scale-invariant under the assumption that $[r_{\min}, r_{\max}] \to [0, \infty]$. Specifically, under the scaling transformation $x' = x/s$ and $y' = y/s$, where s is a scaling parameter, we have $r' = r/s$, and the Poisson process will be distributed in $[r_{\min}/s, r_{\max}/s] \times [a_{\min}, a_{\max}] \times \mathbf{R}^2 \times [0, z_{\max}]$ with respect to the measure $f(sr')sdr' da \lambda s dx' s dy' dz$, which is equal to $f(r')dr' da \lambda dx' dy' dz'$ because $f(r) \propto 1/r^3$. As $[r_{\min}, r_{\max}] \to [0, \infty]$, $[r_{\min}/s, r_{\max}/s] \to [0, \infty]$ too, so the Poisson process is invariant under the scaling transformation. The assumption of Lee et al. (2000) appears to hold for most of the studies of natural image statistics [47, 49, 50, 35].

However, in our experiment, $[r_{\min}, r_{\max}]$ is assumed to be a relatively narrow range. Under the scaling transformation, this range will change to $[r_{\min}/s, r_{\max}/s]$, which is far from being invariant. From this perspective, we may consider that Lee et al. (2001) and the papers cited above are concerned with the marginal statistics by integrating over the whole range of scale. Our work, however, is concerned with the conditional statistics given a narrow range of scale, especially how such conditional statistics change under the scaling transformation. While it is important to look at the marginal statistics over the whole range of the scale, it is perhaps even more important to study the conditional statistics at different scales in order to model different image patterns. Moreover, the conditional statistics at different scales may have to be accounted for by different regimes of statistical models.

2.2. *Image formation and scaling.* Let $O_k \subset \mathbf{R}^2$ be the squared area covered by leaf k in the (x, y) domain of the ivy wall. Then the scene of the ivy wall can be represented by a function $W(x, y) = a_{k(x,y)}$, where $k(x, y) = \arg\max_{k:(x,y)\in O_k} z_k$, i.e., the most forefront leaf that covers (x, y). $W(x, y)$ is a piecewise constant function defined on $\mathbf{R}^2$.

Now let's see what happens if we take a picture of $W(x, y)$ from a distance d. Suppose the scope of the domain covered by the camera is $\Omega \subset \mathbf{R}^2$, where Ω is a finite rectangular region. As noted by Mumford and Gidas (2001) [40], a camera or a human eye only has a finite array of sensors or photoreceptors. Each sensor receives lights from a small neighborhood of Ω. As a simple model of the image formation process, we may divide the continuous domain Ω into a rectangular array of squared windows of length or width σd, where σ is decided by the resolution of the camera. Let $\{\Omega_{ij}\}$ be these squared windows,

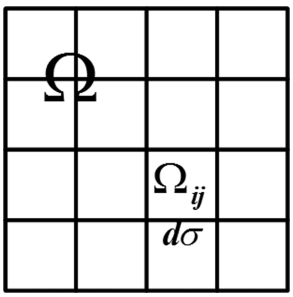


FIG. 2. Illustration of image formation. Each pixel (i, j) corresponds to a squared window Ω_{ij} in the continuous domain Ω. The size of the window is $d\sigma$, where d is the distance between the objects and the camera, and σ is determined by the camera resolution.

with $(i, j) \in D$, where D is a rectangular lattice. See Figure (2) for an illustration, where the domain is covered by 4×4 squared windows, so D in this case is 4×4.

The image $\mathbf{I}$ is defined on D. Let $s = d\sigma$ be the scale parameter of the image formation process. Then

$$\mathbf{I}_s(i, j) = \frac{1}{s^2} \int_{\Omega_{ij}} W(x, y)dxdy, \ (i, j) \in D, \tag{1}$$

which is the average of $W(x, y)$ within window Ω_{ij}. Equation (1) can also be written as

$$w_s(x, y) = \frac{1}{s^2} \int W(x', y')g((x - x')/s, (y - y')/s)dx'dy' = W * g_s, \tag{2}$$

$$\mathbf{I}_s(i, j) = w_s(u + is, v + js), \tag{3}$$

where g is a uniform density function within the window $[-1/2, 1/2] \times [-1/2, 1/2]$, and $g_s(x, y) = g(x/s, y/s)/s^2$. $(u, v) \in [0, s)^2$ denotes the small shifting of the rectangular lattice. There are two operations involved. Equation (2) is smoothing: w_s is a smoothed version of W. Equation (3) is subsampling: $\mathbf{I}_s$ is a discrete sampling of w_s. To be more general, g in Equation (2) can be any density function, for instance, a Gaussian density function. See [40] for a more general mathematical model of the image formation process.

The scale parameter s can be changed by either changing the viewing distance d or the camera resolution σ. If we increase s by increasing the viewing distance or zooming out the camera, then both the size of the scope Ω and the size of the windows Ω_{ij} will increase proportionally. So the resulting image $\mathbf{I}_s$ will change. For example, if we double s to $2s$, then $\mathbf{I}_{2s}$ will cover a scope 4 times as large as the scope of $\mathbf{I}_s$. Because each squared window of size $2s$ contains 4 squared windows of size s, if we look within the portion of $\mathbf{I}_{2s}$ that corresponds to $\mathbf{I}_s$, then the intensity of a pixel in $\mathbf{I}_{2s}$ is the block average of the intensities of the corresponding 2×2 pixels in $\mathbf{I}_s$.

If g is a Gaussian kernel, then the set of $\{w_s(x, y), s > 0\}$ forms a scale space [53, 31]. The scale space theory can account for the change of image intensities due to scaling. But it does not explain the change of statistical properties of the image data under the scaling transformation.

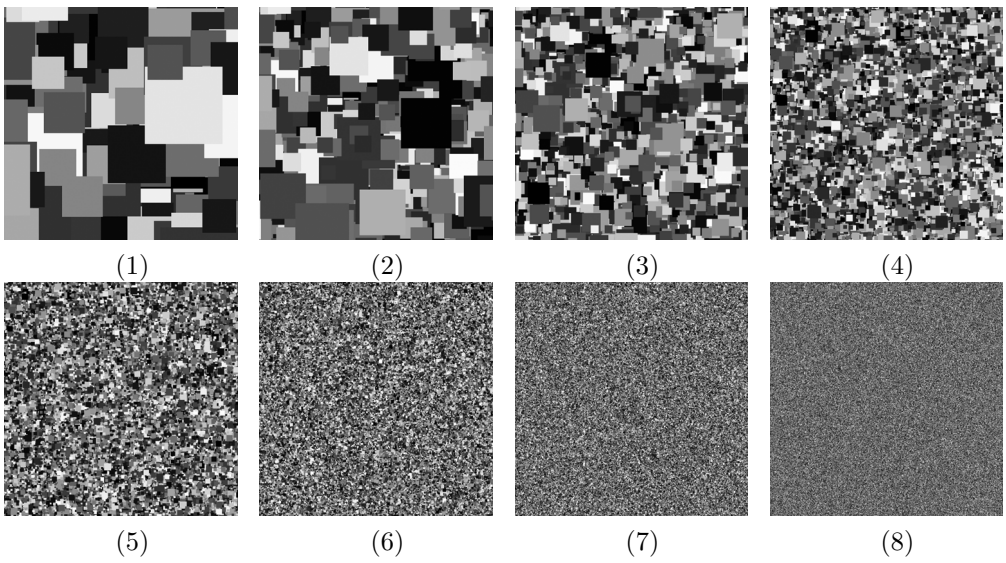


FIG. 3. Pictures of the simulated ivy wall taken at 8 viewing distances. The viewing distance of the $(i+1)$-st image is twice that of the i-th image.

2.3. *Empirical observations on information scaling.*

2.3.1. *Simulated images.* Figure (3) shows a sequence of 8 images of W taken at 8 viewing distances. The images are generated according to Equation (1). The viewing distance of the $(i+1)$-st image is twice that of the i-th image. So the viewing distance of the last image is 128 times that of the first image. Within this wide range of viewing distances, the images display markedly different statistical properties even though they are generated by the same W. The reason is that the square leaves appear at different scales in different images.

(1) For an image taken at a near distance, such as image (1), the window size of a pixel is much less than the average size of the leaves, i.e., $s \ll r$. The image can be represented deterministically by a relatively small number of occluding squares, or by local geometric structures such as edges, corners, etc. The constituent elements of the image are squares or local geometrical structures, instead of pixels.

(2) For an image at an intermediate distance, the window size of a pixel becomes comparable to the average size of the leaves, i.e., $s \approx r$. The image becomes more complex. For images (4) and (5), they can no longer be represented by a small number of geometrical structures. The basic elements have to be pixels themselves. If a simple interpretation of the image is sought, this interpretation has to be some sort of simple summary that cannot code the image intensities deterministically. The summary can be in the form of some spatial statistics of image intensities.

(3) For an image at a far distance, the window size of a pixel can be much larger than the average size of the squares, i.e., $s \gg r$. Each pixel covers a large number of leaves, and its intensity value is the average of many leaves. The image is approaching the white noise.

Computer vision algorithms always start from the analysis of local image patches, often at multiple resolutions. In Figure 4, we take some local 7×7 image patches from the images at different scales shown in Figure (3). These local image patches exhibit very different characteristics. Patches from near distance images are highly structured, corresponding to simple regular structures such as edges and corners, etc. As the distance increases, the patches become more irregular and random. So the local analysis in a computer vision system should be prepared to deal with such local image patches with different regularities and randomness.

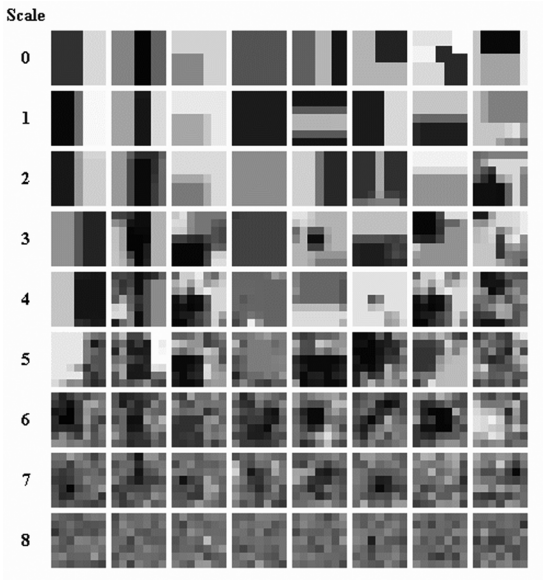
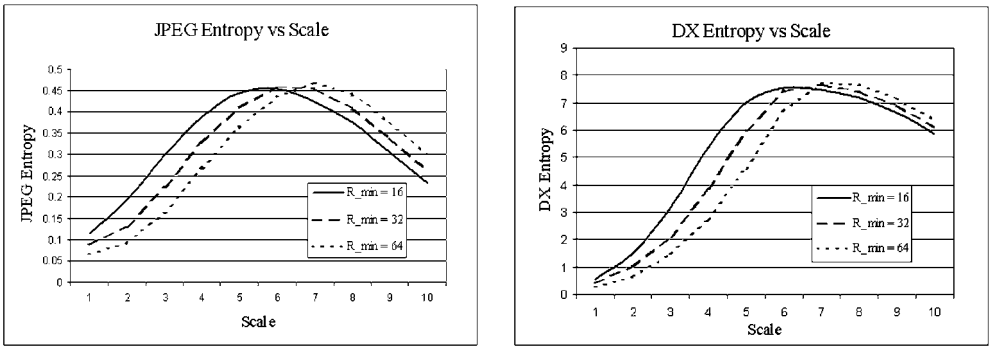


FIG. 4. The 7×7 local patches taken from the images at different scales.



(a) (b)

FIG. 5. The change of statistical properties versus the scale. (a) JPEG compression rate. (b) Entropy of marginal histogram of $\nabla_x \mathbf{I}$.

2.3.2. *Change of compression rate.* We perform some empirical studies on the change of statistical properties of the image data versus the scale. What we care about most is the complexity or randomness of the image, and we measure the complexity rate or randomness empirically by the JPEG 2000 compression rate. Generally speaking, for a simple and regular image, there are a lot of redundancies in the image intensities, so only a small number of bits are needed to store the image without any loss of information up to the discretization precision. For a complex and random image, there is no much regularity or redundancy in the data, so a large number of bits are required to store the image. The reason we use the JPEG 2000 compression rate to measure the complexity rate is two-fold. First, JPEG 2000 is the state of the art image compression standard and currently gives the best approximation to image complexity. Second, given the popularity of JPEG 2000, our results should also be interesting to the image compression community. See [15] for an in-depth treatment of data compression.

The image is compressed by JPEG 2000, and the size of the compressed image file is recorded in terms of the number of bits. This number is then divided by the number of pixels to give the compression rate in terms of bits per pixel. Figure (5.a) plots this measure in the order of viewing distance for images in Figure (3). At the near distance, the randomness is small, meaning that the image is quite regular. Then the randomness starts to increase over distance, because more and more leaves are covered by the scope of the camera. At the far distance, however, the randomness begins to decrease, because the local averaging operation reduces the marginal variance and eventually smoothes the image into a constant image because of the law of large numbers. In this plot, there are three curves. They correspond to three different $r_{\min}$ in our simulation study, while $r_{\max}$ is always fixed at the same value. For smaller $r_{\min}$, the corresponding curve shifts to the left, because the average size of the leaves is smaller.

We also use a simple measure of smoothness as an indicator of randomness or complexity rate. We compute pairwise differences between intensities of adjacent pixels $\nabla_x \mathbf{I}(i,j) = \mathbf{I}(i,j) - \mathbf{I}(i-1,j)$ and $\nabla_y \mathbf{I}(i,j) = \mathbf{I}(i,j) - \mathbf{I}(i,j-1)$. $\nabla \mathbf{I}(i,j) = (\nabla_x \mathbf{I}(i,j), \nabla_y \mathbf{I}(i,j))$ is the gradient of $\mathbf{I}$ at (i,j). The gradient is a very useful local feature that can be used for edge detection [9]. It is also extensively used in image processing. We make a marginal histogram of $\{\nabla_x \mathbf{I}(i,j), (i,j) \in D\}$ and compute the entropy of the histogram. Figure (5.b) plots this entropy over the order of distance for images in Figure (3). The plot behaves similarly as the plot of the JPEG 2000 compression rate.

We also did some experiments on natural images. Figure (6) shows a sequence of images taken at increasing distances from the trees. Figure (7) displays the change of randomness measured by three indicators versus the order of the distance. The dashed line is the JPEG compression rate. The solid line is the smoothness, i.e., the entropy of $\nabla_x \mathbf{I}$. For the black dotted line, we code the image as a linear expansion of a set of local linear bases selected from a large dictionary. We then record the number of the linear bases that need to be included in order to reduce the mean squared error to 30% of the variance of the original image. See Section 4 for more details. The three indicators are linearly normalized so that they fit into the same plot. The change of randomness in Figure (6) is consistent with the simulated example.

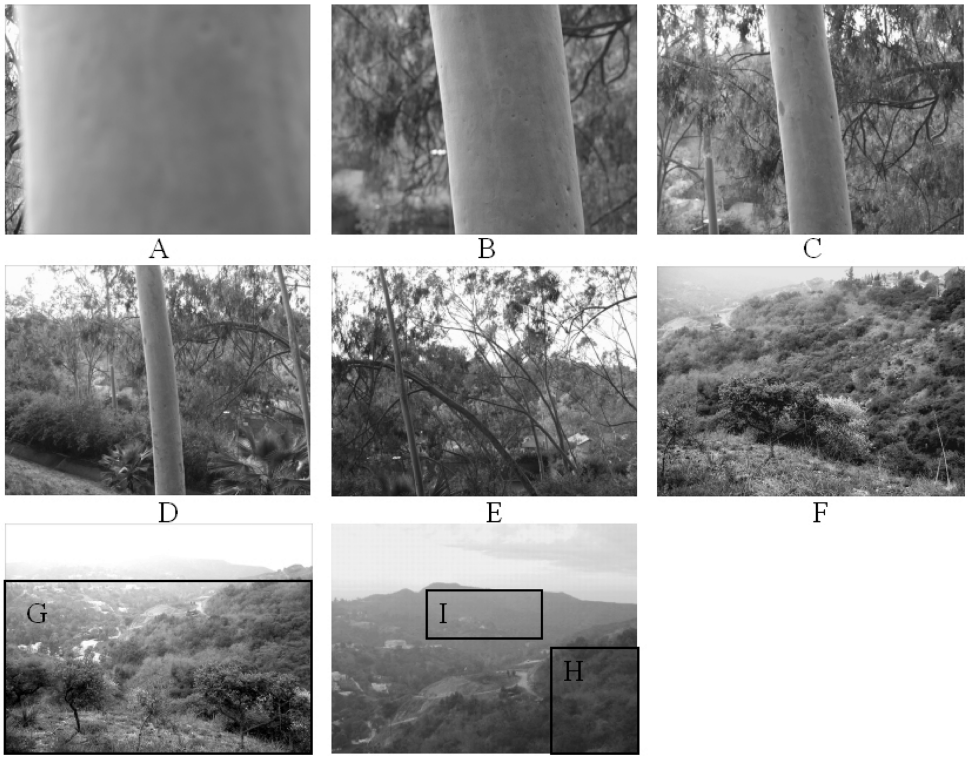


FIG. 6. Natural images taken at different distances from the trees.

