Title
From Information Scaling of Natural Images to Regimes of Statistical Models

Authors
Ying Nian Wu
Song-Chun Zhu
Cheng-en Guo

Publication Date
2011-10-25

From Information Scaling of Natural Images to Regimes of Statistical Models

Ying Nian Wu, Song-Chun Zhu, and Cheng-en Guo

Departments of Statistics and Computer Science

University of California, Los Angeles

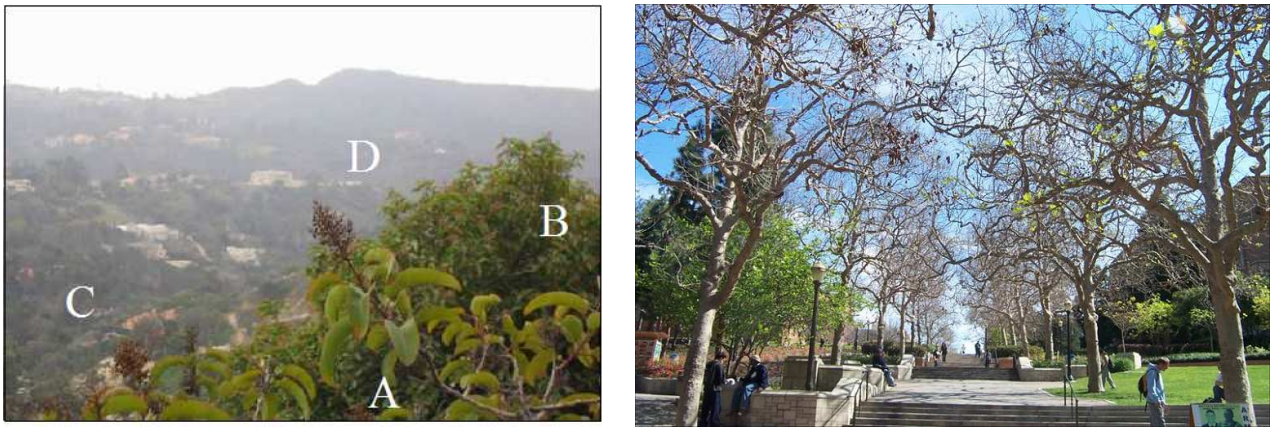
Contact email: ywu@stat.ucla.edu

Abstract

Computer vision can be considered a highly specialized data collection and data analysis problem. We need to understand the special properties of image data in order to construct statistical models for representing the wide variety of image patterns. One special property of vision that distinguishes itself from other sensory data such as speech data is that distance or scale plays a profound role in image data. More specifically, visual objects and patterns can appear at a wide range of distances or scales, and the same visual pattern appearing at different distances or scales produces different image data with different statistical properties, thus entails different regimes of statistical models. In particular, we show that the entropy rate of the image data changes over the viewing distance (as well as the camera resolution). Moreover, the inferential uncertainty changes with viewing distance too. We call these changes information scaling. From this perspective, we examine both empirically and theoretically two prominent and yet largely isolated research themes in image modeling literature, namely, wavelet sparse coding and Markov random fields. Our results indicate that the two models are appropriate on two different entropy regimes: sparse coding targets the low entropy regime, whereas the random fields are suitable for the high entropy regime. Because of information scaling, both models are necessary for representing and interpreting image intensity patterns in the whole entropy range, and information scaling triggers transitions between these two regimes of models. This motivates us to propose a full-zoom primal sketch model that integrates both sparse coding and Markov random fields. In this model, local image intensity patterns are classified into "sketchable regime" and "non-sketchable regime" by a sketchability criterion. In the sketchable regime, the image data are represented deterministically by highly parametrized sketch primitives. In the non-sketchable regime, the image data are characterized by Markov random fields whose sufficient statistics summarize computational results from failed attempts of sparse coding. The contribution of our work is two folded. First, information scaling provides a dimension to chart the space of natural images. Second, the full-zoom modeling scheme provides a natural integration of sparse coding and Markov random fields, thus enables us to develop a new and richer class of statistical models.

1 Introduction

Computer vision can be considered 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 interpret, model, and learn from the image data (Mumford, 1994; Grenander, 1993). However, vision also proves to be a highly specialized data collection and analysis task. We must understand the special characteristics of the image data in order to design adequate models and efficient algorithms.



a b

Figure 1: Image patterns at different distances and scales. (a) Tree leaves at different distances. (b) Twigs and branches of different distances and scales.

One special property of vision that distinguishes itself from other sensory data such as speech data is that distance or scale plays a profound role in image data. More specifically, visual objects and patterns can appear at a wide range of distances or scales. The same visual pattern appearing at different distances or scales produces different image data with different appearances. See Figure (1.a) for an example. It shows tree leaves in four different distance ranges. In region A at near distance, the individual leaves can be perceived. In region B at intermediate distance, the image becomes more complex, and we cannot perceive individual leaves any more. Instead, we only perceive a collective foliage impression. In region C of still farther distance, the image looks like noise. In region D of very far distance, the image

appears to be a smooth region. These local regions have different appearances because the tree leaves appear at different distances, and thus have different sizes on the image. Figure (1.b) shows another example, where tree trunks, branches and twigs appear at different distances and scales, causing different impressions. These two examples show that viewing distance or scale is a key factor in visual perception. In terms of computer vision, the change of distance or scale causes the change of statistical properties of the image intensities, and the change of statistical properties may demand different regimes of statistical models for the image intensities.

In this paper, we study the change of statistical properties, in particular, some information theoretical properties (e.g., Hansen and Yu, 2000), over distance or scale. In particular, we show that the *entropy rate*, defined as entropy per pixel, of the image data changes over the viewing distance (as well as the camera resolution). Moreover, the inferential uncertainty of the underlying visual pattern changes with viewing distance too. We call these changes *information scaling*.

From this perspective, we examine both empirically and theoretically two prominent and yet largely isolated research themes in image modeling literature, namely, wavelet sparse coding (Mallat and Zhang, 1993; Olshausen and Field, 1996; Candes and Donoho, 1999) and Markov random fields (Besag, 1974; Geman and Geman, 1984; Grenander and Miller, 1993). Wavelets originated from harmonic analysis. The key principle is sparsity, where the goal is to find a system of linear basis, so that the class of functions or the ensemble of images can in general be represented or approximated by a small number of linear bases. Markov random fields originated from statistical physics. Instead of coding the image data deterministically with a linear basis, Markov random fields characterize the image data by a set of sufficient statistics.

Our results indicate that the two models are appropriate on two different entropy regimes: sparse coding targets the low entropy regime, whereas the random fields characterize the high entropy regime. Because of information scaling, both models are necessary for representing and interpreting image intensity patterns in the whole entropy range, and information scaling

triggers transitions between the two regimes of models.

This motivates us to propose a full-zoom primal sketch model that integrates both sparse coding and Markov random fields. The term "full-zoom" means that we seek to model image intensity patterns in the full range of scale. The terms "primal sketch" comes from Marr (1982), who, in his monumental book on vision, proposed a symbolic representation of image intensities for the initial stage of visual computation, or low-level vision. In our model, local image intensity patterns are classified into "sketchable regime" and "non-sketchable regime" by a sketchability criterion. In the sketchable regime, the image data are represented deterministically by highly parametrized sketch primitives. In the non-sketchable regime, the image data are characterized by Markov random fields whose sufficient statistics summarize or recycle failed attempts to sketch the image. We fit the full-zoom primal sketch model on natural images, and our experiments suggest that the model captures considerable amount of low-level essence of image data.

The contribution of our work is as follows.

First, the change of image data over distance or scale has been well understood in the literature of scale space theory (Lindeberg, 1994; Mumford and Gidas, 2001). However, the change of statistical properties of the image data over distance or scale, i.e., information scaling, has not been thoroughly studied. Our work on information scaling provides a dimension to chart the space of natural images. Moreover, our work is different from existing results on the statistics of natural images in the literature (Ruderman and Bialek, 1994; Field, 1994; Chi, 2001; Simoncelli and Olshausen, 2001; Sirvastava, Lee, Simoncelli, and Zhu, 2003). Existing results are concerned with the *marginal* statistics while integrating over the scales. Our work, however, is concerned with the *conditional* statistics give the scale, as well as the change of the conditional statistics over 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 the full-zoom modeling scheme provides

a natural integration of these two regimes of models. In our modeling scheme, the random fields summarize the failed attempts of sparse coding.