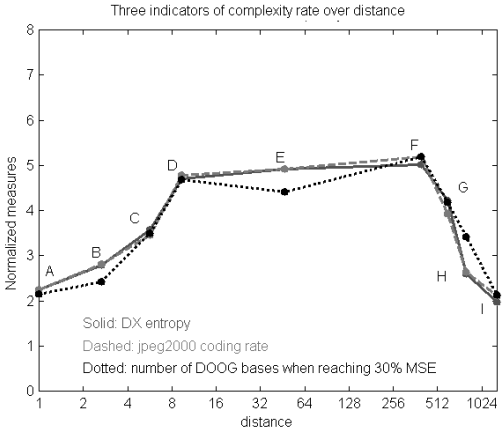
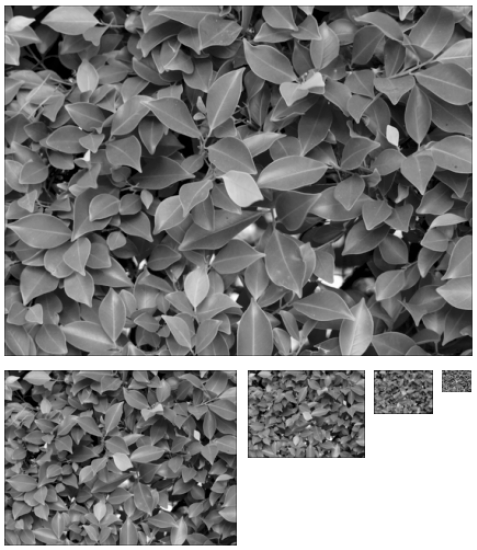
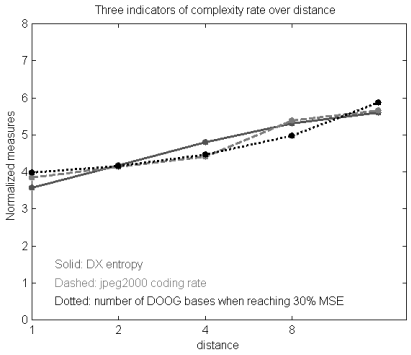


FIG. 7. The change of the randomness of the images in Figure 6 versus the approximate viewing distance.

We also did the same experiment for the pictures in Figure (8). Here we have an image of an ivy wall and its zoomed-out versions. The randomness keeps increasing as



(a)



(b)

FIG. 8. (a) The original image of an ivy wall and its zoomed-out versions. (b) The change of randomness versus the zoom-out factor or equivalently the viewing distance.

we zoom out the image, because the sequence of images does not cover the whole range of the scale.

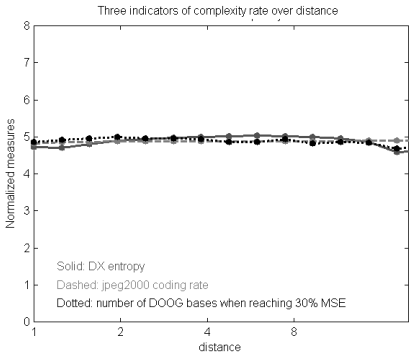


FIG. 9. A scale-invariant image and the change of randomness versus the zoom-out factor or equivalently the viewing distance.

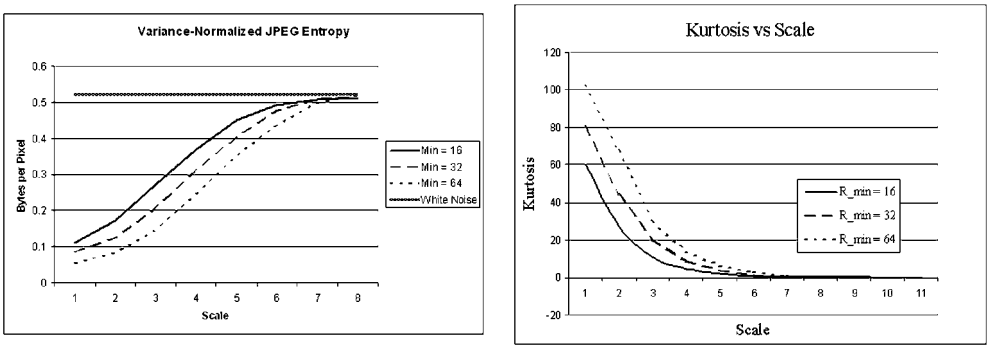
Finally, we repeat the same experiment for the picture in Figure (9) and its zoomed-out versions (not shown). The picture appears to be scale-invariant, and the randomness does not change much as we zoom out the image.

In Figure (8), $[r_{\min}, r_{\max}]$ is very small, so we see a clear change of randomness with respect to the scale. In Figure (9), however, $[r_{\min}, r_{\max}]$ is much larger, and the image is a mixture of objects and patterns of very different scales. In order to model natural

images such as the one in Figure (9), we need to model image patterns over the whole range of the scale.

2.3.3. *Variance normalization.* The local averaging operation in Equation (1) reduces the marginal variance of the image intensities. A more appropriate measure of randomness should be the compression rate of the variance-normalized image, so that this measure is invariant under linear transformations of image intensities. Specifically, for an image $\mathbf{I}$, let σ^2 be the marginal variance of $\mathbf{I}$. Let $\mathbf{I}'(i,j) = \mathbf{I}(i,j)/\sigma$. Then $\mathbf{I}'$ is the variance-normalized version of $\mathbf{I}$, and the marginal variance of $\mathbf{I}'$ is 1. We compute the JPEG compression rates of variance-normalized versions of the images in Figure (3). Figure (10.a) displays the variance-normalized JPEG compression rate versus the order of the distance for the three runs of the simulation study. The compression rate increases monotonically towards an upper bound represented by the horizontal line. This suggests that the scaling process increases the randomness and transforms a regular image to a random image. The upper bound is the JPEG compression rate of the Gaussian white noise process with variance 1.

The convergence of the compression rate of the variance-normalized image to that of the Gaussian white noise image is due to the effect of the central limit theorem. As another illustration, we compute the kurtosis of the marginal distribution of $\{\nabla_x \mathbf{I}(x), x \in D\}$. The kurtosis is decreasing monotonically towards 0, meaning that the image feature becomes closer to a Gaussian distribution (see Figure 10.b).



(a) (b)

FIG. 10. (a) The change of JPEG compression rate of the variance-normalized versions of the images in Figure (3). (b) The change of kurtosis.

3. Theoretical results on information scaling. In this section, we present some theoretical results on information scaling.

3.1. *Basic information-theoretical concepts.* Let $\mathbf{I}(x,y)$ be an image with $(x,y) \in D$, where D is the discrete lattice of pixels (in what follows, we use (x,y) instead of (i,j) to denote discrete pixels). Let $p(\mathbf{I})$ be the distribution of $\mathbf{I}$. We are interested in the following statistical properties [12].

1) *Entropy and entropy rate:* The entropy of p is defined as

$$\mathcal{H}(p) = \mathrm{E}_p[-\log p(\mathbf{I})] = -\int p(\mathbf{I}) \log p(\mathbf{I}) d\mathbf{I},$$

and the entropy rate of p is defined as $\bar{\mathcal{H}}(p) = \mathcal{H}(p)/|D|$, where $|D|$ is the number of pixels in the lattice D.

2) *Relative entropy and relative entropy rate:* For two distributions p and q, the relative entropy, or the Kullback-Leibler divergence between p and q, is defined as

$$\mathcal{K}(p||q) = \mathrm{E}_p\left[\log \frac{p(\mathbf{I})}{q(\mathbf{I})}\right] = -\mathcal{H}(p) - \mathrm{E}_p[\log q(\mathbf{I})] \geq 0.$$

The relative entropy rate is $\bar{\mathcal{K}}(p||q) = \mathcal{K}(p||q)/|D|$.

3) *Relative entropy with respect to Gaussian white noise:* For an image distribution p, let

$$\frac{1}{|D|} \sum_{(x,y)\in D} \mathrm{E}[\mathbf{I}(x,y)^2] = \sigma^2$$

be the marginal variance. Let q be the Gaussian white noise distribution with mean 0 and variance σ^2, i.e., $\mathbf{I}(x,y) \sim \mathrm{N}(0,\sigma^2)$ independently. Then

$$\mathcal{K}(p||q) = -\mathcal{H}(p) - \mathrm{E}_p[\log q(\mathbf{I})] = \mathcal{H}(q) - \mathcal{H}(p) \geq 0. \tag{4}$$

The second equation in (4) follows from $\mathrm{E}_p[\log q(\mathbf{I})] = \mathrm{E}_q[\log q(\mathbf{I})]$ because $\log q(\mathbf{I})$ is linear in $\sum_{x,y} \mathbf{I}(x,y)^2$, which has the same expectations under both p and q. Because $\mathcal{H}(q) \geq \mathcal{H}(p)$ according to (4), the Gaussian white noise distribution has the maximum entropy among all the image distributions with the same marginal variance.

4) *Entropy rate of variance-normalized image:* Continuing from (4) and calculating the entropy rate of Gaussian white noise explicitly, we obtain the relative entropy rate

$$\bar{\mathcal{K}}(p||q) = \log\sqrt{2\pi e} - [\bar{\mathcal{H}}(p(\mathbf{I})) - \log\sigma] = \log\sqrt{2\pi e} - \bar{\mathcal{H}}(p(\mathbf{I}')),$$

where $\mathbf{I}' = \mathbf{I}/\sigma$ is the variance-normalized version of the image $\mathbf{I}$, and $p(\mathbf{I}')$ denotes the distribution of $\mathbf{I}'$. So the entropy rate of the variance-normalized image $\bar{\mathcal{H}}(p(\mathbf{I}'))$ determines the relative entropy rate $\bar{\mathcal{K}}(p||q)$ of $p(\mathbf{I})$ with respect to the Gaussian white noise $q(\mathbf{I})$. In other words, $\bar{\mathcal{H}}(p(\mathbf{I}'))$ measures the departure of p from the Gaussian white noise hypothesis.

3.2. *Change of entropy rate.* For simplicity, let's study what happens if we double the viewing distance or zoom out the image by a factor of 2. Suppose the current image is $\mathbf{I}(x,y), (x,y) \in D$. If we double the viewing distance, the window covered by a pixel will double its size. So the original $\mathbf{I}$ will be reduced to a smaller image $\mathbf{I}_-$ defined on a reduced lattice D_-, and each pixel of $\mathbf{I}_-$ will be the block average of four pixels of $\mathbf{I}$. More specifically, the process can be accounted for by two steps, similar to Equations (2) and (3).

(1) *Local smoothing:* Let the smoothed image be $\mathbf{J}$. Then $\mathbf{J}(x,y) = \sum_{u,v} \mathbf{I}(x+u, y+v)/4$, where $(u,v) \in \{(0,0),(0,1),(1,0),(1,1)\}$. We can write $\mathbf{J} = \mathbf{I} * g$ where g is the uniform distribution over $\{(0,0),(0,-1),(-1,0),(-1,-1)\}$. In general, g can be any kernel with appropriate bandwidth, such as a Gaussian distribution function.

(2) *Subsampling:* $\mathbf{I}_{-}^{(u,v)}(x,y) = \mathbf{J}(2x+u, 2y+v)$, where, again, $(u,v) \in \{(0,0),(0,1),(1,0),(1,1)\}$. Any of the four $\mathbf{I}_{-}^{(u,v)}$ can be regarded as a subsampled version of $\mathbf{J}$.

THEOREM 1. Smoothing effect: Let D be an $M \times N$ lattice, and let $\mathbf{I}$ be defined on D. Let $\mathbf{J} = \mathbf{I} * g$, where g is a local averaging kernel or a probability distribution. As $\min(M,N) \to \infty$,

$$\bar{\mathcal{H}}(p(\mathbf{J})) - \bar{\mathcal{H}}(p(\mathbf{I})) \to \frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{2\pi} \log|\hat{g}(\omega)| d\omega \leq 0, \tag{5}$$

where $\omega = (\omega_x, \omega_y)$ is the spatial frequency, and $\hat{g}(\omega) = \sum_{x,y} g(x,y)\exp\{-i(\omega_x x + \omega_y y)\}$ is the Fourier transform of the kernel g, where the sum is over the support of g.

Proof. Let $\mathbf{I}$ be the image defined on the integer lattice $[0, M-1] \times [0, N-1]$. The discrete Fourier transform of $\mathbf{I}$ is

$$\hat{\mathbf{I}}(\omega) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} \mathbf{I}(x,y)\exp\{-i(\omega_x x + \omega_y y)\},$$

where $\omega_x \in \{2\pi m/M, m = 0, ..., M-1\}$ and $\omega_y \in \{2\pi n/N, n = 0, ..., N-1\}$. The Fourier transforms of $\mathbf{J}$ and g can be similarly defined. Because $\hat{\mathbf{I}}$ and $\hat{\mathbf{J}}$ are obtained from $\mathbf{I}$ and $\mathbf{J}$ respectively by the same linear transformation, $\mathcal{H}(p(\hat{\mathbf{J}})) - \mathcal{H}(p(\hat{\mathbf{I}})) = \mathcal{H}(p(\mathbf{J})) - \mathcal{H}(p(\mathbf{I}))$.

For a convolution with periodic boundary condition, $\hat{\mathbf{J}}(\omega) = \hat{\mathbf{I}}(\omega)\hat{g}(\omega)$. So

$$\begin{aligned}
\bar{\mathcal{H}}(p(\mathbf{J})) - \bar{\mathcal{H}}(p(\mathbf{I})) &= \frac{1}{|D|}\left[\mathcal{H}(p(\hat{\mathbf{J}})) - \mathcal{H}(p(\hat{\mathbf{I}}))\right] \\
&= \frac{1}{MN}\sum_{\omega}\log|\hat{g}(\omega)| = \frac{1}{4\pi^2}\sum_{\omega}\log|\hat{g}(\omega)|\Delta\omega \\
&\to \frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{2\pi}\log|\hat{g}(\omega)|d\omega,
\end{aligned}$$

as $\min(M,N) \to \infty$, where $\Delta\omega = (2\pi/M) \times (2\pi/N)$.

A smoothing kernel g is a probability distribution function, $\hat{g}$ is the characteristic function of g, and

$$\begin{aligned}
\hat{g}(\omega) &= \sum_{x,y} g(x,y)\exp\{-i(\omega_x x + \omega_y y)\} \\
&= \mathrm{E}_g\left[\exp\{-i(\omega_x X + \omega_y Y)\}\right],
\end{aligned}$$

where $(X,Y) \sim g(x,y)$. Then,

$$\begin{aligned}
|\hat{g}(\omega)|^2 &= \left|\mathrm{E}_g\left[\exp\{-i(\omega_x X + \omega_y Y)\}\right]\right|^2 \\
&\leq \mathrm{E}_g\left[\left|\exp\{-i(\omega_x X + \omega_y Y)\}\right|^2\right] = 1.
\end{aligned}$$

Thus, $\int \log|\hat{g}(\omega)|d\omega \leq 0$. □

The above theorem tells us that there is always loss of information under the smoothing operation. This is consistent with intuition in scale space theory, where the increase in scale results in the loss of fine details in the image. The change of entropy rate under linear filtering was first derived in the classical paper of Shannon (1948) [48].

Next, let's study the effect of subsampling. There are four subsampled versions $\mathbf{I}_{-}^{(u,v)}(x,y) = \mathbf{J}(2x+u, 2y+v)$, where $(u,v) \in \{(0,0),(0,1),(1,0),(1,1)\}$. Each $\mathbf{I}_{-}^{(u,v)}$

is defined on a subsampled lattice D_-, with $|D_-| = |D|/4$. See Figure (11) for an illustration.

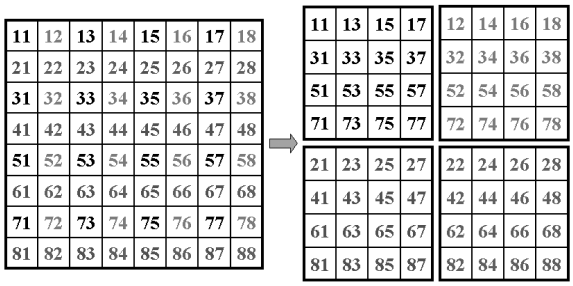


FIG. 11. The four subsampled versions of the original image.

THEOREM 2. Subsampling effect: The average entropy rate of $\mathbf{I}_-^{(u,v)}$ is no less than the entropy rate of $\mathbf{J}$,

$$\frac{1}{4}\sum_{u,v}\bar{\mathcal{H}}(p(\mathbf{I}_-^{(u,v)})) - \bar{\mathcal{H}}(p(\mathbf{J})) = \bar{\mathcal{M}}(\mathbf{I}_-^{(u,v)}, \forall(u,v)) \geq 0, \tag{6}$$

where $\mathcal{M}(\mathbf{I}_-^{(u,v)}, \forall(u,v)) = \mathcal{K}(p(\mathbf{J})||\prod_{u,v} p(\mathbf{I}_-^{(u,v)}))$ is defined as the mutual information among the four subsampled versions, and $\bar{\mathcal{M}} = \mathcal{M}/|D|$.