This paper is intended for the statistical audience. The models, methods, and results reviewed and proposed in this paper are entirely statistical, even though they are mostly developed by researchers in computer vision, neural science and applied mathematics. The plan of the paper is as follows. In Section 2, we will study a simple model treated by Lee, Mumford and Huang (2001) in the context of information scaling. In Section 3, we will prove some theoretical properties on information scaling for general images. Section 4 examines wavelet sparse coding and Markov random fields from the perspective of information scaling. Section 5 presents a full-zoom modeling scheme to cover the whole range of information scaling. Section 6 discusses the limitations of our work, and some directions for future work.

2 Information Scaling of Dead Leaves Model

2.1 The model and the assumptions

To convey the basic idea, we would like to start from the dead leaves model (Matheron, 1975) that has been treated extensively by Lee, Mumford, and Huang (2001) in their investigation of image statistics of natural scenes. The model was also previously used to model natural images by Ruderman (1997) and Alvarez, Gousseau, and Morel (1999), etc. Our use of this model is different from theirs. For our purpose, we may consider that the model describes an ivy wall covered by a large number of leaves of similar sizes.

We assume that the leaves are of squared shape, 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 positions (x, y, z), with the wall serves as the (x, y)-plane, and $z \in [0, z_{\max}]$ is the distance of the leaf from the wall. We assume that $z_{\max}$ is very small, so that z matters only

for deciding the occlusion between the leaves.

For the collection of leaves $\{(r_i, a_i, x_i, y_i, z_i)\}$, we assume that r_i are independent of each other, and so are a_i. (x_i, y_i, z_i) follow a Poisson process in $R^2 \times [0, z_{\max}]$. We assume that the intensity of the Poisson process λ is large enough so that the leaves completely cover the wall. $\{(r_i, a_i, x_i, y_i, z_i)\}$ can be considered a marked point process (Stoyan, Kendall, and Mecke, 1995; Cressie, 1993), where each point (x_i, y_i, z_i) is marked by (r_i, a_i). As noted by Lee et al. (2001), it can also be regarded as 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 measure $f(r)drda\lambda dxdydz$.

Lee et al. (2001) assume that $[r_{\min}, r_{\max}] \to [0, \infty]$. Under scaling transformation, $x' = x/s$ and $y' = y/s$, where s is a scaling parameter, then $r' = r/s$. 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 a measure $f(sr')sdr'da\lambda sdx'sdy'dz$. Under the assumption that $r_{\min} \to 0$ and $r_{\max} \to \infty$, and $f(r) \propto 1/r^3$, the Poisson process is invariance under scaling. This assumption seems to apply to most of the studies of statistics of natural images, such as Ruderman (1994), Field (1994), Chi (2001), Simoncelli and Olshausen (2001), Sirvastava, Lee, Simoncelli, and Zhu (2003), Mandelbrot (1982) and many others.

However, in our experiment, we restrict $[r_{\min}, r_{\max}]$ to a relatively narrow range. Under scaling transformation, the 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 above mentioned authors are concerned with the marginal statistics by integrating over the whole range of scale. So the marginal statistics are invariant under scaling transformation. Our work, however, is concerned with the conditional statistics given a particular scale. The conditional statistics depend on scale, and change over scale. While it is important to look at the marginal statistics of the whole image, it is perhaps even more important to study the conditional statistics in order to model specific image patterns, and the whole range of the conditional statistics may have to be accounted for by different regimes of statistical models.

2.2 Image formation process

Let $T_i \subset \mathbf{R}^2$ be the squared area covered by leaf i in the (x, y) domain of the wall. Then the pattern can be represented by a function $W(x, y) = a_{i(x,y)}$, where $i(x, y) = \arg\max_{i:(x,y)\in T_i} z_i$, i.e., the most forefront leaf that covers (x, y). $W(x, y)$ is a piecewise constant function defined on $\mathbf{R}^2$. For now, we shall assume the wall is infinitely large for convenience.

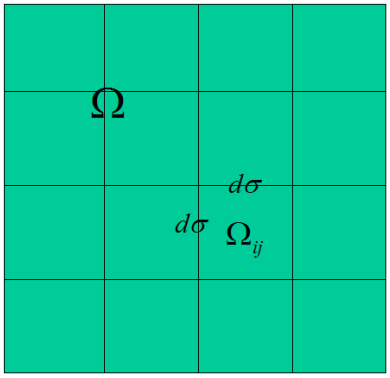


Figure 2: Illustration of image formation: each pixel (i, j) corresponds to a squared window Ω_{ij} in the continuous domain Ω. The size of the window is $d\sigma$, where d is the distance between the pattern and the camera, and σ is the resolution of the camera.

Now let's see what happens if we take 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), a camera (or human eye) only has a finite array of sensors or photoreceptors, each sensor captures lights from a small neighborhood of Ω. As a mathematical model of the image formation process, we may divide the continuous domain Ω into a rectangular array of squared windows of length σd, where σ is decided by the resolution of the camera. Let $\{\Omega_{ij}\}$ be these windows, with $(i, j) \in \Lambda$, where Λ is a rectangular lattice. See Figure (2) for an illustration, where the domain is covered by 4×4 squared windows, so Λ in this case is 4×4.

The image $\mathbf{I}$ is defined on Λ, with

$$\mathbf{I}(i, j) = \int W(x, y)\kappa_{ij}(x, y)dxdy = \langle W, \kappa_{ij}\rangle, \tag{1}$$

where $\kappa_{ij}(x, y)$ is a uniform measure over the squared window Ω_{ij}. One may also replace this uniform density by some other smoothing kernels of bandwidth σd, for instance, a Gaussian kernel whose standard deviation is proportional to σd. See Mumford and Gidas (2001) for more discussions on this issue. In their more general setup, W is a functional, acting on κ_{ij}, which is a test function or a sensor.

Let $s = d\sigma$ be the scale parameter in the above image formation process, which can be written as $\mathbf{I}_s = \gamma_s(W)$ for a functional γ_s. s can be changed by either changing the distance or zooming the camera. If we increase s, the scope Ω and the size of the window will also increase. So the resulting image $\mathbf{I}_s$ will change. Let $W_s(x, y) = W(x/s, y/s)$, then clearly, $\mathbf{I}_s = \gamma_1(W_s)$, i.e., instead of changing the window size, we can also fix the window size for the pixel, but scale the signal W.

Equation (1) can also be written as

$$u_s(x, y) = \int W(x', y')g((x - x')/s, (y - y')/s) = W * g_s, \tag{2}$$

$$\mathbf{I}_s(i, j) = u_s(x_0 + is, y_0 + js), \tag{3}$$

where g_s is a smoothing kernel with bandwidth s, and it corresponds to κ_{ij} in equation (1). There are two operations involved. Equation (2) is smoothing: u_s is a smoothed version of the signal W. Equation (3) is sub-sampling: $\mathbf{I}_s$ is a discrete sampling of u_s.

If g_s is a Gaussian kernel (which is infinitely divisible) with standard deviation s, then the set of $\{u_s(x, y), s > 0\}$ forms a scale space (e.g., Linderberg, 1994). If $s_+ > s$, then $u_{s_+} = u_s * g_{s_+ - s}$, i.e., low resolution image can be obtained from the high resolution image. The scale space has been extensively used in image analysis, where for an image $\mathbf{I}$, we obtain a sequence of smoothed versions by convolving it with Gaussian kernels. This smoothing operation gets ride of small scale details, so that we can concentrate on larger scale structures. The multi-resolution analysis in wavelet theory (Mallat, 1989) also has this smoothing operation and the sub-sampling operation.

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 images under scaling, as well as the change of statistical models for the image data.

2.3 Empirical observations

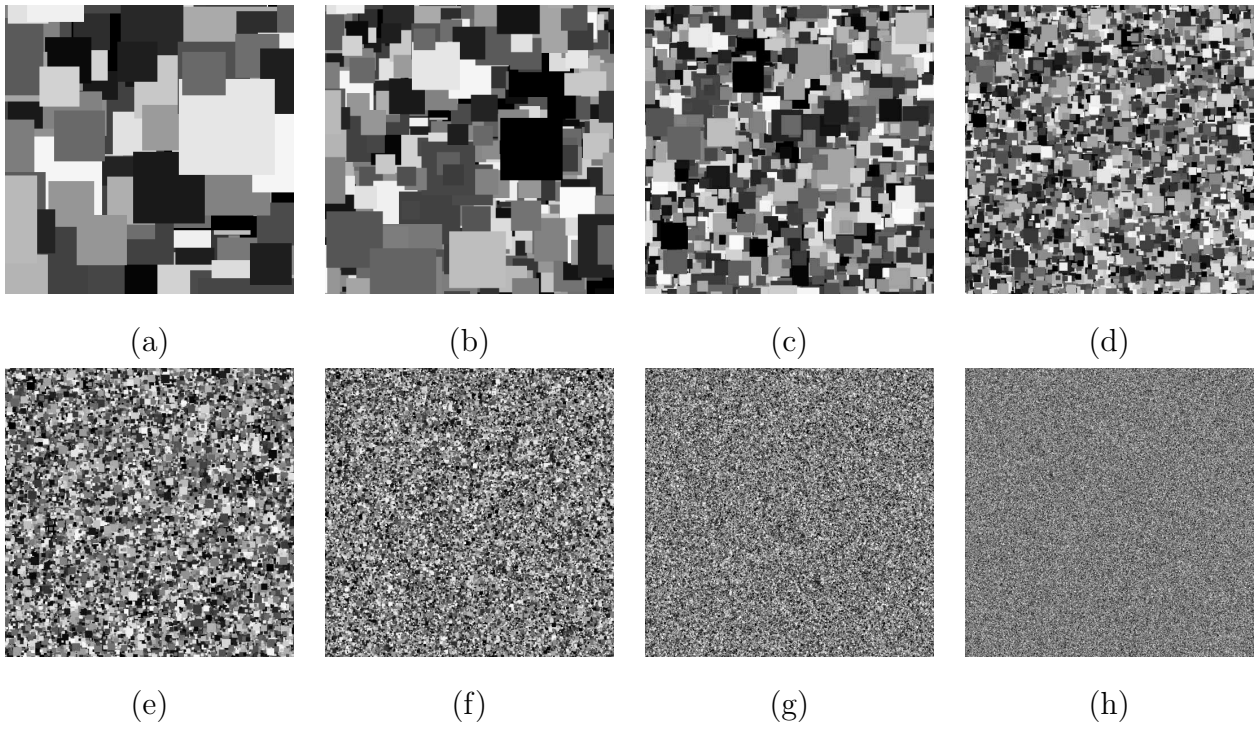


Figure 3: Images of the simulated ivy wall taken at 8 distances, each time, the distance is doubled.

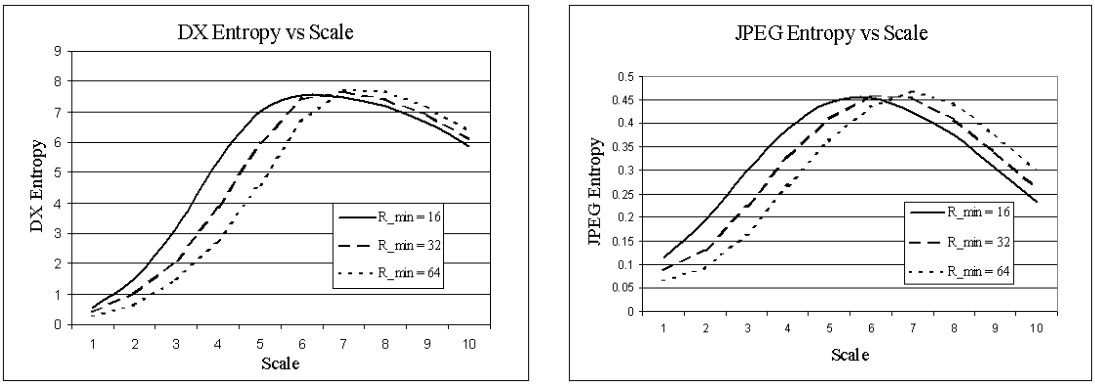
Figure (3) shows a sequence of images taken at 8 distances according to equation (1). Each time we double the distance, so the largest distance is 256 times the smallest distance. Within this wide range of distance, the images show widely different properties even though they are generated by the same signal W. The key is that in $W_s(x, y) = W(x/s, x/y)$, the sizes of leaves changes from r_i to r_i/s, so the distribution of the sizes changes over s.

1) For an image taken at near distance, such as image a), 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 intermediate distance, the window size of a pixel becomes comparable to the average size of leaves, i.e., $s \approx r$. The image becomes more complex. For images (d)

and (e), they cannot be represented by a small number of geometrical structures anymore. The constituent elements have to be pixels themselves. If a simple interpretation of the image is sought after, then 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 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 may cover a large number of leaves, and its intensity value is the local average of the colors of many leaves. The image is approaching the white noise.



(a) (b)

Figure 4: The change of statistical properties over scale. (a) JPEG compression rate. (b) Entropy of marginal histogram of $\nabla_x \mathbf{I}$.

We perform some empirical studies on the change of statistical properties of the image data over distance or scale. What we care most is the complexity or randomness of the image, and we measure the randomness empirically by JPEG 2000 compression rate. We compress the image using JPEG 2000, which is based on wavelet decomposition. Then we compute the number of bits we need to compress the image, and divide it by the number of pixels. That gives us the compression rate in terms of bits per pixel, and it can serve as an indicator of complexity or randomness of the image. We plot this indicator over distance. See Figure (4.a). We can see that at 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 far distance, however, the randomness begins to decrease, because the local averaging effect takes over. 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. 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 $\nabla_y \mathbf{I}(i,j) = \mathbf{I}(i,j) - \mathbf{I}(i,j-1)$. Then $\nabla \mathbf{I}(i,j) = (\nabla_x \mathbf{I}(i,j), \nabla_y \mathbf{I}(i,j))$ is the gradient of $\mathbf{I}$ at (i,j). The gradient is a very useful local feature, and can be used for edge detection (Canny, 1986). In our simulation study, we look at the statistics of $\nabla_x \mathbf{I}$. We make a marginal histogram of $\{\nabla_x \mathbf{I}(i,j), (i,j) \in \Lambda\}$. We then compute the entropy of the histogram. We plot this entropy over distance. See Figure (4.b). We can see that the plot behaves similarly as the plot of the JPEG 2000 compression rate.

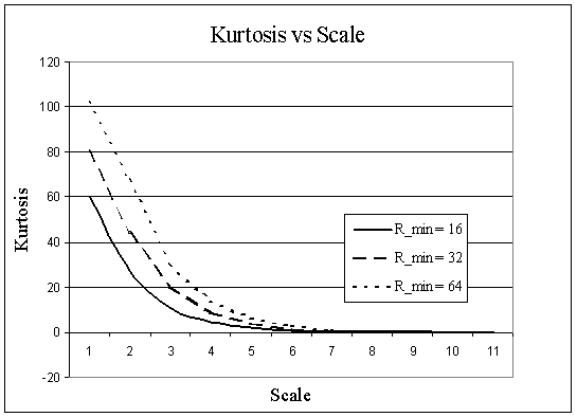


Figure 5: The change of kurtosis over scale.

The local averaging operation in equation (1) pushes the marginal distribution of the image intensities towards the Gaussian distribution because of the central limit theorem. See a recent paper of Johnson (2004) on an information-theoretical centeral limit theorem of random fields, which applies to the situation here. We compute the kurtosis of the marginal empirical distribution of the image intensities to measure how close the marginal distribution

is to the Gaussian distribution. We can see that the kurtosis is decreasing. Meaning that the image feature becomes closer to Gaussian distribution.

The local averaging operation reduces the variance. If we normalize the marginal variance of the image intensities to 1, then both the compression rate and the smoothness increase over the entire distance range. In fact, the image will eventually become white noise, which is the maximum entropy distribution under fixed marginal mean and variance of image intensities.