Proof.

$$\begin{aligned}\sum_{u,v}\mathcal{H}(p(\mathbf{I}_-^{(u,v)})) - \mathcal{H}(p(\mathbf{J})) &= \mathrm{E}\left[\log\frac{p(\mathbf{J})}{\prod_{u,v}p(\mathbf{I}_-^{(u,v)})}\right] \\ &= \mathcal{K}(p(\mathbf{J})||\prod_{u,v}p(\mathbf{I}_-^{(u,v)})) \\ &= \mathcal{M}(\mathbf{I}_-^{(u,v)}, \forall(u,v)) \geq 0,\end{aligned}$$

where the expectation is with respect to the distribution of $\mathbf{J}$, which is also the joint distribution of $\mathbf{I}_-^{(u,v)}$. □

The scaling of the entropy rate is a combination of Equations (5) and (6):

$$\left\{\frac{1}{4}\sum_{u,v}\bar{\mathcal{H}}(p(\mathbf{I}_-^{(u,v)})) - \bar{\mathcal{H}}(p(\mathbf{I}))\right\} - \left\{\bar{\mathcal{M}}(\mathbf{I}_-^{(u,v)}) + \frac{1}{4\pi^2}\int \log|\hat{g}(\omega)|d\omega\right\} \to 0. \tag{7}$$

For regular image patterns, the mutual information per pixel can be much greater than $-\int \log|\hat{g}(\omega)|d\omega/4\pi^2$, so the entropy rate increases with distance, or in other words, the image becomes more random. For very random patterns, the reverse is true. When the mutual information rate equals $-\int \log|\hat{g}(\omega)|d\omega/(4\pi^2)$, we have scale-invariance. More careful analysis is needed to determine when this is true.

Next we study the change of entropy rate of the variance-normalized image $\bar{\mathcal{H}}(p(\mathbf{I}'))$. For simplicity, let's assume that $p(\mathbf{I})$ comes from a stationary process, and $\mathbf{I}_-$ can be any subsampled version of $\mathbf{J} = \mathbf{I} * g$, which is also stationary. Let $\sigma^2 = \mathrm{Var}[\mathbf{I}(x,y)]$ and

$\sigma_-^2 = \text{Var}[\mathbf{I}_-(x,y)]$ be the marginal variances of $\mathbf{I}$ and $\mathbf{I}_-$ respectively. Let $\mathbf{I}' = \mathbf{I}/\sigma$ and $\mathbf{I}'_- = \mathbf{I}_-/\sigma_-$ be the variance-normalized versions of $\mathbf{I}$ and $\mathbf{I}_-$ respectively. It is easy to show that

$$\rho^2 = \frac{\sigma_-^2}{\sigma^2} = \frac{1}{4}\sum_{u,v} \text{corr}(\mathbf{I}(x,y), \mathbf{I}(x+u, y+v)) \leq 1, \ (u,v) \in \{(0,0),(0,1),(1,0),(1,1)\},$$

so the smoothing operation reduces the marginal variance. Therefore, we can modify (7) into

$$\bar{\mathcal{H}}(p(\mathbf{I}'_-)) - \bar{\mathcal{H}}(p(\mathbf{I}')) \approx \bar{\mathcal{M}}(\mathbf{I}_-^{(u,v)}) - \log\rho + \frac{1}{4\pi^2}\int \log|\hat{g}(\omega)|d\omega, \tag{8}$$

where the difference between the left-hand side and the right-hand size converges to 0 as $|D| \to \infty$. In (8), the term $-\log\rho$ is positive, and it compensates for the loss of entropy rate caused by smoothing, i.e., $\int \log|\hat{g}(\omega)|d\omega/(4\pi^2)$, which is negative. As a matter of fact, the first two terms, i.e., the mutual information term and the $-\log\rho$ term on the right-hand side of (8) balance each other, in the sense that if one is small, then the other tends to be large. However, we have not been able to identify conditions under which the right-hand side of (8) is always positive, which would have established the monotone increase of the entropy rate of the variance-normalized image or the monotone decrease of the departure from Gaussian white noise.

The entropy rate of the variance-normalized image is expected to eventually converge to that of Gaussian white noise, which has the maximum entropy rate among all image distributions with fixed marginal variance. The central limit theorem has been proved by Newman (1980) [41].

THEOREM 3. Newman's central limit theorem: Let $\mathbf{I}(x,y), (x,y) \in \mathbf{Z}^2$ be a stationary spatial process with $\text{E}[\mathbf{I}(x,y)] = 0$ and satisfying the following two conditions:

1) Finite susceptibility:

$$V = \sum_{(x,y)\in\mathbf{Z}^2} \text{Cov}(\mathbf{I}(0,0), \mathbf{I}(x,y)) < \infty. \tag{9}$$

2) Positive association:

$$\text{Cov}(F(\mathbf{I}(x_1,y_1),...,\mathbf{I}(x_n,y_n)), G(\mathbf{I}(x_1,y_1),...,\mathbf{I}(x_n,y_n))) \geq 0, \tag{10}$$

for any $\{(x_i,y_i), i = 1,...,n\}$, where F and G are coordinate-wise increasing functions.

For a positive integer $s \in \mathbf{Z}$, and $(x,y) \in \mathbf{Z}^2$, define the block window

$$\Omega_{s,x,y} = \{(x',y') : x' \in \{xs,...,(x+1)s-1\}, y' \in \{ys,...,(y+1)s-1\}\},$$

and let

$$\mathbf{I}_s(x,y) = \sum_{(x',y')\in\Omega_{s,x,y}} \mathbf{I}(x',y')/s. \tag{11}$$

Then as $s \to \infty$, $\mathbf{I}_s$ converges weakly to Gaussian white noise with mean 0 and marginal variance V.

Natural scenes consist of objects with occluding surfaces of smooth colors, and the colors of different objects are more or less independent. Therefore, pixels that sample the

same object tend to have similar intensities, whereas the intensities of pixels that sample different objects tend to be independent. Therefore, the positive association condition (10) is reasonable for natural images. For the finite susceptibility condition (9) to be true, we need to require that the sizes of objects have an upper bound, so that there is no long range dependence between pixel intensities.

In Newman's theorem, the size of the block $\Omega_{s,x,y}$ is s^2, but we divide the block sum by s instead of s^2 in (11), so that the marginal variance of $\mathbf{I}_s(x, y)$ converges to the constant V. That is, dividing the block sum by s instead of s^2 amounts to asymptotic variance-normalization.

The convergence of the entropy of $\mathbf{I}_s(x, y)$ to that of a Gaussian distribution has been established by [28]. But the monotone convergence has only been established for the iid case by [4].

The change of the entropy rate of the image data versus the scale can be used to explain the transition from a deterministic interpretation to a statistical interpretation of the image intensities. We only need to assume a bound on the complexity of the allowable interpretation. If a local image patch has a low entropy rate, we can code this pattern with a small number of variables deterministically. But if the local image patch has a high entropy rate, a small number of variables will not be able to account for the image intensities deterministically, and we have to interpret the image pattern statistically, by leaving the unaccounted complexity to randomness.

3.3. *Change of inferential uncertainty.* The above analysis on the entropy rate is only about the observed image $\mathbf{I}$ alone. The goal of computer vision is to interpret the observed image in order to recognize the objects in the outside world. In this subsection, we shall go beyond the statistical properties of the observed image itself and study the interaction between the observed image and the outside scene that produces the image.

Again, we would like to use the dead leaves model in Section 2 to convey the basic idea. Suppose our attention is restricted to a finite scope $\Omega \subset \mathbf{R}^2$, and let $W = \{(x_i, y_i, r_i, a_i), i = 1, ..., N\}$ be the leaves in Ω that are not completely occluded by other leaves. Then we have $W \sim p(W)$ and $\mathbf{I} = \gamma(W)$, where $p(W)$ comes from the Poisson process that generates the dead leaves, and γ represents the transformation defined by Equation (1) for a scale parameter s.

For convenience, assume that both W and $\mathbf{I}$ are properly discretized. For any joint distribution $p(W, \mathbf{I})$, the conditional entropy $\mathcal{H}(p(W \mid \mathbf{I}))$ is defined as

$$\mathcal{H}(p(W \mid \mathbf{I})) = -\sum_{W,\mathbf{I}} p(W, \mathbf{I}) \log p(W \mid \mathbf{I}). \tag{12}$$

$\mathcal{H}(p(W \mid \mathbf{I}))$ measures the inferential uncertainty or imperceptibility of W from the image $\mathbf{I}$.

PROPOSITION 1. If $W \sim p(W)$ and $\mathbf{I} = \gamma(W)$, then $\mathcal{H}(p(W|\mathbf{I})) = \mathcal{H}(p(W)) - \mathcal{H}(p(\mathbf{I}))$. That is, imperceptibility = scene entropy - image entropy.

This proposition is easy to prove. The marginal distribution of $\mathbf{I}$ is $p(\mathbf{I}) = \sum_{W:\gamma(W)=\mathbf{I}} p(W)$. The posterior distribution of W given $\mathbf{I}$ is $p(W|\mathbf{I}) = p(W, \mathbf{I})/p(\mathbf{I}) = p(W)/p(\mathbf{I})$. Here $p(W, \mathbf{I}) = p(W)$ because $\mathbf{I}$ is determined by W. Following the definition

in (12), $\mathcal{H}(p(W \mid \mathbf{I})) = -\sum_W p(W)(\log p(W) - \log p(\mathbf{I})) = \mathcal{H}(p(W)) - \mathcal{H}(p(\mathbf{I}))$. Here $\mathrm{E}_W[\log p(\mathbf{I})] = \mathrm{E}_{\mathbf{I}}[\log p(\mathbf{I})]$ since $\mathbf{I}$ is determined by W.

If we increase the viewing distance or equivalently zoom out the camera while fixing the scope $\Omega \subset \mathbf{R}^2$, i.e., fixing W, then we obtain a zoomed-out version $\mathbf{I}_- = R(\mathbf{I})$, where R represents the zooming-out operation of smoothing and subsampling, and is a many-to-one transformation. During the process of zooming out, the total entropy of the image will decrease, i.e., $\mathcal{H}(p(\mathbf{I}_-)) \leq \mathcal{H}(p(\mathbf{I}))$, even though the entropy per pixel can increase as we have shown in the previous subsection. Therefore, we have the following result.

PROPOSITION 2. If $W \sim p(W)$, $\mathbf{I} = \gamma(W)$, and $\mathbf{I}_- = R(\mathbf{I})$, where R is a many-to-one mapping, then $\mathcal{H}(p(W|\mathbf{I}_-)) \geq \mathcal{H}(p(W|\mathbf{I}))$, i.e., the imperceptibility increases as the image is reduced.

What does this result tell us in terms of interpreting the image $\mathbf{I}$ or $\mathbf{I}_-$? Although the model $W \sim p(W)$ and $\mathbf{I} = \gamma(W)$ is the right physical model for all scales s, this model is meaningful in interpreting $\mathbf{I}$ only within a limited range, say $s \leq s_{\text{bound}}$, so that the imperceptibility $\mathcal{H}(p(W \mid \mathbf{I}))$ is below a small threshold. In this regime, the representation $\mathbf{I} = \gamma(W)$ is good for both recognition and coding. For recognition, $\mathcal{H}(p(W \mid \mathbf{I}))$ is small, so W can be accurately determined from $\mathbf{I}$. For coding, we can first code W according to $p(W)$, with a coding cost $\mathcal{H}(p(W))$. Then we code $\mathbf{I}$ using $\mathbf{I} = \gamma(W)$ without any coding cost. The total coding cost would be just $\mathcal{H}(p(W))$. If the imperceptibility $\mathcal{H}(p(W \mid \mathbf{I}))$ is small, $\mathcal{H}(p(W)) \approx \mathcal{H}(p(\mathbf{I}))$, so coding W will not incur coding overhead.

But if s is very large, the imperceptibility $\mathcal{H}(p(W \mid \mathbf{I}))$ can be large according to Proposition 2. In this case, the representation $\mathbf{I} = \gamma(W)$ is not good for either recognition or coding. For recognition, W cannot be estimated with much certainty. For coding, if we still code W first, and code $\mathbf{I}$ by $\mathbf{I} = \gamma(W)$, this will not be an efficient coding, since $\mathcal{H}(p(W))$ can be much larger than $\mathcal{H}(p(\mathbf{I}))$, and the difference is the imperceptibility $\mathcal{H}(p(W \mid \mathbf{I}))$.

Then what should we do? The regime of $s > s_{\text{bound}}$ is quite puzzling for vision modeling. Our knowledge about geometry, optics, and mechanics enables us to model every phenomenon in our physical environment. Such models may be sufficient for computer graphics as far as generating physically realistic images is concerned. For instance, a garden scene can be constructed by simulating billions of leaves and grass strands, and the image can be produced by projecting these billions of objects onto the image with perspective geometry. A river scene, a fire scene or a smoke scene can be obtained using computational fluid dynamics. A piece of cloth can be generated using a dense set of particles that follow the laws of mechanics. Realistic lighting can be simulated by ray tracing and optics. But such models are hardly meaningful for vision, because the imperceptibilities of the underlying elements or variables are intolerable. When we look at a garden scene, we never really perceive every leaf or every strand of grass. When we look at a river scene, we do not perceive the constituent elements used in fluid dynamics. When we look at a scene with sophisticated lighting and reflection, we do not trace back the light rays. In those situations where physical variables are not perceptible due to

scaling or other aspects of the image formation process, it is quite a challenge to come up with good models for the observed images. Such models do not have to be physically realistic, but they should generate visually realistic images, so that such models can be employed to interpret the observed image at a level of sophistication that is comparable to human vision.

The following are some of our simple theoretical considerations of this problem from the perspectives of recognition and coding. We shall become more concrete on the modeling issue in subsequent sections.

Suppose the image $\mathbf{I}$ is reduced to an image $\mathbf{I}_- = R(\mathbf{I})$, so that W cannot be reliably inferred. Then, instead of pursuing a detailed description W from $\mathbf{I}_-$, we may choose to estimate some aspects of W from $\mathbf{I}_-$. For instance, in the simulated ivy wall example, we may estimate properties of the overall distribution of colors of leaves, as well as the overall distribution of their sizes, etc. Let's call it $W_- = \rho(W)$, with ρ being a many-to-one reduction function. It is possible that we can estimate W_- from $\mathbf{I}_-$ because of the following result.

PROPOSITION 3. Letting $W \sim p(W)$, $\mathbf{I} = \gamma(W)$, and $W_- = \rho(W)$, $\mathbf{I}_- = R(\mathbf{I})$, where both ρ and R are many-to-one mappings, we have

$$(1)\ \ \mathcal{H}(p(W_-|\mathbf{I}_-)) \le \mathcal{H}(p(W|\mathbf{I}_-)),$$

$$(2)\ \ p(\mathbf{I}_-|W_-) = \frac{\sum_{W:\rho(W)=W_-;R(\gamma(W))=\mathbf{I}_-} p(W)}{\sum_{W:\rho(W)=W_-} p(W)}.$$

Result (1) tells us that even if W is imperceptible from $\mathbf{I}_-$, W_- may still be perceptible. Result (2) tells us that although W defines $\mathbf{I}$ deterministically via $\mathbf{I} = \gamma(W)$, W_- may only define $\mathbf{I}_-$ statistically via a probability distribution $p(\mathbf{I}_-|W_-)$. While W represents deterministic structures, W_- may only represent some texture properties. Thus, we have a transition from a deterministic representation of the image intensities $\mathbf{I} = \gamma(W)$ to a statistical characterization $\mathbf{I}_- \sim p(\mathbf{I}_-|W_-)$. See Figure (12) for an illustration.

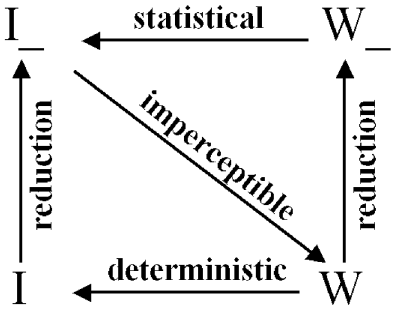


FIG. 12. Transition from deterministic representation to statistical description.

For an image $\mathbf{I}$, we may extract $F(\mathbf{I})$, which can be a dimension reduction or a statistical summary, so that $F(\mathbf{I})$ contains as much information about $\mathbf{I}$ as possible as far as W or W_- is concerned. In the following proposition, we shall not distinguish between $(W, \mathbf{I})$ and $(W_-, \mathbf{I}_-)$ for notational uniformity.

PROPOSITION 4. Let $F = F(\mathbf{I})$.

(1) If $W \sim p(W)$, $\mathbf{I} = \gamma(W)$, then $\mathcal{K}(p(W|\mathbf{I})||p(W|F)) = \mathcal{H}(p(\mathbf{I}|F))$.

(2) If $W \sim p(W)$ and $[\mathbf{I}|W] \sim p(\mathbf{I}|W)$, then $\mathcal{K}(p(W|\mathbf{I})||p(W|F)) = \mathcal{M}(W, \mathbf{I}|F)$, where $\mathcal{M}(W, \mathbf{I}|F) = \mathrm{E}_{W,\mathbf{I}}\{\log[p(W, \mathbf{I}|F)/(p(W|F)p(\mathbf{I}|F))]\}$ is the mutual information between W and $\mathbf{I}$ given F.

Result (1) tells us that for $F(\mathbf{I})$ to contain as much information about W as possible, we want to make $\mathcal{H}(p(\mathbf{I}|F))$ be as small as possible, so that F can be used to reconstruct $\mathbf{I}$ accurately. Result (2) tells us that if we want to estimate W, we want F to be sufficient about $\mathbf{I}$ as far as W is concerned. $\mathcal{M}(W, \mathbf{I}|F)$ can be considered to be a measure of sufficiency.

Now let's study this issue from the coding perspective. Suppose the image $\mathbf{I}$ follows a true distribution $f(\mathbf{I})$, and we use a model $w \sim p(w)$, and $[\mathbf{I} \mid w] \sim p(\mathbf{I} \mid w)$ to code $\mathbf{I} \sim f(\mathbf{I})$. Here the variable w is augmented solely for the purpose of coding. It might be some $w = W_- = \rho(W)$, or it may not have any correspondence to the reality W. In the coding scheme, for an image $\mathbf{I}$, we first estimate w by a sample from the posterior distribution $p(w|\mathbf{I})$, then we code w by $p(w)$ with coding length $-\log p(w)$. After that, we code $\mathbf{I}$ by $p(\mathbf{I}|w)$ with coding length $-\log p(\mathbf{I}|w)$. So the average coding length is $-\mathrm{E}_f\left[\mathrm{E}_{p(w|\mathbf{I})}(\log p(w) + \log p(\mathbf{I}|w))\right]$.

PROPOSITION 5. The average coding length is $\mathrm{E}_f[\mathcal{H}(p(w|\mathbf{I}))] + \mathcal{K}(f(\mathbf{I})||p(\mathbf{I})) + \mathcal{H}(f)$, where $p(\mathbf{I}) = \sum_w p(w)p(\mathbf{I} \mid w)$ is the marginal distribution of $\mathbf{I}$ under the model. So, coding redundancy = imperceptibility + model bias.

The above proposition provides a selection criterion for models with latent variables. The imperceptibility term comes up because we assume a coding scheme where w must be coded first, and then $\mathbf{I}$ is coded based on w. Given the latent variable structure of the model, it is very natural to assume such a coding scheme.

4. Wavelet sparse coding and Markov random fields. In this section, we shall examine two concrete classes of image models and analyze their entropy behaviors. Before doing that, we shall briefly describe the Gabor wavelets, which are mathematical models of simple neuron cells in the primary visual cortex. The Gabor wavelets play an important role in both types of models.

4.1. *Gabor wavelets.* Huber and Wiesel (1962) [27] discovered that simple neuron cells in the primary visual cortex (or what is called the V1 area) selectively respond to visual stimuli such as bars and edges at different locations, scales, and orientations. Daugman (1980) [13] proposed a mathematical model for the response properties of these simple cells using Gabor wavelets. These wavelets are translated, dilated and rotated versions of the following function:

$$G(x, y) \propto \frac{1}{\sigma_x \sigma_y} \exp\{-\frac{x^2}{2\sigma_x^2}\} e^{i\omega x}, \tag{13}$$

which is a pair of local sine and cosine waves propagating along the x-axis, where the localization is achieved by multiplying the sine waves by a Gaussian function. σ_y is

larger than σ_x, so $G(x, y)$ is elongated along the y-axis. ω and σ_x are chosen so that the amplitude of the sine or cosine wave decays to 0 very quickly.

Another model [56] comes from the derivatives of the Gaussian function,

$$G(x, y) \propto \frac{\partial^k}{\partial x^k} \frac{1}{\sigma_x \sigma_y} \exp\{-\frac{x^2}{2\sigma_x^2}\}, \tag{14}$$

where $k = 1$ and 2, i.e., the first and second derivatives of an elongate Gaussian function. The function (14) is similar to the Gabor function in (13); in particular, the first derivative in (14) is similar to the Gabor sine component, and the second derivative is similar to the Gabor cosine component.

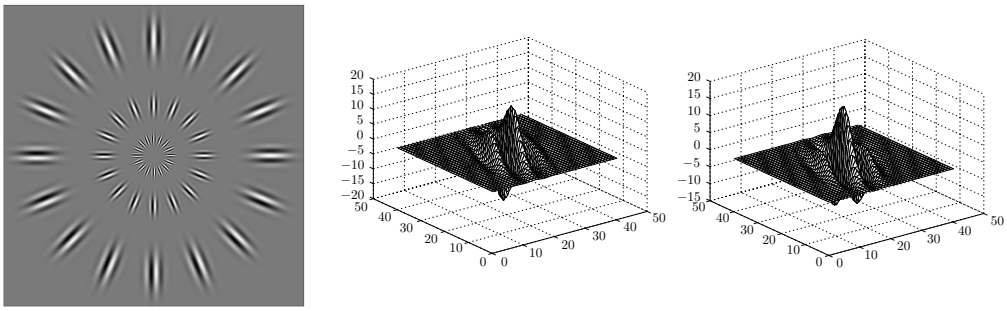


FIG. 13. (a) A sample of Gabor wavelets at different locations, scales, and orientations. (b) An example of a Gabor sine wavelet. (c) An example of a Gabor cosine wavelet.

We can dilate and rotate the Gabor function with scale s and orientation α,

$$G_{s,\alpha}(x, y) \propto G\left([x \cos \alpha + y \sin \alpha]/s, [-x \sin \alpha + y \cos \alpha]/s\right). \tag{15}$$

We can then translate $G_{s,\alpha}$ to make it centered at (x, y),

$$B_{x,y,s,\alpha}(x', y') = G_{s,\alpha}(x' - x, y' - y). \tag{16}$$

See Figure (13) for an illustration of a set of $B_{x,y,s,\alpha}$ at different locations (x, y), scales s and orientations α [29].

For an image $\mathbf{I}(x, y)$, one can define the inner product or filter response as

$$r_{x,y,s,\alpha} = \langle \mathbf{I}, B_{x,y,s,\alpha} \rangle = \sum_{x',y'} \mathbf{I}(x', y') B_{x,y,s,\alpha}(x', y').$$

$(r_{x,y,s,\alpha},\ (x, y) \in D)$ is said to be the filtered image where the filter is indexed by scale and orientation (s, α). $\langle \mathbf{I}, B_{x,y,s,\alpha} \rangle$ has a large magnitude if $B_{x,y,s,\alpha}$ lies on an edge or a bar.

We can sample (x, y, s, α) to form a large but finite dictionary $\{B_{x,y,s,\alpha}\}$. The dictionary can be overcomplete in the sense that the number of bases in $\{B_{x,y,s,\alpha}\}$ is larger than the dimensionality of the image $\mathbf{I}$.

4.2. *Sparse coding.* Field and Olshausen (1996) [42] proposed an elegant explanation for Gabor wavelets. The principle they adopted is the sparsity principle. The question they asked was: for the ensemble of natural images, can we find a dictionary of linear bases, so that for every image in that ensemble, we can almost always find a small number of linear bases from this dictionary to represent this image?

Field and Olshausen (1996) collected a sample of natural image patches (of size 12×12), $\mathbf{I}_m, m = 1, ..., M$. Then they estimated image bases $\{B_i, i = 1, ..., N\}$ (which are also images of 12×12, with $N > 12 \times 12$, i.e., the dictionary is overcomplete) by minimizing

$$\sum_{m=1}^{M} \left[\|\mathbf{I}_m - \sum_{i=1}^{N} c_{m,i} B_i\|^2 + \lambda \sum_{i=1}^{N} \delta(c_{m,i}) \right], \tag{17}$$

over all possible $\{B_i\}$, where $\delta()$ is a measure of sparsity, and λ is a tuning constant. In the objective function (17), the first term requires that the linear expansion $\sum_i c_{m,i} B_i$ should be close to the observed image $\mathbf{I}_m$. The second term requires that only a small number of $c_{m,i}$ are significantly different from 0. The simplest measure of sparsity is to count the number of nonzero $\{c_{m,i}\}$, i.e., $\delta(c) = 1$ if $c \neq 0$, and $\delta(c) = 0$ if $c = 0$. But this measure is not differentiable, making it hard for optimization. For computational convenience, it can be replaced by some measure such as the l_p-norm of the sequence $\{c_{m,i}, i = 1, ..., N\}$, with $0 < p \leq 1$. Using a gradient algorithm, Field and Olshausen (1996) were able to learn localized, scaled, and oriented base functions very similar to the Gabor wavelets shown in Figure (13). That is, each learned that B_i can be approximated by a $B_{x,y,s,\alpha}$ defined by (15) and (16), where, again, (x, y) is the center, s is the scale, and α is the orientation.

This problem can be formulated in terms of a statistical model [30, 43]:

$$c_i \sim p(c) \text{ independently}, \tag{18}$$

$$\mathbf{I} = \sum_{i=1}^{N} c_i B_i + \epsilon, \tag{19}$$

where $p(c)$ is assumed to be a heavy-tailed distribution. The model used by [43] for $p(c)$ is a mixture of two Gaussian distributions $\rho \mathrm{N}(0, \sigma_1^2) + (1 - \rho)\mathrm{N}(0, \sigma_0^2)$. The two mixture components represent two states of the coefficients. One is the active state, with probability ρ, which is very small, and the variance σ_1^2 is very large. The other state is the inactive state, with probability $1 - \rho$, which is very large, and the variance σ_0^2 is very small or even 0 [45].

The independence assumption in (18) is only for convenience. In general, one can write the wavelet sparse coding model in the following form:

$$C = \{c_i\} \sim p(C), \tag{20}$$

$$\mathbf{I} = \sum_{i=1}^{N} c_i B_i + \epsilon, \tag{21}$$

where $C = \{c_i\}$ are coefficients, and ϵ is assume to be Gaussian white noise. We can rewrite the model (20) and (21) in the matrix form $C \sim p(C)$, $\mathbf{J} = BC$, and $\mathbf{I} = \mathbf{J} + \epsilon$,

where $\mathbf{I}$ and $\mathbf{J}$ become vectors, B is the matrix whose column vectors are the bases B_i, and C is the vector consisting of all the c_i.

The uncertainty caused by the overcompleteness can easily be seen via the singular value decomposition of $B = U(\Lambda, 0)(V_1, V_0)'$. B is a $|D| \times N$ matrix, where $|D|$ is the size of the image lattice D, and N is the total number of bases. Because of overcompleteness, $|D| < N$. U is a $|D| \times |D|$ orthogonal matrix. Λ is a $|D|$-dimensional diagonal matrix of singular values. $V = (V_1, V_0)$ is the $N \times N$ orthogonal matrix, where V_1 is $N \times |D|$, and V_0 is $N \times (N - |D|)$. Let $\tilde{C} = (\tilde{C}_1 = V_1'C, \tilde{C}_0 = V_0'C)$. Then $\mathbf{J} = BC = U\Lambda\tilde{C}_1$. That is, only $\tilde{C}_1$ can be solved from $\mathbf{J}$, while $\tilde{C}_0$ cannot be determined.

For an analysis of entropy,

$$\mathcal{H}(p(C)) = \mathcal{H}(p(\tilde{C})) = \mathcal{H}(p(\tilde{C}_1)) + \mathcal{H}(p(\tilde{C}_0|\tilde{C}_1)),$$
$$\mathcal{H}(p(\mathbf{J})) = \log|\det(\Lambda)| + \mathcal{H}(p(\tilde{C}_1)) = \frac{1}{2}\log|\det(BB')| + \mathcal{H}(p(\tilde{C}_1)).$$

PROPOSITION 6. In the above notation,

$$\mathcal{H}(p(\mathbf{J})) = \mathcal{H}(p(C)) - \mathcal{H}(p(\tilde{C}_0|\mathbf{J})) + \frac{1}{2}\log|\det(BB')|.$$

Low entropy regime: If $p(C)$ is very sparse, for instance, the parameter ρ in the mixture model $\rho\mathrm{N}(0, \sigma_1^2) + (1-\rho)\mathrm{N}(0, \sigma_0^2)$ is very small, then $\mathcal{H}(p(C))$ is small; thus $\mathcal{H}(p(\mathbf{J}))$ is also small. So the sparse coding model targets the low entropy component $\mathbf{J}$ of an image $\mathbf{I}$.

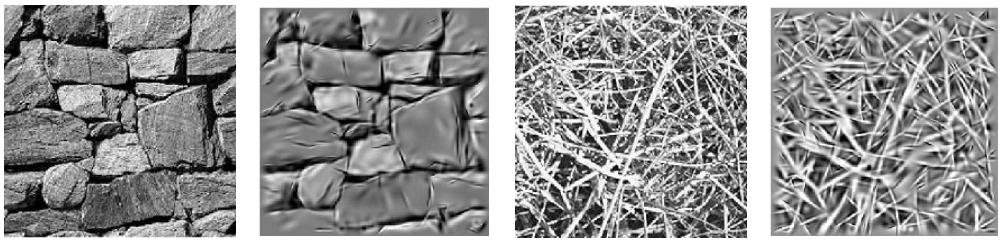
If the image $\mathbf{I}$ comes from a high entropy distribution such as random texture, the sparse coding model may not be able to account for the high entropy by the signal part $\mathbf{J}$. As a result, all the remaining entropy will be absorbed by the white noise ϵ, but the white noise model cannot capture texture information. If we force ϵ to be close to 0, then the representation will not be sparse any more.

Suppose we are given a dictionary of Gabor wavelet bases $\{B_{x,y,s,\alpha}\}$. Then we may write the model (20) and (21) in a more geometrically explicit form:

$$S = \{(x_j, y_j, s_j, \alpha_j), j = 1, ..., n\} \sim p(S), \tag{22}$$
$$(c_1, ..., c_n) \sim p(c_1, ..., c_n), \tag{23}$$
$$\mathbf{I} = \sum_{j=1}^{n} c_j B_{x_j, y_j, s_j, \alpha_j} + \epsilon, \tag{24}$$

where $S = \{(x_j, y_j, s_j, \alpha_j), j = 1, ..., n\}$ is the set of n bases selected from the dictionary to code $\mathbf{I}$, where n is a small number. Compared to the general form (20) and (21), model (22)–(24) captures the sparsity of $p(C)$ in (20) explicitly by the selection of the small set S of bases, while the small coefficients of the inactive bases are ignored. $S = \{(x_j, y_j, s_j, \alpha_j), j = 1, ..., n\}$ forms a sketch of image $\mathbf{I}$. For different images, different S will be selected, and the size of S, i.e., n, can vary. Under the independence assumption of model (18) and (19), S follows a Poisson process that penalizes the number of bases n. In general, S should be modeled by a more sophisticated spatial point process $p(S)$ that describes the geometric pattern formed by the selected bases [54, 57].

Mallat and Zhang (1993) [34] proposed a greedy algorithm called a matching pursuit algorithm for finding a sparse representation $\mathbf{I} = \sum_1^n c_j B_{x_j,y_j,s_j,\alpha_j} + \epsilon$, while not assuming any sophisticated model $p(S)$ beyond sparsity. The algorithm starts from an empty set of bases. Each time, it selects a base that leads to the largest reduction in the l_2-norm of error. The algorithm stops when the error is smaller than a threshold. This algorithm fits the model (18) and (19) approximately. Wu, Zhu, and Guo (2002) [54] proposed a Markov chain Monte Carlo version of the matching pursuit algorithm that rigorously samples from the posterior distribution of model (18) and (19).



(a) (b) (c) (d)

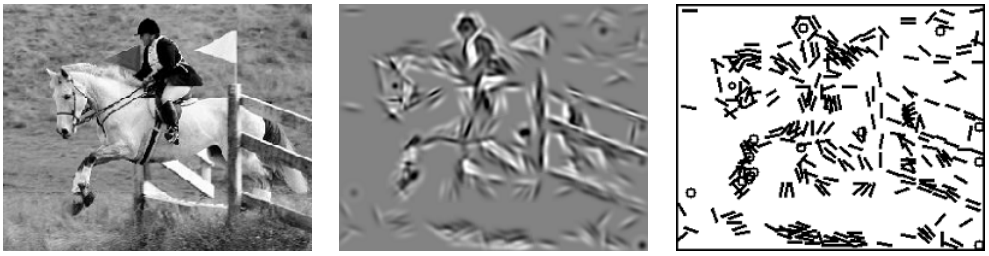
FIG. 14. Sparse coding. (a) and (c) are observed images of 128×128 pixels. (b) and (d) are respectively the reconstructed images using 300 bases.

Now let's examine the sparse coding model empirically by some experiments. In the experiments, we use an overcomplete dictionary of linear bases such as those depicted in Figure (13). At each pixel, there are localized bases of different scales and orientations. So the set of bases is highly overcomplete. We use the matching pursuit algorithm to construct the sparse coding of the observed images.

Figure (14) shows two examples of sparse coding. (a) and (c) are observed images of 128×128 pixels, and (b) and (d) are images reconstructed by 300 bases. We can see that sparse coding is very effective for images with sparse structures, such as image (a). However, the texture information is not well represented. We can continue to add more bases in the matching pursuit process if we want to code texture, but then the representation will no longer be sparse.

There is one more problem with the sparse coding model (18) and (19), which does not have a sophisticated $p(S)$. See Figure (15). (a) is the observed image of 300×200 pixels. (b) is the image reconstructed using 500 bases. (c) is a symbolic representation where each base in the sparse coding is represented by a bar at the same location with the same elongation and orientation as the corresponding base (we also include some isotropic bases in the dictionary, and they are represented by circles). As shown by this experiment, the bases do not line up very well, indicating that we need a stronger model $p(S)$ for the spatial organization of the local bases, so that they line up into more regular structures. To conclude, sparsity alone cannot capture the low entropy of very regular patterns.