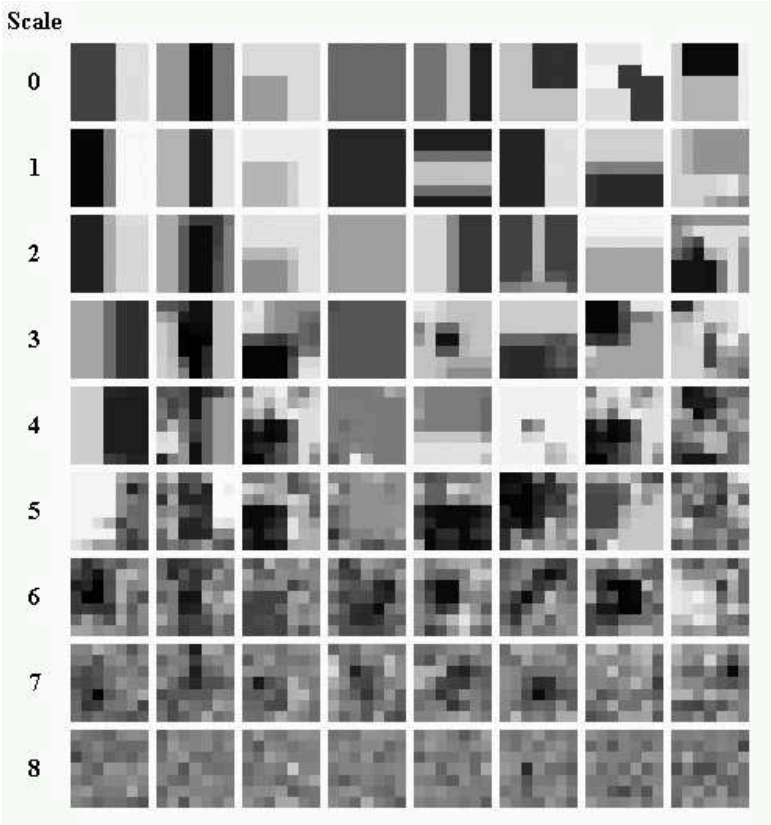


Figure 6: The 7×7 local patches taken from the images at different scales.

Computer vision algorithms always start from local processing and representation (Geman and Koloydenko, 1999). We take some local 7×7 image intensity patches from the images at different scales. 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 operators in a computer vision system should be prepared to deal with such local image patches with different regularities and randomness.

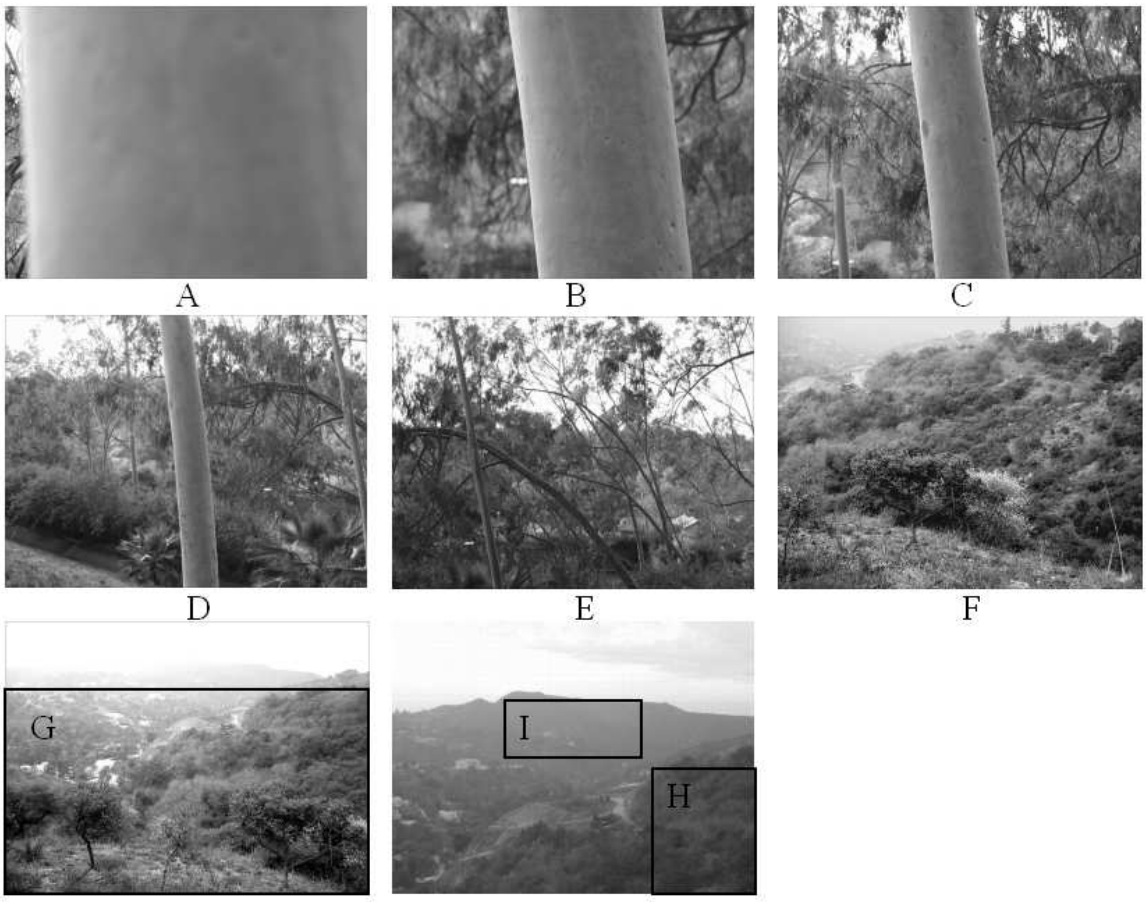


Figure 7: Natural images taken at different distances from the trees.

We also did some experiments on natural images. Figure (7) shows a sequence of images taken at increasing distances from the tree. Figure (8) displays the change of randomness measured by three indicators. The red dashed line is the JPEG compression rate. The blue solid line is the smoothness, i.e., the entropy of $\nabla_x \mathbf{I}$. For the black dotted line, we code the image as the linear expansion of a set of local linear bases selected from a vocabulary. We then record the number of bases we need to include in order to reduce the mean squared error to 30% of the variance of the original image. See Section 4 for more details. We linearly normalize the three indicators so that they fit into the same plot. We can see that the change of randomness in Figure (7) is consistent with the simulated example.

We also did the same experiment for the pictures in Figure (9). Here we have an image

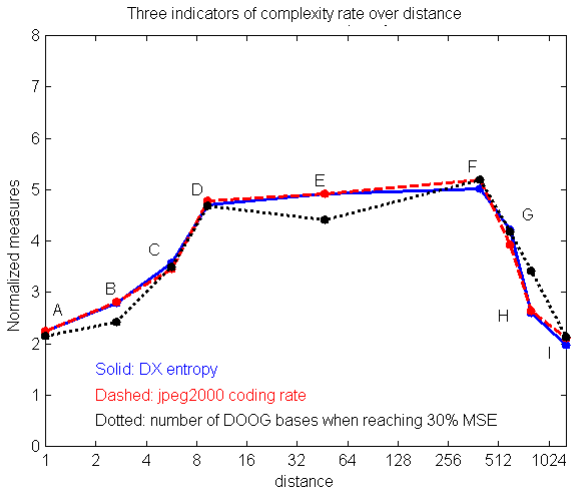


Figure 8: The change of the randomness over distance in Figure 7.

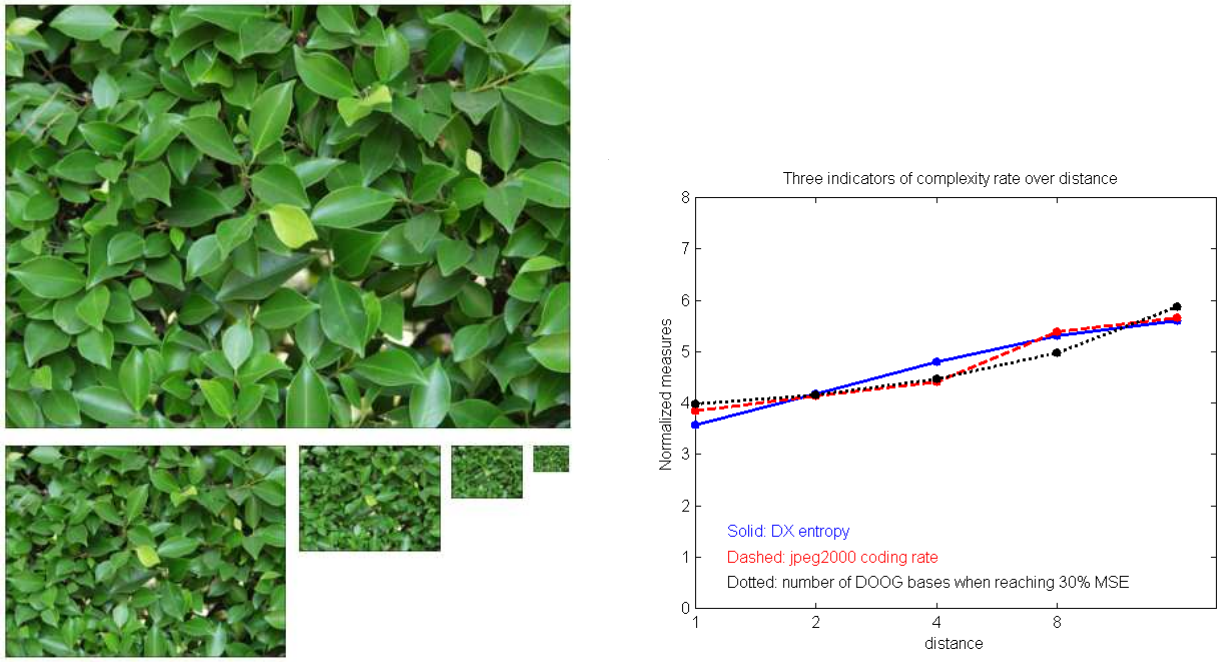


Figure 9: The figure on the left displays the original image of ivy wall and its downscaled versions. The plot on the right shows the change of randomness over the order of downscaling.

of ivy wall and its downscaled versions. We can see that the randomness keeps increasing, because the sequence of images does not cover the whole range of scale.