4.3. *Markov random fields.* The Markov random fields originated in statistical physics, and they were first introduced to statistics by Besag (1974) [6]. Geman and Geman



(a) (b) (c)

FIG. 15. Sparse coding. (a) is the observed 300×200 image. (b) is the image reconstructed using 500 bases. (c) is a symbolic representation where each base is represented by a bar at the same location with the same elongation and orientation.

(1984) [20] and many other researchers used Markov random fields for image processing and modeling. Zhu and Mumford (1997) [60] connected Markov random fields to partial differential equations and the variational approaches to image processing.

The Markov property of a Markov random field is defined with respect to a neighborhood system, where for each pixel $(x, y) \in D$, there is a set of neighboring pixels $\partial(x, y) \subset D$. The neighborhood relationship is a mutual relationship; that is, if (x, y) is a neighbor of (x', y'), then (x', y') is also a neighbor of (x, y). From the neighborhood system $\partial = \{\partial(x, y) : (x, y) \in D\}$, one can define the set of cliques. A clique A is a set of pixels so that any two pixels in A are neighbors.

$p(\mathbf{I})$ is a Markov random field with respect to the neighborhood system ∂, if for all $(x, y) \in D$,

$$p(\mathbf{I}(x, y) \mid \mathbf{I}(D \backslash (x, y))) = p(\mathbf{I}(x, y) \mid \mathbf{I}(\partial(x, y))). \tag{25}$$

By convention, for $A \subset D$, we define $\mathbf{I}(A)$ as the intensities of all the pixels in A. $D \backslash (x, y)$ denotes all the pixels in D except (x, y). The Markov property (25) means that the distribution of the pixel intensity only depends on the intensities of the neighboring pixels.

According to the Hammersley-Clifford theorem [25], a Markov random field with respect to the neighborhood system ∂ can be written as a Gibbs distribution:

$$p(\mathbf{I}) = \frac{1}{Z} \exp\{-\sum_{A} U_A(\mathbf{I}(A))\},$$

where $U_A()$ is a potential function defined on the clique A, and Z is the normalizing constant to make $p(\mathbf{I})$ sum or integrate to 1.

For modeling purposes, if A has many pixels, then U_A is a high-dimensional function, and it can be difficult to specify it and estimate it from the image data. In statistical physics as well as in early research on image processing, people often assume pairwise potentials, that is, all the U_A with the cardinality of the clique $|A| > 2$ are set to 0. However, for natural images, pairwise relationship can hardly be an adequate description.

Zhu, Wu, and Mumford (1997) [60] proposed a modeling strategy to get around this problem: replacing the high dimension $\mathbf{I}(A)$ by low-dimensional features. The key to their

model construction is to match the marginal distributions of filter responses. Specifically, suppose the image $\mathbf{I}$ is a random sample from an unknown distribution $f(\mathbf{I})$, which we want to estimate or approximate. Let $f_{x,y,s,\alpha}$ be the marginal distribution of $\langle \mathbf{I}, B_{x,y,s,\alpha} \rangle$ under $f(\mathbf{I})$, and let

$$f_{s,\alpha} = \frac{1}{|D|} \sum_{(x,y) \in D} f_{x,y,s,\alpha}$$

be the marginal distribution pooled over D. If $f(\mathbf{I})$ is stationary, then $f_{x,y,s,\alpha} = f_{s,\alpha}$ for all (x, y). $f_{x,y}$ can be estimated from image $\mathbf{I}$ by pooling the marginal histogram of filter responses $\{\langle \mathbf{I}, B_{x,y,s,\alpha} \rangle, \forall (x, y) \in D\}$. The basic idea of Zhu et al. (1997) is to select a set of filters (s_k, α_k), $k = 1, ..., K$, and construct a distribution $p(\mathbf{I})$ so that

$$p_{s_k,\alpha_k} = f_{s_k,\alpha_k}, \ k = 1, ..., K, \tag{26}$$

where p_{s_k,α_k} is the marginal distribution under $p(\mathbf{I})$, defined in a similar way as f_{s_k,α_k}.

There can be many $p(\mathbf{I})$ that satisfy the constraint (26). The one that achieves the maximum entropy is in the form of the following Gibbs distribution or Markov random field:

$$p_\lambda(\mathbf{I}) = \frac{1}{Z(\lambda)} \exp\{\sum_{k=1}^{K} \sum_{(x,y) \in D} \lambda_k(\langle \mathbf{I}, B_{x,y,s_k,\alpha_k} \rangle)\}, \tag{27}$$

where $\lambda = \{\lambda_k(), k = 1, ..., K\}$ is a set of functions of filter responses, $Z(\lambda)$ is the normalizing constant depending on $\{\lambda_k()\}$, and $\lambda = \{\lambda_k(), k = 1, ..., K\}$ is chosen so that the constraint (26) is satisfied. The reason for choosing a maximum entropy distribution among all the $p(\mathbf{I})$ that satisfy (26) is that this distribution is the most random and therefore introduces the least amount of prejudice in approximating $f(\mathbf{I})$.

PROPOSITION 7. Let $p(\mathbf{I})$ be any distribution such that (26) is satisfied. Let $p_\lambda(\mathbf{I})$ be defined as (27), and assume that $p_\lambda(\mathbf{I})$ satisfies (26). Then $\mathcal{H}(p_\lambda) - \mathcal{H}(p) = \mathcal{K}(p||p_\lambda) \geq 0$.

Proof. The key to the proof is the observation that

$$\mathrm{E}_p \left[\sum_{x,y} \lambda_k(\langle \mathbf{I}, B_{x,y,s_k,\alpha_k} \rangle) \right] = \mathrm{E}_{p_\lambda} \left[\sum_{x,y} \lambda_k(\langle \mathbf{I}, B_{x,y,s_k,\alpha_k} \rangle) \right] = |D| \int \lambda_k(r) f_{s_k,\alpha_k}(r) dr,$$

because both p and p_λ share the same marginal distribution f_{s_k,α_k}. Thus $\mathrm{E}_p[\log p_\lambda(\mathbf{I})] = \mathrm{E}_{p_\lambda}[\log p_\lambda(\mathbf{I})]$. Therefore,

$$\begin{aligned} \mathcal{H}(p_\lambda) - \mathcal{H}(p) &= \mathrm{E}_p[\log p(\mathbf{I})] - \mathrm{E}_{p_\lambda}[\log p_\lambda(\mathbf{I})] \\ &= \mathrm{E}_p[\log p(\mathbf{I})] - \mathrm{E}_p[\log p_\lambda(\mathbf{I})] = \mathcal{K}(p||p_\lambda) \geq 0. \quad \square \end{aligned}$$

This proposition leads to the following conclusions:

Maximum entropy: For a fixed set of filters $(s_k, \alpha_k, k = 1, ..., K)$, p_λ of (27) achieves the maximum entropy among all those $p(\mathbf{I})$ satisfying (26), because $\mathcal{H}(p_\lambda) - \mathcal{H}(p) \geq 0$.

Minimum entropy: The true unknown distribution $f(\mathbf{I})$ also satisfies (26), so the above result also holds if we replace $p(\mathbf{I})$ by $f(\mathbf{I})$, that is, $\mathcal{H}(p_\lambda) - \mathcal{H}(f) = \mathcal{K}(f||p_\lambda)$. If we want to find the set of filters $(s_k, \alpha_k, k = 1, ..., K)$ to minimize $\mathcal{K}(f \| p_\lambda)$, we need to minimize the entropy of p_λ.

High entropy regime: p_λ approaches the true distribution f from above in terms of entropy. That is, p_λ is always more random than f. So Markov random field models are capable of modeling high entropy patterns.

Discretization and exponential family model: One can further parametrize $\lambda_k()$ by step functions over a set of bins $\Delta_t, t = 1, ..., T$, so that $\lambda_k(\langle \mathbf{I}, B_{x,y,s_k,\alpha_k}\rangle) = \lambda_{kt}$ if $\langle \mathbf{I}, B_{x,y,s_k,\alpha_k}\rangle \in \Delta_t$. Then model (27) can be written as

$$p_\lambda(\mathbf{I}) = \frac{1}{Z(\lambda)} \exp\{\sum_k \sum_{x,y} \sum_t \lambda_{k,t} \delta(\langle \mathbf{I}, B_{x,y,s_k,\alpha_k}\rangle \in \Delta_t)\} \tag{28}$$

$$= \frac{1}{Z(\lambda)} \exp\{\sum_k \sum_t \lambda_{k,t} H_{k,t}(\mathbf{I})\} \tag{29}$$

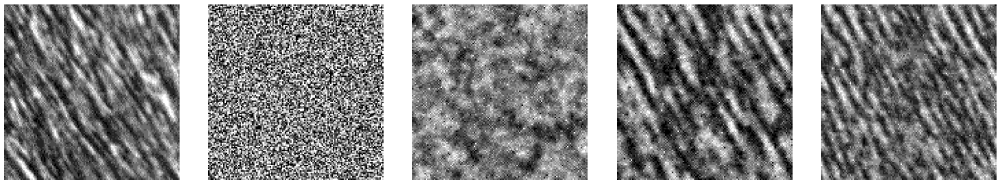
$$= \frac{1}{Z(\lambda)} \exp\{\sum_k \langle \lambda_k, H_k(\mathbf{I})\rangle\}. \tag{30}$$

In (28), $\delta(\langle \mathbf{I}, B_{x,y,s_k,\alpha_k}\rangle \in \Delta_t) = 1$ if $\langle \mathbf{I}, B_{x,y,s_k,\alpha_k}\rangle \in \Delta_t$, and $\delta = 0$ otherwise. In (29), $H_{k,t}(\mathbf{I}) = \sum_{x,y} \delta(\langle \mathbf{I}, B_{x,y,s_k,\alpha_k}\rangle \in \Delta_t)$, i.e., the number of $\langle \mathbf{I}, B_{x,y,s_k,\alpha_k}\rangle$ falling into bin Δ_t. In (30), $H_k = (H_{k,t}, \forall t)$ is the marginal histogram of $\{\langle \mathbf{I}, B_{x,y,s_k,\alpha_k}\rangle, \forall (x, y)\}$. Let $h_k(\mathbf{I}) = H_k(\mathbf{I})/|D|$ be the normalized histogram. If $\mathbf{I} \sim f(\mathbf{I})$, then $h_k(\mathbf{I})$ is an estimate of the marginal distribution f_{s_k,α_k}.

Model (30) is the so-called exponential family model. $H_k(\mathbf{I})$ are the sufficient statistics. $\lambda_k = (\lambda_{k,t}, \forall t)$ are the parameters. $\lambda = (\lambda_k, \forall k)$ can be estimated from the observed image $\mathbf{I}_{\text{obs}}$ by solving the following estimation equation:

$$\mathrm{E}_\lambda[H_k(\mathbf{I})] = H_k(\mathbf{I}_{\text{obs}}), \forall k. \tag{31}$$

Filter pursuit: Zhu, et al. (1997) proposed a filter pursuit procedure to add one filter at a time, so that the added filter leads to the maximum reduction of the entropy of the fitted model p_λ. Figure (16) displays an example of a filter pursuit procedure on a homogeneous texture. With the $K = 0$ filter, the sampled image is white noise. With the $K = 7$ filters, the sampled image in (e) is perceptually equivalent to the input image.



(a) observed (b) $K = 0$ (c) $K = 2$ (d) $K = 3$ (e) $K = 7$

FIG. 16. Filter pursuit: adding one filter at a time to reduce the entropy.

Micro-canonical ensemble: Wu, Zhu, and Liu (2000) [55] considered the following ensemble, which is called a micro-canonical ensemble in statistical physics (Chandler, 1987):

$$\Omega(h) = \{\mathbf{I} : h_k(\mathbf{I}) = h_k, \forall k\}, \tag{32}$$

where $h = (h_k, \forall k)$ can be estimated from the observed image. This is a deterministic concept of equivalent class, where all the images in this ensemble share the same set of spatial statistics.

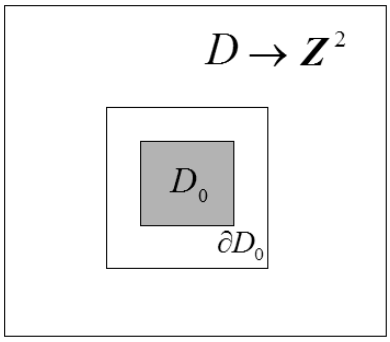


FIG. 17. The deterministic concept of a micro-canonical ensemble defined on $D \to \mathbf{Z}^2$ produces the probabilistic concept of a Markov random field on a fixed patch D_0, according to the equivalence of ensembles in statistical physics.

As observed by Wu, et al. (2000), the uniform distribution over $\Omega(h)$, $\text{Unif}(\Omega(h))$, can be made equivalent to a Markov random field, thanks to two of the most profound results in statistical physics and information theory respectively.

1) According to the equivalence of ensembles in statistical physics [11], under $\text{Unif}(\Omega(h))$, for any fixed part of the image lattice $D_0 \subset D$, as $D \to \mathbf{Z}^2$, the image intensities of D_0 converge to

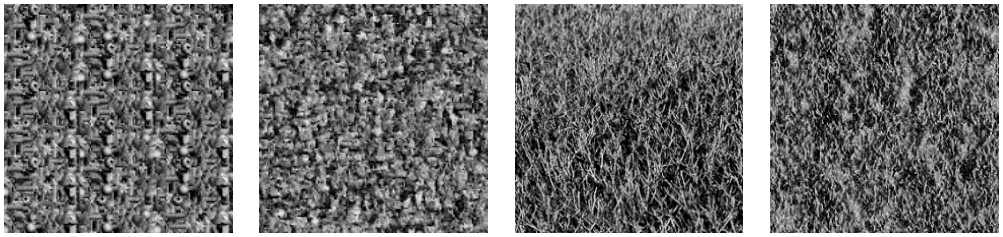
$$p(\mathbf{I}_{D_0} \mid \mathbf{I}_{\partial D_0}) = \frac{1}{Z(\lambda)} \exp\{\sum_k \sum_{x,y \in D_0} \lambda_k(\langle \mathbf{I}, B_{x,y,s_k,\alpha_k} \rangle)\},$$

where ∂D_0 are the neighboring pixels of D_0 so that pixels in ∂D_0 and pixels in D_0 may be covered by the same filters. λ can be solved from Equation (31).

2) According to the asymptotic equipartition property in information theory [1, 3], as $D \to \mathbf{Z}^2$, the Markov random field model (30) is equivalent to the uniform distribution over a micro-canonical ensemble (32) in the absence of a phase transition. One can show that the entropy rate of the Markov random field model approaches $\log |\Omega(h)|/|D|$ asymptotically, where $|\Omega(h)|$ is the volume of $\Omega(h)$.

The following are some experiments with a fixed set of filters. These experiments show that the filter statistics are quite effective in representing stochastic textures. Figure (18) shows two examples. (a) and (c) are observed images, and (b) and (d) are respectively the "reconstructed" images. Here the reconstruction is of a statistical nature: (b) and (d) are sampled from the respective micro-canonical ensembles (32) by matching feature statistics. See [26, 46] for more discussions on feature statistics.

We need to stress that, under the Markov random field model or equivalently the micro-canonical ensemble, the filter responses $\langle \mathbf{I}, B_{x,y,s,\alpha} \rangle$ are not independent of each other, because the number of bases $B_{x,y,s,\alpha}$ far exceeds the number of pixels. Although only marginal distributions are specified, the dependencies among adjacent responses

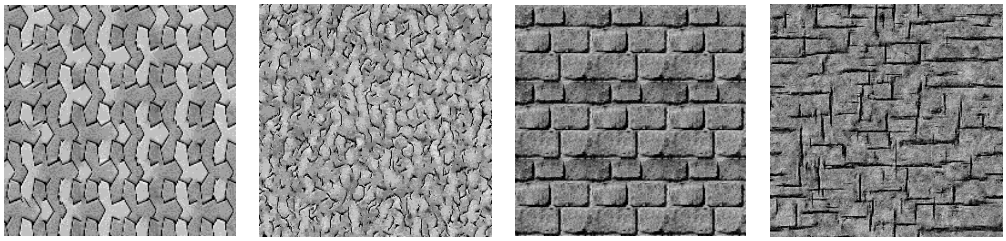


(a) (b) (c) (d)

FIG. 18. (a) and (c) are observed images. (b) and (d) are simulated by matching marginal histograms.

from the same filter can be accounted for implicitly by the distributions of the responses from other filters. Sometimes, long range patterns can emerge by matching statistics of local features.

But still, since the model only specifies the marginal distributions of filter responses, it cannot represent large regular structures very well. See Figure (19) for two examples with line structures. In order to model regular structures, we need to represent these structures explicitly. Moreover, we also need to model the spatial organizations of these structures.



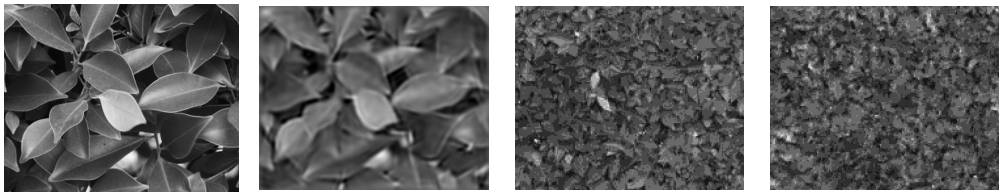
(a) (b) (c) (d)

FIG. 19. (a) and (c) are observed images. (b) and (d) are simulated by matching marginal histograms.