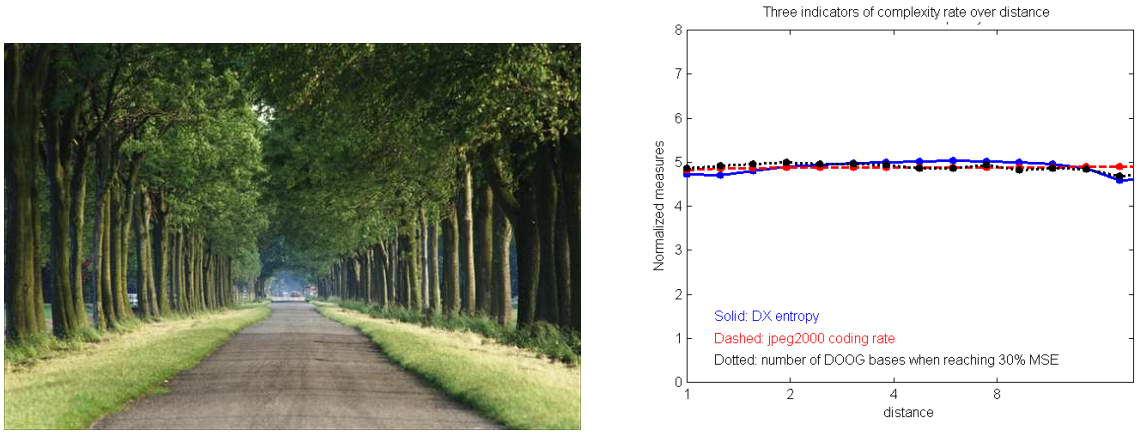


Figure 10: A scale invariant image and the change of randomness over the order of downscaling.

Finally, we repeat the same experiment for the picture in Figure (10) and its downscaled versions (not shown). The picture appears to be scale invariant, and the randomness does not change much across the scale.

In Figure (9), $[r_{\min}, r_{\max}]$ is very small, so we see clear change of randomness over scale. In Figure (10), $[r_{\min}, r_{\max}]$ is much larger, and the image is a mixture of objects and patterns of different distances or scales. So we need to be able to model the image patterns in the whole range of distances or scales in order to model images like the one in Figure (10).

3 Information Scaling

3.1 Notation

In this article, we will make use of the following information theoretical concepts (see Cover and Thomas, 1994; Rissanen, 1989; Hansen and Yu, 2000):

1) *Entropy*: for a distribution $p(x)$, its entropy is

$$\mathcal{H}(p) = -\mathrm{E}[\log p(x)] = -\int p(x) \log p(x) dx.$$

If $p(x)$ is a discrete distribution, then the integral becomes sum. The entropy measures the randomness of p. It also measures the average description length if we are to code the

instances generated by $p(x)$. This entropy is called Shannon entropy.

2) *Conditional entropy*: for a joint distribution $p(x, y)$, let $p(x|y)$ be the conditional distribution of x given y. The conditional entropy of x given y is defined as

$$\mathcal{H}(p(x|y)) = -\mathrm{E}[\log p(x|y)] = -\int p(x, y) \log p(x|y) dx dy,$$

where the integral (or sum in discrete case) is over both x and y.

3) *Kullback-Leibler divergence*: for any two distributions $p(x)$ and $q(x)$, the Kullback-Leibler divergence or distance is

$$\mathcal{D}(p\|q) = \mathrm{E}_p[\log p(x)/q(x)] = \int \log \frac{p(x)}{q(x)} p(x) dx \geq 0.$$

If the random variables are discrete, then the integral should be replaced by the sum.

4) *Mutual information:* if $X = (x_1, ..., x_d)$, then the mutual information among the d components are

$$\mathcal{M}(x_1, ..., x_d) = \mathcal{D}(p(X)\| \prod_{i=1}^{d} p_i(x_i)) = \mathrm{E}_p \left[\log \frac{p(X)}{\prod_{i=1}^{d} p_i(x_i)} \right] \geq 0,$$

i.e., the Kullback-Leibler distance between the joint distribution $p(X)$ and the product of the marginal distributions $p_1(x_1)$, ..., $p_d(x_d)$.

The book by Rissanen (1989) discusses other notions of complexity, and describes the minimum description length for statistical modeling. See also Hansen and Yu (2000) and Lee (2001).

3.2 Complexity scaling

For simplicity, let's study what happens if we double the distance between the camera and the visual pattern. Suppose the image of the visual pattern at near distance is $\mathbf{I}(i, j), (i, j) \in \Lambda$. If we double the distance, according to our previous discussions, the window of a pixel will also double its length and width. So the original $\mathbf{I}$ will be reduced to a smaller image $\mathbf{I}_-$ defined on a reduced lattice Λ_-, and each pixel of $\mathbf{I}_-$ will be the average of four pixels of $\mathbf{I}$ (assuming the windows are aligned at the two distances). We call this process *downscaling*, and we can account for it by two steps, similar to equations (2) and (3).

1) Local smoothing. Let the smoothed image be $\mathbf{J}$, then $\mathbf{J}(i,j) = [\mathbf{I}(i,j) + \mathbf{I}(i+1,j) + \mathbf{I}(i,j+1) + \mathbf{I}(i+1,j+1)]/4$. In general, we can convolve $\mathbf{I}$ with a smoothing kernel g (or a density function such as Gaussian in the scale space theory) to get $\mathbf{J}$, i.e., $\mathbf{J} = \mathbf{I} * g$.

2) Subsampling. $\mathbf{I}_{-}^{(t_x,t_y)}(i,j) = \mathbf{J}(2i+t_x, 2j+t_y)$, where $(t_x,t_y) \in \{(0,0),(0,1),(1,0),(1,1)\}$. Any of the four $\mathbf{I}_{-}^{(t_x,t_y)}$ can be regarded a downsampled version of $\mathbf{J}$.

We use entropy rate to quantify the randomness of an image (or a part of an image). For an image $\mathbf{I} \sim p(\mathbf{I})$ defined on lattice domain Λ, its entropy rate is defined as $\overline{\mathcal{H}}(p(\mathbf{I})) = \mathcal{H}(p(\mathbf{I}))/|\Lambda|$, i.e., entropy per pixel.

Let's first study the effect of local smoothing, $\mathbf{J} = \mathbf{I} * g$.

Theorem 1 *Smoothing effect: As the lattice $\Lambda \to Z^2$,*

$$\overline{\mathcal{H}}(p(\mathbf{J})) - \overline{\mathcal{H}}(p(\mathbf{I})) \to \int \log|\hat{g}(\omega)| d\omega \leq 0, \tag{4}$$

where $\hat{g}$ is the Fourier transform of the kernel g, $\omega \in [-\pi/2, \pi/2] \times [-\pi/2, \pi/2]$ is the spatial frequency.

Proof: In the Fourier domain, we have $\hat{\mathbf{J}}(\omega) = \hat{\mathbf{I}}(\omega)\hat{g}(\omega)$, where $\hat{\mathbf{J}}$ and $\hat{\mathbf{I}}$ are Fourier transforms of $\mathbf{J}$ and $\mathbf{I}$ respectively. For finite rectangular lattice Λ, the spatial frequency ω takes values in a finite grid. Since the Fourier transform is orthogonal, we have $\mathcal{H}(p(\mathbf{I})) = \mathcal{H}(p(\hat{\mathbf{I}}))$, and $\mathcal{H}(p(\mathbf{J})) = \mathcal{H}(p(\hat{\mathbf{J}}))$. Thus

$$\frac{1}{|\Lambda|}\mathcal{H}(p(\mathbf{J})) = \frac{1}{|\Lambda|}\mathcal{H}(p(\mathbf{I})) + \frac{1}{|\Lambda|}\sum_{\omega} \log|\hat{g}(\omega)|.$$

As $\Lambda \to Z^2$, the second term on the right hand side goes to $\int \log|\hat{g}(\omega)| d\omega$.