In the end, we would like to mention that if the linear bases form a complete system, i.e., the number of bases is the same as the number of pixels, then both the wavelet model (18) and (19) (with $\epsilon = 0$) and the Markov random field model (27) reduce to the independent component analysis model [5].

5. Integrating different regimes of models.

5.1. *Motivation.* Our examination of wavelet sparse coding and Markov random fields indicates that they are appropriate for different entropy regimes. Because information scaling transforms a low-entropy image to a high-entropy image, we need both models to represent natural images over the whole range of the scale. For instance, Figure (20) displays results for two images of leaves. (a) is the observed 300×200 image of leaves at a near distance. (b) is the image reconstructed by the matching pursuit algorithm using 1,000 wavelet bases. (c) is the observed image at a far distance. (d) is obtained



(a) (b) (c) (d)

FIG. 20. From sparse coding to random field. (a) Observed 300 × 200 image at a near distance. (b) Reconstructed by sparse coding with 1,000 bases. (c) Observed 300 × 200 image at a far distance. (d) Synthesized by matching marginal histograms.

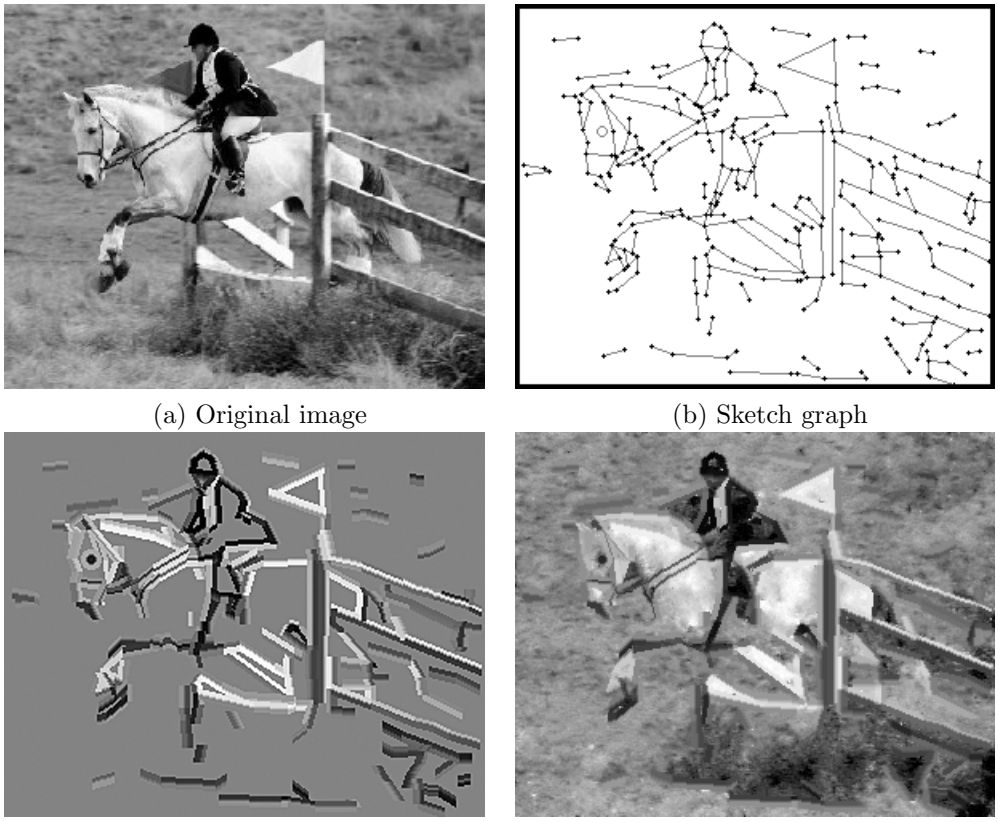
by matching the histograms of filter responses from a set of filters. (d) is not an exact reconstruction of (a), but it captures the texture appearance of (c).

Because visual objects can appear at different scales in the same image, we need to combine the two regimes of models into an integrated model. In what follows, we shall first propose such a model, which we call the "primal sketch" model. Experiments on this model show that it can represent a large variety of natural images. After that we propose a statistical theory that embraces different regimes of models in a common theoretical framework.

5.2. *Primal sketch model.* The term "primal sketch" comes from Marr (1982) [36], who, in his book on vision, proposed a symbolic representation of image intensities for the initial stage of visual computation.

As shown by Figure (15) as well as Figure (20.b), the wavelet sparse coding model (18) and (19) with independence assumptions on the wavelet coefficients is not efficient for coding geometric structures such as edges and bars, as well as lines, curves, junctions, and corners. The reason is that the independence assumption does not capture the low entropy of these geometric structures. There are two schemes to improve upon model (18) and (19). (1) Replace the wavelet bases by some more sophisticated image functions or primitives for sparse coding. (2) Model the joint distributions of the wavelet coefficients. These two schemes are closely related. In the primal sketch model, we introduce explicit geometric sketch primitives as the elements for sparse coding. In the next subsection, we shall study the connection between the two schemes.

Figure (21) illustrates the basic idea of the primal sketch model. Figure (21.a) is the observed image, the same as the observed image in Figure (15). It is represented by a small number of sketch primitives, which form a sketch graph; see Figure (21.b). The nodes are end points, corners, and junctions. The nodes are connected by edges and bars. These sketch primitives generate what we call the "sketchable" part of the image; see Figure (21.c). The image intensities generated by these primitives are very close to the corresponding image intensities of the original image. That is, we seek a sparse deterministic coding for the sketchable part of the image, which captures the low entropy portion of the image. For the remaining "nonsketchable" part of the image, we fill in textures by matching the marginal histograms of filter responses, or more formally, we use Markov random fields to characterize the high entropy portion of the image. The



(a) Original image (b) Sketch graph

(c) Sketchable part (d) Synthesized image

FIG. 21. Primal sketch model. (a) Observed image. (b) "Sketchable" part is described by a geometric sketch graph. (c) The sketchable part of the image. (d) Fill in the "nonsketchable" part by matching feature statistics.

Markov random fields fill in the nonsketchable part while using the sketchable part as the boundary condition. So the final result in Figure (21.d) is a seamless integration of structures and textures.

Geometric sketch primitives: A most common sketch primitive is an oriented and elongated structure such as an edge or a bar:

$$\Phi(x, y) = h(-(x - x_0)\sin\alpha + (y - y_0)\cos\alpha), \quad (x, y) \in D_0,$$

where $h()$ is a one-dimensional profile function, and D_0 is an oriented rectangle set of pixels along direction α. The low entropy is achieved by the fact that the two-dimensional image patch $\mathbf{I}(D_0)$ can be represented accurately by a one-dimensional profile $h()$ along a direction α. Moreover, the profile $h()$ can be further modeled by some parametric functions. For edges, Elder and Zucker (1998) [16] proposed the following profile. Let $h_0(x)$ be a step edge: $h_0(x) = 1/2$ for $x \leq 0$, and $h_0(x) = -1/2$ for $x > 0$. Let $h(x) = a + bh_0 * g_s$, where g_s is a Gaussian kernel with bandwidth s. Here the parameter of

such an edge primitive is $\theta = (x_0, y_0, l, w, s, \alpha, a, b)$ with location (x_0, y_0), length l, width w, scale s, orientation α, local intensity level a, edge contrast b. The convolution with Gaussian kernel of scale s is used to reflect the blurred transition of intensity values across the edge, caused by the three-dimensional shape of the underlying physical structure that produces the edge, as well as the resolution and focus of the camera. A bar structure is a composition of two edges. Junctions and corners are compositions of edges and bars.

Integrated model: For an image $\mathbf{I}$ defined on a lattice D, let $S = \{\Phi_i(x, y \mid \theta_i), i = 1, ..., n\}$ be the set of sketch primitives to be used to model the sketchable part of $\mathbf{I}$. Let D_S be the sketchable part of the lattice. Let $D_{S,i}$ be the pixels covered by $\Phi_i(x, y \mid \theta_i)$. Then $D_S = \bigcup_{i=1}^n D_{S,i}$. Let $\bar{D}_S = D \backslash D_S$ be the nonsketchable part of the image. The primal sketch model is as follows:

$$\mathbf{I}(x, y) = \Phi_i(x, y|\theta_i) + \epsilon, \ \epsilon \sim \text{iid N}(0, \tau^2), \ \ (x, y) \in D_{S,i}, i = 1, ..., n; \tag{33}$$

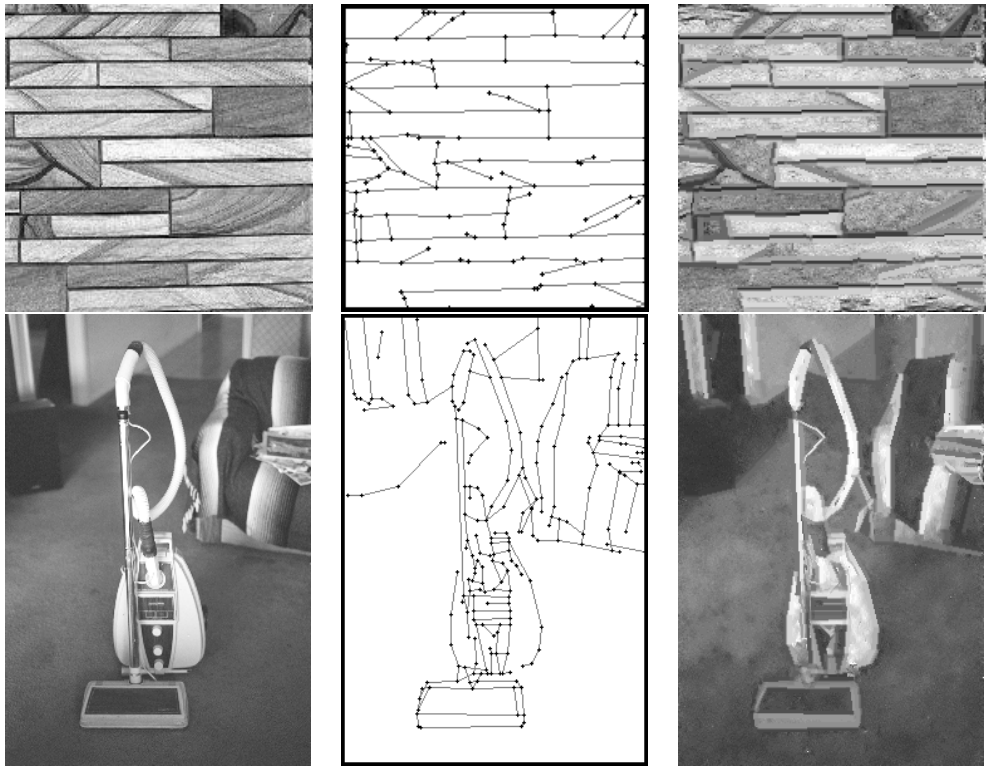
$$p(\mathbf{I}(\bar{D}_S) \mid \mathbf{I}(D_S)) = \frac{1}{Z(\lambda)} \exp\{\sum_k \sum_{(x,y)\in \bar{D}_S} \lambda_k(\langle \mathbf{I}, B_{x,y,s_k,\alpha_k} \rangle)\}. \tag{34}$$

In the above model, the sketchable part of the image is represented by a small number of sketch primitives. The nonsketchable part of the image is described by the Markov random field model. According to that model, the nonsketchable part $\mathbf{I}(\bar{D}_S)$ is generated conditionally on the sketchable part $\mathbf{I}(D_S)$. In the language of Markov random fields, $\mathbf{I}(D_S)$ serves as the boundary conditions, because some of the bases B_{x,y,s_k,α_k} in the above Markov random field model can cover both $\mathbf{I}(D_S)$ and $\mathbf{I}(\bar{D}_S)$. One may also consider $p(\mathbf{I}(\bar{D}_S) \mid \mathbf{I}(D_S))$ as an inpainting $\mathbf{I}(\bar{D}_S)$ by interpolating $\mathbf{I}(D_S)$, where λ_k functions control the overall smoothness as well as other texture properties. See [10] for more details on inpainting. See also [44] for an image decomposing scheme in terms of cartoons and textures.

The nonsketchable part may consist of several regions of different textures with different marginal histograms. If this is the case, we need to segment the nonsketchable part of the image and fit a separate Markov random field for each segmented region.

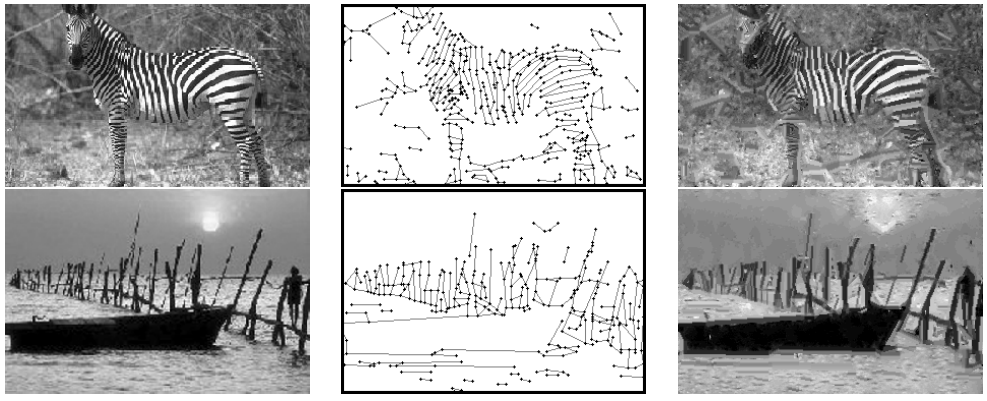
The prior model for S is of the following form: $p(S) \propto \exp\{\beta(S)\}$, where $\beta(S)$ is specified to favor sketch graphs with extended and connected primitives by penalizing the number of primitives and the number of free end points. We refer to the companion paper [24] for details on this issue.

Connection between sparse coding and Markov random field: The primal sketch model combines sparse coding (33) and the Markov random field (34). These two components are closely connected. On the sketchable part of the image, which can be accurately represented by sketch primitives, the filter responses $\langle \mathbf{I}, B_{x,y,s,\alpha} \rangle$ exhibit very regular spatial patterns. As a matter of fact, the sketch primitives such as edge and bar segments are detected from such regular patterns of filter responses. In particular, those bases $B_{x,y,s,\alpha}$ that achieve a local maximum are typically located on the edge or bar segments with their orientations aligned with those of the corresponding edge and bar segments. On the remaining part of the image where there are no such joint patterns formed by filter responses, the image cannot be sketched, and can only be summarized by marginal histograms of filter responses due to the lack of joint patterns. In this sense, the marginal



(a) (b) (c)

FIG. 22. Examples of primal sketch model. (a) Observed image. (b) Sketch graph. (c) Synthesized image from the fitted model.



(a) (b) (c)

FIG. 23. Examples of primal sketch model. (a) Observed image. (b) Sketch graph. (c) Synthesized image from the fitted model.

histograms recycle the filter responses that fail to detect sketch primitives or fail to form joint patterns.

Under the scaling transformation of zooming out the image, the sketch primitives will become smaller and eventually out of zoom. Consequently, the joint patterns of the filter responses gradually diminish, with only the marginal distributions left to be used for characterizing the image.

A stage-wise procedure is used to fit the above model. In the first stage, the sketch primitives are identified by minimizing $\sum_i \sum_{(x,y)\in D_{S,i}} (\mathbf{I}(x,y) - \Phi_i(x,y|\theta_i))/2\tau^2 - \beta(S)$, after the filter responses identify the candidates for the sketch primitives. In the second stage, the remaining nonsketchable part of the image lattice is segmented into homogeneous regions, and a random field model is fitted in each region by reproducing the marginal histograms of filter responses. The reader is referred to [24] for a more sophisticated version of the algorithm with all the technical details. Figures (22) and (23) show some examples.

5.3. *Towards a unified theoretical framework.* Following the theory of Della Pietra, Della Pietra, and Lafferty [14], we propose a theoretical framework for image modeling and learning. This framework embraces different regimes of models.

Let $f(\mathbf{I})$ be an unknown distribution that we want to estimate or approximate based on random samples from $f(\mathbf{I})$. For instance, $f(\mathbf{I})$ may be the distribution of a texture pattern, or the distribution of geometric primitives, or the distribution of a class of objects such as faces or cars.

From test statistics to models: We start from a reference model or a null hypothesis $H_0 : \mathbf{I} \sim q(\mathbf{I})$. For example, $q(\mathbf{I})$ can be the uniform distribution, or the Gaussian white noise distribution, or the current approximation to $f(\mathbf{I})$. We then modify $q(\mathbf{I})$ to a model $p(\mathbf{I})$ by identifying test statistics to reject the null hypothesis H_0.

Specifically, we extract a set of low-dimensional distributions from the high-dimensional $f(\mathbf{I})$. Let's denote them by $\varphi(f)$. $\varphi(f)$ can be estimated using random samples from f and can serve as the test statistics that reveal the departure of $f(\mathbf{I})$ from $q(\mathbf{I})$. We want to choose the dimensions of $\varphi(f)$ so that $\varphi(f)$ exposes the most glaring departure from $q(\mathbf{I})$ or provides the strongest evidence against H_0. From such $\varphi(f)$, a model can be constructed by improving upon $q(\mathbf{I})$ along the dimensions of $\varphi(f)$ while leaving the remaining dimensions unchanged.