A smoothing kernel g is a probability distribution function, $\hat{g}$ is the characteristic function of g, and $\hat{g}(\omega) = \sum_x g(x) e^{-i\omega x} = \mathrm{E}_g[e^{-i\omega x}]$, with $x \sim g(x)$. So $|\hat{g}(\omega)|^2 = |\mathrm{E}_g[e^{-i\omega x}]|^2 \leq \mathrm{E}_g[|e^{-i\omega x}|^2] = 1$. Thus, $\int \log|\hat{g}(\omega)| d\omega \leq 0$. QED

Remark: the above theorem tells us that we always lose information under the smoothing operation. This is consistent with the intuition in scale space theory, which holds that increasing the scale in the scale space will result in the loss of fine details in the image.

Next, let's study the effect of subsampling. According to our discussion before, there are four sub-sampled versions, $\mathbf{I}_{-}^{(t_x,t_y)}(i,j) = \mathbf{J}(2i+t_x, 2j+t_y)$, where $(t_x,t_y) \in \{(0,0),(0,1),(1,0),(1,1)\}$. For simplicity, let's denote them by $\mathbf{I}_{-}^{(k)}, k = 1,2,3,4$, each defined on a subsampled lattice Λ_{-}, with $|\Lambda_{-}| = |\Lambda|/4$. See Figure (11) for an illustration.

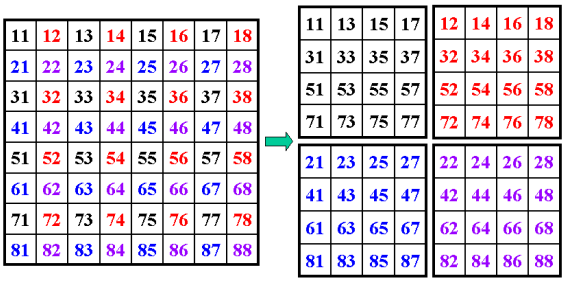


Figure 11: The four subsampled versions of the original image.

Theorem 2 *Subsampling effect: The entropy rate of* $\mathbf{I}_{-}^{(k)}$ *is no less than the entropy rate of* $\mathbf{J}$*,*

$$\frac{1}{4}\sum_{k=1}^{4}\overline{\mathcal{H}}(p(\mathbf{I}_{-}^{(k)})) - \overline{\mathcal{H}}(p(\mathbf{J})) = \frac{1}{|\Lambda|}\mathcal{M}(\mathbf{I}_{-}^{(k)}, k=1,...,4) \geq 0, \tag{5}$$

where $\mathcal{M}() = \mathrm{E}\left[\log p(\mathbf{J})/\prod_k p(\mathbf{I}_{-}^{(k)})\right]$ *denotes mutual information among the four down-sampled versions.*

Proof:

$$\begin{aligned}\sum_{k=1}^{4}\mathcal{H}(p(\mathbf{I}_{-}^{(k)})) - \mathcal{H}(p(\mathbf{J})) &= \mathrm{E}\left[\log\frac{p(\mathbf{J})}{\prod_k p(\mathbf{I}_{-}^{(k)})}\right] \\ &= \mathcal{M}(\mathbf{I}_{-}^{(k)}, k=1,...,4) \geq 0. \ \text{QED}\end{aligned}$$

The complexity scaling is a combination of equations (4) and (5):

$$\left\{\frac{1}{4}\sum_{k=1}^{4}\overline{\mathcal{H}}(p(\mathbf{I}_{-}^{(k)})) - \overline{\mathcal{H}}(p(\mathbf{I}))\right\} - \left\{\frac{1}{|\Lambda|}\mathcal{M}(\mathbf{I}_{-}^{(k)}, k=1,2,3,4) + \int \log|\hat{g}(\omega)|d\omega\right\} \to 0.$$

For regular image patterns, the mutual information per pixel can be much greater than $-\int \log|\hat{g}(\omega)|d\omega$, 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 to $-\int \log |\hat{g}(\omega)| d\omega$, we have scale invariance. More careful analysis is needed to determine when this is true.

Since local averaging reduces the marginal variance of the image intensities, if we renormalize the image intensities to keep the marginal variance constant, we will increase the entropy rate of the image. Therefore, renormalization cancels the effect of local averaging. So for renormalized image, we expect that the entropy rate keeps increasing.

There is another important notions of entropy, the Kolmogorov algorithmic complexity. For an image $\mathbf{I}$ (properly discretized), its algorithmic complexity $\mathcal{H}_{\text{alg}}(\mathbf{I})$ can be defined as the shortest binary code of a Turing machine to generate $\mathbf{I}$. See Rissanen (1989) for more details. Then we can also defined the algorithm complexity rate $\overline{\mathcal{H}}_{\text{alg}}(\mathbf{I}) = \mathcal{H}_{\text{alg}}(\mathbf{I})/|\Lambda|$. Obviously, the length of the shortest code for $\mathbf{J}$ is equal to or smaller than the sum of the lengths of the shortest codes for $\mathbf{I}_{-}^{(k)}$, $k = 1, ..., 4$, so we have the following

Proposition 1 *The algorithmic complexity rate increases with sub-sampling, i.e.,*

$$\frac{1}{4}\sum_{k=1}^{4} \overline{\mathcal{H}}_{\text{alg}}(\mathbf{I}_{-}^{(k)}) \geq \overline{\mathcal{H}}_{\text{alg}}(p(\mathbf{J})).$$

One can define mutual algorithmic information as the difference between the left hand side and the right hand side.

The change of entropy rate of the image data over distance can be used to explain the transition from a deterministic interpretation to statistical interpretation of the image intensities. We only need to postulate 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 parameters deterministically. But if the local image patch has a high entropy rate, a small number of parameters 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 Perceptibility scaling

The above analysis on complexity 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 and patterns in the outside world. In this section, we shall go beyond the statistical properties of the observed image itself, and study the interaction between the observed image and the outside world 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 range $D \subset \mathbf{R}^2$, and let $\mathbf{W} = ((x_i, y_i, r_i, a_i), i = 1, ..., N)$ be the leaves in D that are not completely occluded by other leaves. Then we have $\mathbf{W} \sim p(\mathbf{W})$, and $\mathbf{I}_s = \gamma_s(\mathbf{W})$, where $s = d\sigma$ is the scale parameter in the image formation process.

For convenience, assume that both $\mathbf{W}$ and $\mathbf{I}_s$ are properly discretized. Then the marginal distribution of $\mathbf{I}_s$ is $p(\mathbf{I}_s) = \sum_{\mathbf{W}:\gamma_s(\mathbf{W})=\mathbf{I}_s} p(\mathbf{W})$. The posterior distribution of $\mathbf{W}$ given $\mathbf{I}_s$ is $p(\mathbf{W}|\mathbf{I}_s) = p(\mathbf{W})/p(\mathbf{I}_s)$ because $\mathbf{I}_s$ is fully determined by $\mathbf{W}$. We can define the entropy of $p(\mathbf{W}|\mathbf{I}_s)$ as the *imperceptibility* of $\mathbf{W}$ from the image $\mathbf{I}_s$.

Proposition 2 *In the above notation,* $\mathcal{H}(p(\mathbf{W}|\mathbf{I}_s)) = \mathcal{H}(p(\mathbf{W})) - \mathcal{H}(p(\mathbf{I}_s))$. *That is, imperceptibility = scene entropy - image entropy.*

Remark: Here we may consider the inference of $\mathbf{W}$ as inverting $\mathbf{I}_s = \gamma_s(\mathbf{W})$ under the prior knowledge $\mathbf{W} \sim p(\mathbf{W})$. The imperceptibility can be considered a measure of how ill-posed the inversion problem is.