According to the Stein-Chernoff Lemma [12], $\mathcal{K}(p||q)$ measures the optimal exponential decay rate of type I error for testing the hypotheses: $H_0 : \mathbf{I} \sim q(\mathbf{I})$ versus $H_1 : \mathbf{I} \sim p(\mathbf{I})$, if we let the type II error goes to 0. $\mathcal{K}(p||q)$ is also the expected log-likelihood ratio under H_1. Therefore, we use $\mathcal{K}(p||q)$ to measure the departure of $p(\mathbf{I})$ from $q(\mathbf{I})$.

(1) *Minimum divergence:* For a given φ, among all the distributions $p(\mathbf{I})$ that satisfy $\varphi(p) = \varphi(f)$, we choose the one that minimizes $\mathcal{K}(p||q)$. Specifically, let

$$p_\varphi = \arg \min_{p:\varphi(p)=\varphi(f)} \mathcal{K}(p||q)$$

be such a minimum divergence distribution. p_φ is the model resulting from matching the low-dimensional distributions $\varphi(f)$.

(2) *Maximum divergence:* Among all possible φ, we choose the one that maximizes $\mathcal{K}(p_\varphi||q)$, where for each φ, p_φ is the minimum divergence distribution obtained by (1).

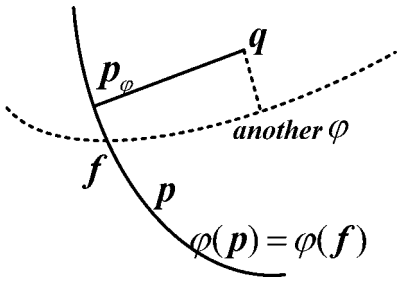


FIG. 24. Illustration of the max-min divergence principle. Each distribution is a point. The solid curve represents $\{p : \varphi(p) = \varphi(f)\}$ for some φ. The dotted curve corresponds to another φ, which is less preferable than the φ of the solid curve.

This max-min divergence principle is a generalized version of the minimax entropy principle studied by [60]. Given φ, we want $\varphi(p)$ to match $\varphi(f)$, but other than that, we want p to stay as close to q as possible. One may interpret this as a "least action principle" in modeling, where, other than $\varphi(f)$, we should refrain from introducing artificial evidence against H_0. Thus $\mathcal{K}(p_\varphi||q)$ measures the strength of evidence in $\varphi(f)$ alone. Among all possible φ, we want to choose φ so that the resulting least action or most conservative p_φ is as far from q as possible. Figure (24) illustrates the basic idea, where each distribution is a point. For each φ, the set of distributions $\{p : \varphi(p) = \varphi(f)\}$ is represented by a curve. p_φ can be imagined as the "projection" of q onto this curve, and $\mathcal{K}(p||q)$ can be imagined as squared distance. In Figure (24), the φ of the solid curve is preferred to the φ of the dotted curve.

The following are two types of $\varphi(f)$.

(1) *Marginal distributions:* Let f_{x,y,s_k,α_k} be the distribution of $\langle \mathbf{I}, B_{x,y,s_k,\alpha_k}\rangle$ under $f(\mathbf{I})$, for a selected set of filters $\{(s_k, \alpha_k), k = 1, ..., K\}$. Let $f_{s_k,\alpha_k} = \sum_{(x,y)\in D} f_{x,y,s_k,\alpha_k}/|D|$. If $f(\mathbf{I})$ is stationary within the lattice D, then $f_{x,y,s_k,\alpha_k} = f_{s_k,\alpha_k}$ for all $(x, y) \in D$, and f_{s_k,α_k} can be estimated by the marginal histogram $H_k(\mathbf{I})$ in (30). f_{s_k,α_k} involves spatial pooling over pixel locations. We can write $\varphi = \{(s_k, \alpha_k), k = 1, ..., K\}$.

(2) *Joint distributions:* Let $S = (B_{x_j,y_j,s_j,\alpha_j}, j = 1, ..., n)$ be a small number of select bases. Let f_S be the joint distribution of $(\langle \mathbf{I}, B_{x_j,y_j,s_j,\alpha_j}\rangle, j = 1, ..., n)$ under $f(\mathbf{I})$. f_S can also be estimated using random samples from f. f_S involves bases at selected locations. We can write $\varphi = S = (B_{x_j,y_j,s_j,\alpha_j}, j = 1, ..., n)$.

The following propositions give specific forms of p_φ, which are parametrized by $p_\lambda(\mathbf{I})$.

PROPOSITION 8. Let $p(\mathbf{I})$ be a distribution such that $p_{s_k,\alpha_k} = f_{s_k,\alpha_k}, k = 1, ..., K$. Then among all such $p(\mathbf{I})$, suppose there is a distribution of the following form:

$$p_\lambda(\mathbf{I}) = \exp\{\sum_{k=1}^{K} \sum_{(x,y)\in D} \lambda_k(\langle \mathbf{I}, B_{x,y,s_k,\alpha_k}\rangle)\} q(\mathbf{I}). \tag{35}$$

Then $\mathcal{K}(p||q) - \mathcal{K}(p_\lambda||q) = \mathcal{K}(p||p_\lambda) \geq 0$. That is, among all such p, p_λ achieves the minimum of $\mathcal{K}(p||q)$. Moreover, $\mathcal{K}(f||p_\lambda) = \mathcal{K}(f||q) - \mathcal{K}(p_\lambda||q)$. Thus by maximizing $\mathcal{K}(p_\lambda||q)$ among all possible sets of filters $\{(s_k, \alpha_k), k = 1, ..., K\}$, we minimize $\mathcal{K}(f||p_\lambda)$.

Proof. Because both p and p_λ satisfy $p_{s_k,\alpha_k} = f_{s_k,\alpha_k}$, we have

$$\begin{aligned} \mathrm{E}_{p_\lambda}\left[\sum_{(x,y)\in D} \lambda_k(\langle \mathbf{I}, B_{x,y,s_k,\alpha_k}\rangle)\right] &= \mathrm{E}_{p}\left[\sum_{(x,y)\in D} \lambda_k(\langle \mathbf{I}, B_{x,y,s_k,\alpha_k}\rangle)\right] \\ &= |D| \int \lambda_k(r) f_{s_k,\alpha_k}(r) dr. \end{aligned}$$

Then

$$\begin{aligned} \mathcal{K}(p_\lambda||q) &= \mathrm{E}_{p_\lambda}\left[\log \frac{p_\lambda(\mathbf{I})}{q(\mathbf{I})}\right] = \mathrm{E}_{p}\left[\log \frac{p_\lambda(\mathbf{I})}{q(\mathbf{I})}\right] \\ &= \mathrm{E}_{p}\left[\log \frac{p(\mathbf{I})}{q(\mathbf{I})} - \log \frac{p(\mathbf{I})}{p_\lambda(\mathbf{I})}\right] = \mathcal{K}(p||q) - \mathcal{K}(p||p_\lambda). \end{aligned}$$

The above equation is still true if we replace p by f. Therefore, $\mathcal{K}(f||p_\lambda) = \mathcal{K}(f||q) - \mathcal{K}(p_\lambda||q)$. □

Connection to Markov random field: Compared to model (27), the normalizing constant $Z(\lambda)$ is absorbed into the λ_k functions in (35). Also, in model (27), $q(\mathbf{I})$ is assumed to be a uniform measure, whereas in model (35), $q(\mathbf{I})$ can be any distribution or measure.

PROPOSITION 9. For $S = (B_{x_j,y_j,s_j,\alpha_j}, j = 1, ..., n)$, let $p(\mathbf{I})$ be a distribution such that $p_S = f_S$. Then among all such $p(\mathbf{I})$, there is a distribution

$$p_\lambda(\mathbf{I}) = \exp\{\lambda(\langle \mathbf{I}, B_{x_j,y_j,s_j,\alpha_j}\rangle, j = 1, ..., n)\} q(\mathbf{I}), \tag{36}$$

where $\lambda(r_1, ..., r_n) = \log(f_S(r_1, ..., r_n)/q_S(r_1, ..., r_n))$, $r_j = \langle \mathbf{I}, B_{x_j,y_j,s_j,\alpha_j}\rangle$, $j = 1, ..., n$. $\mathcal{K}(p||q) - \mathcal{K}(p_\lambda||q) = \mathcal{K}(p||p_\lambda) \geq 0$, and $\mathcal{K}(p_\lambda||q) = \mathcal{K}(f_S||q_S) = \mathcal{K}(f||q) - \mathcal{K}(f||p_\lambda)$. Thus by maximizing $\mathcal{K}(f_S||q_S)$ among all possible $S = (B_{x_j,y_j,s_j,\alpha_j}, j = 1, ..., n)$, we minimize $\mathcal{K}(f||p_\lambda)$. In other words, we identify S so that the hypothesis testing $H_0 : (r_1, ..., r_n) \sim q_S$ versus $H_1 : (r_1, ..., r_n) \sim f_S$ has the maximum expected log-likelihood ratio.

The above proposition can be proved in a similar way as the proof of Proposition 8. But it can be written in a more explicit form. Consider a linear change of variable $\mathbf{I} \to (R, \bar{R})$, where $R = (r_1, ..., r_n)'$, and $\bar{R}$ consists of the coordinates of $\mathbf{I}$ in the subspace that is orthogonal to the space spanned by $S = (B_{x_j,y_j,s_j,\alpha_j}, j = 1, ..., n)$. We can choose any orthonormal basis in this subspace. Under such a linear transformation, let $q(R, \bar{R})$ and $f(R, \bar{R})$ be the joint distributions of $(R, \bar{R})$ under q and f respectively. Then under p_λ, the distribution of $(R, \bar{R})$ is $f_S(R)q(\bar{R} \mid R)$, where $q(\bar{R} \mid R) = q(R, \bar{R})/q_S(R)$ is the conditional distribution of $\bar{R}$ given R under q. That is, p_λ is constructed by replacing the distribution $q_S(R)$ by $f_S(R)$, while maintaining $q(\bar{R} \mid R)$.

Model (36) is actually a generalized version of projection pursuit [19]. In model (36), multiple bases can be selected at once, and the selected bases can form specific patterns modeled by $p(S)$ and $f_S(R)$. $\langle \mathbf{I}, B_{x_j,y_j,s_j,\alpha_j}\rangle$ can be discretized, or f_S and q_S can be estimated by histograms. $\langle \mathbf{I}, B_{x_j,y_j,s_j,\alpha_j}\rangle$ can also be replaced by nonlinear transforms.

Connection to sparse coding model: Model (36) can be written in the form of the sparse coding model (23) and (24). In matrix form, let $S = (B_{x_j,y_j,s_j,\alpha_j}, j = 1, ..., n)$ be the $|D| \times n$ matrix whose columns are vectorized versions of B_{x_j,y_j,s_j,α_j}. So $R = S'\mathbf{I}$. Let $\bar{S}$ be the $(|D| \times (|D| - n))$-dimensional matrix whose columns are orthonormal and are

orthogonal to the columns of S, so $\bar{R} = \bar{S}'\mathbf{I}$. Let $C = (c_1, ..., c_n)' = (S'S)^{-1}R$ be the least squares coefficients that project $\mathbf{I}$ onto the subspace spanned by $S = (B_{x_j,y_j,s_j,\alpha_j}, j = 1, ..., n)$. Then $\mathbf{I} = \sum_j c_j B_{x_j,y_j,s_j,\alpha_j} + \epsilon$, where $\epsilon = \bar{S}\bar{R}$ is the residual image residing in the subspace spanned by $\bar{S}$. If $q(\mathbf{I})$ is a Gaussian white noise model with mean 0 and variance σ^2, then $\bar{R}$ is independent of R, and each component of $\bar{R}$ follows $\mathrm{N}(0, \sigma^2)$. Therefore, model (36) can be written as

$$(c_1, ..., c_n) \sim f(c_1, ..., c_n); \tag{37}$$

$$\mathbf{I} = \sum_{j=1}^{n} c_j B_{x_j,y_j,s_j,\alpha_j} + \epsilon; \tag{38}$$

$$\epsilon = \bar{S}\bar{R}; \quad \bar{R} \sim \text{iid } \mathrm{N}(0, \sigma^2), \tag{39}$$

where $f(c_1, ..., c_n)$ in (37) is the distribution of $(c_1, ..., c_n)$ under $f(\mathbf{I})$ and can be obtained from $f_S(R)$ via the linear transformation $C = (S'S)^{-1}R$. The above model is essentially the same as the sparse coding model (23) and (24). The only difference is that $C = (c_1, ..., c_n)$ becomes a deterministic transform of $\mathbf{I}$, and ϵ in (38) and (39) is a white noise model in the residual $(|D| - n)$-dimensional space that is orthogonal to $(B_{x_j,y_j,s_j,\alpha_j}, j = 1, ..., n)$. Model (36) or model (37)–(39) has the advantage that the log-likelihood is in closed form, so the model can be easily fitted to the observed data. The original sparse coding model (23) and (24) treats C as a latent variable; thus the model fitting involves integrating out C.

One can also take $q(\mathbf{I})$ as the uniform distribution over the sphere $\Omega(\sigma^2) = \{\mathbf{I} : \|\mathbf{I}\|^2/|D| = \sigma^2\}$, where σ^2 is the marginal variance of the observed image. We can get a model similar to the above, except that in (39), σ^2 needs to be replaced by the unbiased estimate of the residual variance.

Connection to primal sketch model and sketchability: In model (36), $S = (B_{x_j,y_j,s_j,\alpha_j}, j = 1, ..., n)$ can be assumed to follow a distribution $p(S)$. There are various $p(S)$ and the corresponding $f_S(R)$. For the primal sketch model (33) and (34), the $p(S)$ and $f_S(R)$ are of very low entropy. Consider a simple example where there is an edge segment $\Phi(x, y \mid \theta)$, and $\mathbf{I}(x, y) = \Phi(x, y \mid \theta) + \mathrm{N}(0, \tau^2)$. If we choose $S = (B_{x_j,y_j,s_j,\alpha_j}, j = 1, ..., n)$, so that they form a straight line segment that is identical to Φ, or more specifically, $\{B_{x_j,y_j,s_j,\alpha_j}\}$ are connected, (x_j, y_j) lie on the central line of Φ and are equally spaced, and s_j and α_j are all identical, where α_j is the same as the orientation of the edge segment, then $f_S(r_1, ..., r_n) = \mathrm{N}(\mu, S'S\tau^2)$, where r_j has the same expectation μ because Φ has a constant step-edge profile. If τ^2 is small, this $f_S(r_1, ..., r_n)$ has very low entropy. So detecting Φ amounts to finding $S = (B_{x_j,y_j,s_j,\alpha_j}, j = 1, ..., n)$ to maximize $\mathcal{K}(f_S||q_S)$. If $q(\mathbf{I})$ is a white noise model with iid $\mathrm{N}(0, \sigma^2)$, then $q_S(r_1, ..., r_n) = \mathrm{N}(0, S'S\sigma^2)$. In other words, we are searching for the following "sketchability" test:

$$H_0 : (r_1, ..., r_n) \sim \mathrm{N}(0, S'S\sigma^2) \text{ versus } H_1 : (r_1, ..., r_n) \sim \mathrm{N}(\mu, S'S\tau^2) \tag{40}$$

with the maximum likelihood ratio. The resulting $S = (B_{x_j,y_j,s_j,\alpha_j}, j = 1, ..., n)$ then gives us the line segment. See also [38] for a hypothesis testing approach to detecting line segments based on meaningful alignment.

On the part of the image where sketchability tests have failed, i.e., we fail to find the joint distributions of the form (40) in H_1 to reject H_0, we can only reject H_0 by pooling the marginal distributions, which can be highly non-Gaussian. This gives us model (35).

Scaling triggers transitions: Both the sparse coding model and the Markov random field model are special cases of the common modeling scheme proposed in the beginning of this subsection. The sparse coding model identifies the joint distribution of bases at selected locations. These selected bases form the low entropy foreground. The Markov random field model pools the marginal distributions pooled over locations. These marginal distributions characterize the high entropy background where no low entropy joint patterns can be detected.

Under the scaling process of zooming out the image, the joint patterns will be gradually weakened and eventually out of zoom, so that only the marginal distributions are available to reject the Gaussian white noise hypothesis. As the scaling process goes on, the image will eventually converge to the Gaussian white noise. The modeling process is the process of identifying departures from the Gaussian white noise hypothesis. We shall study the scaling transition in more depth in future work.

Mid-entropy regime: The transition from low entropy geometric patterns to high entropy texture patterns is a gradual one during the scaling process. There is a mid-entropy regime between the low entropy regime and the high entropy regime. For images in this regime, if we look at a 30×30 image patch, we do not see long lines and big regions, neither do we see random textures, but we see various types of objects, such as faces and cars. Let $f(\mathbf{I})$ be such a mid-entropy distribution. $f(\mathbf{I})$ can still be modeled by (36), except that $p(S)$ is not a low entropy geometric pattern where all the $(B_{x_j,y_j,s_j,\alpha_j}, j = 1, ..., n)$ form simple straight lines. Instead, $S = (B_{x_j,y_j,s_j,\alpha_j}, j = 1, ..., n)$ captures the more complex overall shapes of the objects, and $p(S)$ should account for shape deformation by allowing the B_{x_j,y_j,s_j,α_j} to actively shift their locations and orientations. Moreover, $p(S)$ and $f_S(R)$ can only be learned from multiple training images. We shall report our work on this model elsewhere.