If we have images of $\mathbf{W}$ at two scales, $\mathbf{I}_{s_+}$ and $\mathbf{I}_s$ with $s_+ > s$, according to scale space theory, $\mathbf{I}_{s_+}$ can be obtained from $\mathbf{I}_s$ by a deterministic downscaling transformation. Since $\mathbf{I}_{s_+}$ is of lower dimension than $\mathbf{I}_s$, this transformation is a many to one reduction. During the process of image scaling, the overall entropy of the image will decrease (even though the entropy per pixel can increase as we show in the previous subsection). Therefore, we have the following result.

Proposition 3 *Imperceptibility increases with downscaling, i.e.,* $\mathcal{H}(p(\mathbf{W}|\mathbf{I}_{s_+}) \geq \mathcal{H}(p(\mathbf{W}|\mathbf{I}_s))$ *for* $s_+ > s$*, if* $\mathbf{I}_{s_+} = R(\mathbf{I}_s)$ *for a many to one reduction* $R()$*.*

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

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

Then what should we do? The regime of $s > s_{\text{bound}}$ proves to be most baffling for vision modeling. Our knowledge about geometry, optics, and mechanics enables us to model every phenomenon in our visual environment. Such models may be sufficient for computer graphics to generate physically realistic images. For instance, researchers in graphics can generate a garden scene by placing billions of leaves and grass strands under perspective geometry. They can generate a river scene, a fire scene or smoke scene using computational fluid dynamics. They can generate clothes using a set of particles under the law of mechanics. They can generate sophisticated lighting using 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 image formation process, it is a big 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 patterns, so that the computer vision system can interpret the observed image with comparable sophistication to human vision.

The following are some of our simple theoretical considerations of this problem from the perspectives of recognition and description. We shall get into more depth on the modeling issue in later sections.

For recognition, instead of pursuing a detailed description $\mathbf{W}$, we may choose to estimate some rough summary of $\mathbf{W}$. For instance, in the simulated ivy wall example, we may care about properties of the overall distribution of colors of leaves, as well as the overall distribution of their sizes, etc. Let's call it $\mathbf{W}_- = \rho(\mathbf{W})$, with ρ being a many to one reduction function. It is possible that we can estimate $\mathbf{W}_-$ because of the following result.

Proposition 4 *For* $\mathbf{W} \sim p(\mathbf{W})$, $\mathbf{I}_s = \gamma_s(\mathbf{W})$, *and* $\mathbf{W}_- = \rho(\mathbf{W})$, *we have*

$$\begin{aligned} &1)\ \ \mathcal{H}(p(\mathbf{W}_-|\mathbf{I}_s)) \leq \mathcal{H}(p(\mathbf{W}|\mathbf{I}_s)). \\ &2)\ \ p(\mathbf{I}_s|\mathbf{W}_-) = \frac{\sum_{\mathbf{W}:\rho(\mathbf{W})=\mathbf{W}_-;R(\gamma(\mathbf{W}))=\mathbf{I}_s} p(\mathbf{W})}{\sum_{\mathbf{W}:\rho(\mathbf{W})=\mathbf{W}_-} p(\mathbf{W})}. \end{aligned}$$

Result 2) tells us that although $\mathbf{W}$ defines $\mathbf{I}_s$ deterministically via $\mathbf{I}_s = \gamma_s(\mathbf{W})$, $\mathbf{W}_-$ may only define $\mathbf{I}_s$ statistically via a probability distribution $p(\mathbf{I}_s|\mathbf{W}_-)$. While $\mathbf{W}$ represents deterministic structures, $\mathbf{W}_-$ may only represent some statistical properties. Thus, we have a transition from a deterministic representation of the image intensities $\mathbf{I}_s = \gamma_s(\mathbf{W})$ to a statistical representation $\mathbf{I}_s \sim p(\mathbf{I}_s|\mathbf{W}_-)$.

It is also possible that for an image $\mathbf{I}_s$ of large s, we may just summarize it by some $F(\mathbf{I}_s)$, so that the summary $F(\mathbf{I}_s)$ contains as much information about $\mathbf{I}_s$ as possible as far as $\mathbf{W}$ or $\mathbf{W}_-$ is concerned.

Proposition 5 *Let $F = F(\mathbf{I}_s)$ be a summary of $\mathbf{I}_s$,*

1) If $\mathbf{W} \sim p(\mathbf{W})$, $\mathbf{I}_s = \gamma_s(\mathbf{W})$, then

$$\begin{aligned}\mathcal{D}(p(\mathbf{W}|\mathbf{I}_s)||p(\mathbf{W}|F)) &= \mathrm{E}_{\mathbf{W}}\left[\log \frac{p(\mathbf{W}|\mathbf{I}_s)}{p(\mathbf{W}|F)}\right] \\ &= \mathcal{H}(\mathbf{W}|F) - \mathcal{H}(\mathbf{W}|\mathbf{I}_s) = \mathcal{H}(\mathbf{I}_s|F).\end{aligned}$$

2) If $\mathbf{W}_- \sim p(\mathbf{W}_-)$ and $[\mathbf{I}_s|\mathbf{W}_-] \sim p(\mathbf{I}|\mathbf{W}_-)$, then

$$\begin{aligned}\mathcal{D}(p(\mathbf{W}_-|\mathbf{I}_s)||p(\mathbf{W}_-|F)) &= E_{\mathbf{W}_-,\mathbf{I}_s}\left[\log \frac{p(\mathbf{W}_-|\mathbf{I}_s)}{p(\mathbf{W}_-|F)}\right] \\ &= \mathcal{H}(p(\mathbf{W}_-|F)) - \mathcal{H}(\mathbf{W}_-|\mathbf{I}_s) = \mathcal{M}(\mathbf{W}_-, \mathbf{I}_s|F).\end{aligned}$$

Here $\mathcal{D}()$ denotes Kullback-Leibler divergence, and $\mathcal{M}()$ denotes mutual information.

Remark: Result 1) tells us that for $F(\mathbf{I}_s)$ to contain as much information about $\mathbf{W}$ as possible, we want to make $\mathcal{H}(\mathbf{I}_s|F)$ as small as possible. Result 2) tells us that if we want to estimate some $\mathbf{W}_-$, then we want F to be sufficient about $\mathbf{I}_s$ as far as $\mathbf{W}_-$ is concerned. $\mathcal{M}(\mathbf{W}_-, \mathbf{I}_s|F)$ can be considered a measure of sufficiency.

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

Proposition 6 *The average coding length is $\mathrm{E}_p[\mathcal{H}(f(w|\mathbf{I}_s))]+\mathcal{D}(p||f)+\mathcal{H}(p)$. That is, coding redundancy = imperceptibility + error. Here $\mathcal{H}(f(w|\mathbf{I}_s))$ is the entropy of $f(w|\mathbf{I}_s)$ conditional on $\mathbf{I}_s$, and $D(p||f)$ is the Kullback-Leibler distance.*

Remark: The above proposition provides a selection criterion for models with latent variables. It can be considered a generalization of the minimum description length of Rissanen (1989), see also Hansen and Yu (2000), and Lee (2001). The imperceptibility term comes

up because we assume a coding scheme where w must be coded first, and then $\mathbf{I}_s$ is coded based on w. Given the latent variable structure of the model, it is very natural to assume such a coding scheme.

4 Two modeling schemes and entropy analysis

4.1 Modeling natural image patterns

The reason we study information scaling is because we want to model natural image patterns. In natural scene images, there are bewildering varieties of patterns. Just like the wide varieties of physical phenomena can be described by simple and unified physics laws, one may ask whether we can find simple and unified mathematical and statistical models to describe the wide variety of visual patterns in natural environmental scenes.

In the last section, we show that changing the distance between the camera and the visual pattern will cause the change of the entropy rate of the resulting observed image intensities, which in turn may trigger the transition between deterministic representation and statistical interpretation of the image intensities. In this section, we will get more concrete in terms of modeling, and examine existing image models and their empirical and theoretical properties from the perspective of entropy. Before doing that, we shall first briefly describe the mathematical models of simple neuron cells in primitive visual cortex. Since these cells perform the first step of visual computation, they will shed light on image modeling at the early stage of visual processing, or low-level vision.

4.2 Simple visual cells and Gabor wavelets

Hubel and Wiesel (1962), in their Nobel prize winning work, discovered that simple neuron cells in cat's primitive visual cortex (or what is called V1 area) selectively respond to visual stimuli such as bars and edges at different locations, scales, and orientations. Daugmann (1980) 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 functions:

$$G(x,y) \propto \frac{1}{\sigma_x \sigma_y} \exp\{-\frac{x^2}{2\sigma_x^2}\} e^{i\omega x}, \tag{6}$$

which is essentially a pair of local sine and cosine waves propagating along the x-axis, where the localization is achieved by multiplying the waves with Gaussian functions. Such Gabor functions fit the observed data on neuron cells reasonably well. In the fitted model, the σ_y is larger than σ_x, so it is elongated along the y-axis. ω_x and σ_x are such that the amplitude of the sine or cosine wave decays to 0 very quickly, so essentially only one cycle of the wave survive.