Connection to AdaBoost: For mid-entropy $f(\mathbf{I})$, if we do not model the deformation, and if we discretize $\langle \mathbf{I}, B_{x_j,y_j,s_j,\alpha_j} \rangle$ into binary values by thresholding, then we obtain the following model:

$$p_\lambda(\mathbf{I}) = \frac{1}{Z(\lambda)} \exp\{\sum_{j=1}^{n} \lambda_j \delta(\langle \mathbf{I}, B_{x_j,y_j,s_j,\alpha_j} \rangle)\} q(\mathbf{I}), \tag{41}$$

where we further simplify the λ function in (36) into an additive form. $\delta(\langle \mathbf{I}, B_{x_j,y_j,s_j,\alpha_j} \rangle) = 1$ if $|\langle \mathbf{I}, B_{x_j,y_j,s_j,\alpha_j} \rangle| > \xi$, where ξ is a threshold. $\delta = 0$ otherwise. $\lambda = (\lambda_j, j = 1, ..., n)$ are parameters, and $Z(\lambda)$ is the normalizing constant. This model can be considered to be a generative version of the AdaBoost method [18] of Viola and Jones (2004) [52], who used Harr wavelets instead of Gabor wavelets for computational efficiency. In AdaBoost, $\delta(\langle \mathbf{I}, B_{x_j,y_j,s_j,\alpha_j} \rangle)$ are called weak classifiers, and they are introduced one by one. We can also introduce these weak classifiers one by one into the model (41) in a fashion very similar to AdaBoost, by following Della Pietra, Della Pietra, and Lafferty [14].

Specifically, let $q_n(\mathbf{I})$ be the current model of the form (41) after n weak classifiers are introduced. We can update $q_n(\mathbf{I})$ to a model

$$q_{n+1}(\mathbf{I}) = q_n(\mathbf{I}) \frac{f(\delta(\langle \mathbf{I}, B_{x,y,s,\alpha} \rangle))}{q_n(\delta(\langle \mathbf{I}, B_{x,y,s,\alpha} \rangle))}, \tag{42}$$

by choosing a new weak classifier $\delta(\langle \mathbf{I}, B_{x,y,s,\alpha} \rangle)$, so that $\mathcal{K}(f(\delta)||q_n(\delta))$ is maximized among all (x, y, s, θ). Here $f(\delta)$ and $q_n(\delta)$ are the Bernoulli distributions of the binary variable $\delta(\langle \mathbf{I}, B_{x,y,s,\alpha} \rangle)$ under f and q_n respectively. More specifically, $f(1)$ is the probability that $|\langle \mathbf{I}, B_{x,y,s,\alpha} \rangle| > \xi$ under f, and $f(0) = 1 - f(1)$. $q_n(1)$ is the probability that $|\langle \mathbf{I}, B_{x,y,s,\alpha} \rangle| > \xi$ under q_n, and $q_n(0) = 1 - q_n(1)$. $\mathcal{K}(f(\delta)||q_n(\delta)) = f(1) \log(f(1)/q_n(1)) + f(0) \log(f(0)/q_n(0))$. One can write (42) as $q_{n+1}(\mathbf{I}) = f(\delta) q_n(\mathbf{I} \mid \delta)$, where $q_n(\mathbf{I} \mid \delta)$ is the conditional distribution of $\mathbf{I}$ given $\delta(\langle \mathbf{I}, B_{x,y,s,\alpha} \rangle)$ under q_n.

After choosing $B_{x,y,s,\alpha}$ to maximize $\mathcal{K}(f(\delta)||q_n(\delta))$, we can write (42) as

$$q_{n+1}(\mathbf{I}) = q_n(\mathbf{I}) \exp\{\lambda \delta(\langle \mathbf{I}, B_{x,y,s,\alpha} \rangle)\}/Z(\lambda),$$

where $\lambda = \log(f(1)/q_n(1)) - \log(f(0)/q_n(0))$ and $Z(\lambda) = q_n(0)/f(0)$. Then we let $\lambda_{n+1} \leftarrow \lambda$, $(x_{n+1}, y_{n+1}, s_{n+1}, \alpha_{n+1}) \leftarrow (x, y, s, \alpha)$, $n \leftarrow n + 1$, and iterate.

Let $\mathbf{I}_1, ..., \mathbf{I}_M$ be random samples from $f(\mathbf{I})$. $f(1)$ and $f(0)$ can be estimated as frequencies. If we generate random samples $\mathbf{J}_1, ..., \mathbf{J}_M$ from $q_n(\mathbf{I})$, then $q_n(1)$ and $q_n(0)$ can also be estimated as frequencies. The new weak classifier $\delta(\langle \mathbf{I}, B_{x,y,s,\alpha} \rangle)$ is chosen to tell apart $\mathbf{I}_1, ..., \mathbf{I}_M$ and $\mathbf{J}_1, ..., \mathbf{J}_M$ with maximum $\mathcal{K}(f(\delta)||q_n(\delta)) = \mathcal{K}(q_{n+1}||q_n) = \mathcal{K}(f||q_n) - \mathcal{K}(f||q_{n+1})$.

We can also replace $\mathcal{K}(f(\delta)||q_n(\delta)) = \mathcal{K}(f||q_n) - \mathcal{K}(f||q_{n+1})$ by

$$-\left[\frac{\partial \mathcal{K}(f||q_{n+1})}{\partial \lambda}\right]_{\lambda=0} = f(1) - q_n(1),$$

which is the misclassification rate.

So in each step of the above procedure, a new weak classifier is trained by $\mathbf{I}_1, ..., \mathbf{I}_M$ and $\mathbf{J}_1, ..., \mathbf{J}_M$, where $\mathbf{J}_1, ..., \mathbf{J}_M$ serve as negative examples. Unlike the original AdaBoost, where the samples are reweighted at each step, in the above procedure, the negative examples are sampled at each step from the current model q_n. If the selected bases do not overlap, then $q_n(1)$ and $q_n(0)$ can be calculated in closed form without simulation.

The above procedure is valid for any types of weak classifiers, not limited to the form of $\delta(\langle \mathbf{I}, B_{x,y,s,\alpha} \rangle)$. The reader is also referred to the pioneering work of Della Pietra, Della Pietra, and Lafferty (1997) [14] on introducing features, where model (41) is a special case. We would also like to point out that simultaneous to our work, Tu (2007) [51] independently explored the connection between AdaBoost and the generative model and obtained interesting experimental results.

6. Discussion. This paper studies entropy rate, inferential uncertainty, hypothesis testing, and statistical modeling from the perspective of scaling, which is ubiquitous in natural images. The hope is that the theory presented in this paper will eventually lead to robust and efficient procedures and algorithms for learning and recognizing the whole spectrum of visual patterns and objects. The following are some points we want to make regarding our work.

Nonlinear transforms: The linear filter $\langle \mathbf{I}, B_{x,y,s,\alpha} \rangle$ in the model can be replaced by nonlinear local operators, such as the local gradients or local orientations at different scales. Such nonlinear operators can be built on $\langle \mathbf{I}, B_{x,y,s,\alpha} \rangle$. Generalizing linear filters to nonlinear operators does not affect the validity of the theoretical results derived in the previous sections, in particular, the previous subsection.

Multi-scale analysis and compositional relationships: An image can be analyzed at multiple resolutions [7], or be analyzed by Gabor wavelets $B_{x,y,s,\alpha}$ within a large range of scale. Ideally, image understanding should involve recognizing patterns at multiple resolutions, and these patterns form recursive whole-part compositional relationships [22]. Our results on information scaling also apply to the change of the analysis resolution in addition to the change of the camera resolution and the viewing distance. However, our current model does not account for the compositional relationships of patterns at multiple resolutions. This issue is treated extensively in Zhu and Mumford (2007) [59]. Such relationships are important constraints that help resolve ambiguities in recognizing patterns at multiple resolutions.

Model complexity versus entropy rate: It is important to distinguish between image complexity and model complexity. The complexity of the image data can be measured by the entropy rate, with or without variance-normalization. The complexity of the model can be measured by the dimensionality or the number of parameters in the model. For instance, an image generated by Gaussian white noise has the maximum entropy rate among images with fixed marginal variance. But the Gaussian white noise model is a very simple model with only one parameter for the marginal variance. Although the variance-normalized entropy rate tends to increase over the scaling process, the model complexity does not always increase. As a matter of fact, the image changes from simple regularity such as long straight edges and large smooth regions to simple randomness such as the Gaussian process and white noise. In between, it goes through more complex regularity such as object patterns and more complex randomness such as highly non-Gaussian texture patterns. The model complexity is expected to peak in the mid-resolution, where the entropy rate of the image data is in the medium range. This is the regime where we obtain most of the visual information. Our work on this mid-entropy regime will be reported elsewhere.

Acknowledgement. We thank S. Bahrami for his assistance on the experiments presented in Section 2. We thank Z. Tu and A. Yuille for discussions. A shorter version of the paper has appeared in the *Workshop on Generative Model Based Vision 2004*, organized by A. Pece. The work is supported by NSF DMS-0707055, NSF IIS-0713652, and ONR N-00014-05-1-0543.

References

[1] P. H. Algoet and T. M. Cover, "A sandwich proof of the Shannon-McMillan-Breiman theorem," *Annals of Probability*, 16, 899-909, 1988. MR929085 (89b:94011)

[2] L. Alvarez, Y. Gousseau, and J. M. Morel, "The size of objects in natural and artificial images," *Advances in Imaging and Electron Physics*, 111, 167-242, 1999.

[3] A. R. Barron, "The strong ergodic theorem for densities: Generalized Shannon-McMillan-Breiman theorem," *Annals of Probability*, 13, 1292-1303, 1985. MR806226 (86k:94023)

[4] A. R. Barron, "Entropy and the central limit theorem," *Annals of Probability*, 14, 336-342, 1986. MR815975 (87h:60048)
[5] A. Bell, and T. J. Sejnowski, "The 'independent components' of natural scenes are edge filters," *Vision Research*, 37, 3327-3338, 1997.
[6] J. Besag, "Spatial interaction and the statistical analysis of lattice systems (with discussion)," *Journal of Royal Statistics Society*, B, 36, 192-236, 1974. MR0373208 (51:9409)
[7] P. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Transactions on Communication*, 31, 532-540, 1983.
[8] E. J. Candès and D. L. Donoho, "Curvelets - a surprisingly effective nonadaptive representation for objects with edges," *Curves and Surfaces*, L. L. Schumakeretal. (eds), Vanderbilt University Press, Nashville, TN, 1999.
[9] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8, 679–698, 1986.
[10] T. F. Chan and J. Shen, "Mathematical models for local nontexture inpaintings," *SIAM Journal of Applied Mathematics*, 62(3), 1019-1043, 2001. MR1897733 (2003f:65110)
[11] D. Chandler, *Introduction to Modern Statistical Mechanics*, The Clarendon Press, Oxford University Press, New York, 1987. MR913936 (89d:82001)
[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991. MR1122806 (92g:94001)
[13] J. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *Journal of Optical Society of America*, 2, 1160-1169, 1985.
[14] S. Della Pietra, V. Della Pietra, and J. Lafferty, "Inducing features of random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4), 380–393, 1997.
[15] D. L. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies, "Data compression and harmonic analysis," *IEEE Trans. Information Theory*. 6, 2435-2476, 1998. MR1658775 (99i:94028)
[16] J. H. Elder and S. W. Zucker, "Local scale control for edge detection and blur estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(7), 699-716, 1998.
[17] D. J. Field, "What is the goal of sensory coding?" *Neural Computation*, 6, 559-601, 1994.
[18] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, 55, 119-139, 1997. MR1473055 (99g:68172)
[19] J. H. Friedman, "Exploratory projection pursuit," *Journal of the American Statistical Association*, 82, 249, 1987. MR883353 (88c:62004)
[20] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741, 1984.
[21] S. Geman and C. Graffigne, "Markov random field image models and their applications to computer vision," *Proceedings of the International Congress of Mathematicians*, 1, 1496-1517, 1987. MR934354
[22] S. Geman, D. F. Potter, and Z. Chi, "Composition system," *Quarterly of Applied Math*, 60(4), 707-736, 2002. MR1939008 (2003i:68129)
[23] U. Grenander, *General Pattern Theory*, The Clarendon Press, Oxford Univ Press, New York, 1993. MR1270904 (96e:68118)
[24] C. Guo, S. C. Zhu, and Y. N. Wu, "Primal sketch: Integrating structure and texture," *Computer Vision and Image Understanding*, 106, 5-19, 2007.
[25] J. Hammersley and P. Clifford, *Markov Fields on Finite Graphs and Lattices*, Preprint, UC. Berkeley, 1968.
[26] D. J. Heeger and J. R. Bergen, "Pyramid based texture analysis/synthesis," *Computer Graphics Proceedings*, 229-238, 1995.
[27] D. Huber and T. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *Journal of Physiology*, 160, 1962.
[28] O. Johnson, "An information theoretical central limit theorem for finitely susceptible FKG systems," technical report, 2004.
[29] T. S. Lee, "Image representation using 2D Gabor wavelets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10, 959-971, 1996.
[30] M. S., Lewicki and B. A. Olshausen, "Probabilistic framework for the adaptation and comparison of image codes," *Journal of the Optical Society of America*, 16(7), 1587-1601, 1999.

[31] T. Lindeberg, *Scale-Space Theory in Computer Vision*, Kluwer Academic Publishers, 1994.
[32] A. Lee, D. Mumford, and J. Huang, "Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model," *International Journal of Computer Vision*, 41(1/2), 35-59, 2001.
[33] S. Mallat, "A theory of multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7), 674-693, 1989.
[34] S. Mallat and Z. Zhang, "Matching pursuit in a time-frequency dictionary," *IEEE Transactions on Signal Processing*, 41, 3397-415, 1993.
[35] B. B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, San Francisco, CA, 1982. MR665254 (84h:00021)
[36] D. Marr, *Vision,* W. H. Freeman and Company, San Francisco, CA, 1982.
[37] S. G. Matheron, *Random Sets and Integral Geometry*, John Wiley and Sons, 1975. MR0385969 (52:6828)
[38] L. Moisan, A. Desolneux, and J.-M. Morel, "Meaningful alignments," *International Journal of Computer Vision*, 40, 1, 7-23, 2000.
[39] D. B. Mumford, "Pattern theory: A unifying perspective," *Proceedings of 1st European Congress of Mathematics*, Birkhäuser-Boston, 1994. MR1341824
[40] D. Mumford and B. Gidas, "Stochastic models for generic images", *Quarterly of Applied Math*, 59(1), 85-111, 2001. MR1811096 (2001m:68166)
[41] C. M. Newman, "Normal fluctuations and the FKG inequalities," *Communications in Mathematical Physics*, 74(2), 119-128, 1980. MR576267 (81i:82070)
[42] 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.
[43] B. A. Olshausen and K. J. Millman, "Learning sparse codes with a mixture-of-Gaussians prior," *Advances in Neural Information Processing Systems*, 12, 841-847, 2000.
[44] S. Osher, A. Sole, and L. Vese, "Image decomposition and restoration using total variation minimization and the H^{-1} norm," *Multiscale Modeling and Simulation: A SIAM Interdisciplinary Journal*, 1(3), 349-370, 2003. MR2030155 (2004k:49004)
[45] A. Pece, "The problem of sparse image coding," *Journal of Mathematical Imaging and Vision*, 17(2), 89-108, 2002. MR1950863 (2004a:94008)
[46] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *International Journal of Computer Vision*, 40(1):49-71, 2000.
[47] D. L. Ruderman and W. Bialek, "Statistics of natural images: Scaling in the Woods," *Physical Review Letters*, 73, 1994.
[48] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, 27, 379-423, 623-656, 1948. MR0026286 (10:133e)
[49] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, 24, 1193-1216, 2001.
[50] A. Srivastava, A. Lee, E. Simoncelli, and S. Zhu, "On advances in statistical modeling of natural images," *Journal of Mathematical Imaging and Vision*, 18(1), 17-33, 2003. MR1966173
[51] Z. Tu, "Learning generative models via discriminative approaches," *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, 2007.
[52] P. A. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, 57(2), 137-154, 2004.
[53] A. Witkin, "Scale-space filtering," *Proceedings of International Joint Conference on Artificial Intelligence*, Karlsruhe, 1983.
[54] Y. N. Wu, S. C. Zhu, and C. Guo, "Statistical modeling of texture sketch," *Proceedings of European Conference of Computer Vision,* 2002.
[55] Y. N. Wu, S. C. Zhu, and X. W. Liu, "Equivalence of Julesz ensemble and FRAME models," *International Journal of Computer Vision*, 38(3), 245-261, 2000.
[56] R. A. Young, "The Gaussian derivative model for spatial vision: I. Retinal mechanism," *Spatial Vision*, 2(4), 273-293, 1987.
[57] S. C. Zhu, C. E. Guo, Y. Z. Wang, and Z. J. Xu, "What are textons?" *International Journal of Computer Vision*, 62(1/2), 121-143, 2005.
[58] S. C. Zhu and D. B. Mumford, "Prior learning and Gibbs reaction-diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(11), 1236-1250, 1997.

[59] S. C. Zhu and D. B. Mumford, "Quest for a stochastic grammar of images," *Foundations and Trends in Computer Graphics and Vision*, to appear.
[60] S. C. Zhu, Y. N. Wu, and D. Mumford, "Minimax entropy principle and its applications in texture modeling," *Neural Computation*, 9(8), 1627-1660, 1997.