Another model (e.g., Malik and Perona, 1989) is the derivatives of Gaussian filters,

$$G(x,y) \propto \frac{\partial^k}{\partial x^k} \frac{1}{\sigma_x \sigma_y} \exp\{-\frac{x^2}{2\sigma_x^2}\}, \tag{7}$$

where $k = 1$ and 2, i.e., the first and second derivatives of an elongate Gaussian. The function (7) is similar to Gabor function in (6), in particular, the first derivative in (7) is similar to Gabor sine component, and the second derivative is similar to Gabor cosine component. The derivative of Gaussian filter is essentially a gradient operator.

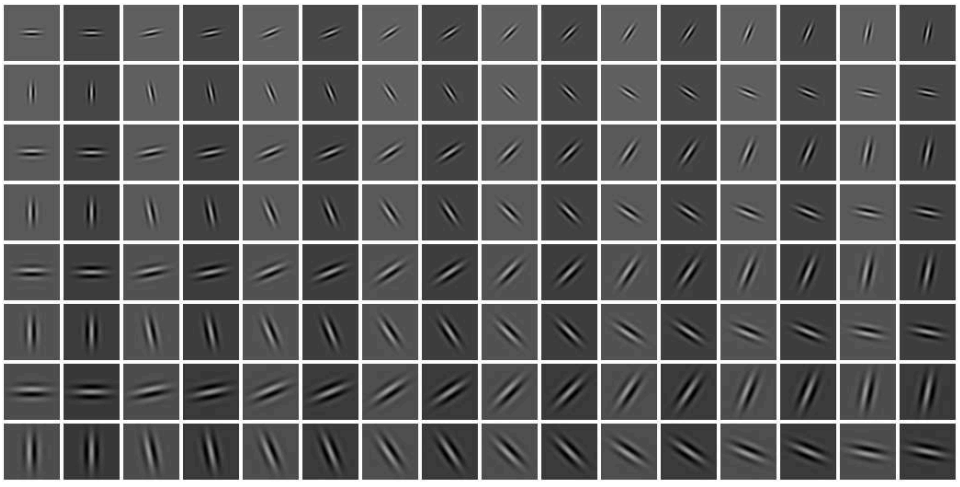


Figure 12: Gabor functions of different scales and orientations.

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

$$G_{s,\theta}(x,y) = G\left([x\cos\theta + y\sin\theta]/s, [-x\sin\theta + y\cos\theta]/s\right).$$

See Figure (12) for an illustration of a set of Gabor filters of different scales and orientations. Finally, we can translate $G_{s,\theta}$ to make it centered at (x,y), $B_{x,y,s,\theta}(x',y') = G_{s,\theta}(x'-x, y'-y)$.

For a function $\mathbf{I}(x,y)$ defined on $\mathbf{R}^2$, one can define the inner product or filter response as

$$r_{x,y,s,\theta} = \langle \mathbf{I}, B_{x,y,s,\theta} \rangle = \int\int I(x',y') B_{x,y,s,\theta}(x',y') dx'dy'. \tag{8}$$

That is, we project $\mathbf{I}$ on $B_{x,y,s,\theta}$, which can be considered a linear base in the image space. For an image $\mathbf{I}$ defined on a discrete lattice, we can discretize the (x,y) domain of $B_{x,y,s,\theta}$, and replace the integral by sum. In what follows, we will stick to the notation (x,y) instead of (i,j) for labeling the pixels.

These linear operators detect elongate local image structures such as edges and bars. Also, they are overcomplete, in the sense that the number of filters $\{B_{x,y,s,\theta}\}$ is much larger than the dimensionality of the image $\mathbf{I}$. Lee (1996) designed an overcomplete set of Gabor bases $\{B_{x,y,s,\theta}\}$ that form a so-called tight frame. That is, for $r_{x,y,s,\theta}$ computed as equation (8), we can reconstruct the image $\mathbf{I}$ by

$$\mathbf{I} = \sum_{x,y,s,\theta} r_{x,y,s,\theta} B_{x,y,s,\theta}, \tag{9}$$

even though $\{B_{x,y,s,\theta}\}$ is overcomplete, and does not form an orthogonal basis. There is biological evidence that the visual cells in cat's primitive visual cortex are close to a tight frame.

One may ask, what is the purpose of these visual cells? Whether we can make use of them for constructing low-level models for computer vision?

4.3 Sparse coding

Field and Olshausen (1996) proposed an elegant explanation for Gabor wavelets. In their work, these functions are estimated as parameters in a statistical model for natural image

data. The principle they adopt is the sparsity principle. The question they ask is: for an ensemble of natural image data, can we find a vocabulary of linear bases, so that for every image in that ensemble, we can almost always find a small number of linear bases from this vocabulary to represent this image?

Field and Olshausen (1996) collected an ensemble of natural image patches (of size 12×12), $\mathbf{I}_1, ..., \mathbf{I}_M$. Then they estimate image bases $B_1, ..., B_K$ (which are also images of 12×12, with $K > 12 \times 12$, i.e., the basis is overcomplete) by minimizing

$$\sum_{m=1}^{M} \left\{ \min_{\{c_{m,k}\}} \left[\|\mathbf{I}_m - \sum_{k=1}^{K} c_{m,k} B_k\|^2 + \lambda \sum_{k=1}^{K} S(c_{m,k}) \right] \right\}, \tag{10}$$

over all possible basis $\{B_k\}$, where $S()$ is a measure of sparsity, and λ is a tuning constant. In objective function (10), the first term requires that the linear explanation $\sum_k c_{m,k} B_k$ should be close to the observed image $\mathbf{I}_m$. The second term requires that only a small number of $c_{m,k}$ are significantly different from 0. The simplest measure of sparsity is to count the number of non-zero $\{c_{m,k}\}$, i.e., $S(c) = 1$ if $c \neq 0$, and $S(c) = 0$ if $c = 0$. But this measure does not allow for very small c. Moreover, it is not differentiable, making it hard for optimization. So it can be replaced by some continuous measure such as l_p norm of the sequence $\{c_{m,k}, k = 1, ..., K\}$, with $p \leq 1$. Using a simple 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 12. That is, one can write $k = (x, y, s, \theta)$ where (x, y) is the location (on the 12×12 lattice), s is the scale, and θ is the orientation.

Lewki and Olshausen (1999) and Olshausen and Millman (2001) posed this problem explicitly in a statistical model

$$c_{m,k} \sim p(c) \text{ independently}, \tag{11}$$

$$\mathbf{I}_m = \sum_k c_{m,k} B_k + \epsilon_m, \tag{12}$$

where $p(c)$ is assumed to be a long tailed distribution. The model used by Olshausen and Millman (2001) for $p(c)$ is a mixture of two Gaussian distributions $\rho N(0, \sigma_1^2) + (1-\rho) N(0, \sigma_0^2)$. 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 inactive state, with probability $1-\rho$, which is very large, and the variance σ_0^2 is very small, meaning that most of the times, the coefficient is close to 0. See also Pece (2002).

In this two-level hierarchical model, $c_{m,k}$ are latent variables or missing data. Olshausen and Millman (2001) used a stochastic approximation type of algorithm (similar to the EM algorithm) to compute the maximum likelihood estimate of $\{B_k\}$. The above model is also similar to the Bayesian variable selection problem in linear regression (e.g., George and McCullogh, 1997), except that $\{B_k\}$, the set of regressors themselves, need to be estimated from the data. See also Chipman, Kolaczyk, and McCulloch (1997).

The independence assumption in (11) is only for convenience. In general, one can write the wavelet sparse coding model in the following form:

$$C = \{c_k\} \sim p(C), \tag{13}$$

$$\mathbf{I} = \sum c_k B_k + \epsilon, \tag{14}$$

where $C = \{c_k\}$ are coefficients, and ϵ is assume to be Gaussian white noise. We can rewrite the model (13) and (14) in matrix form $C \sim p(C)$, $\mathbf{I}_1 = BC$, and $\mathbf{I} = \mathbf{I}_1 + \epsilon$, where $\mathbf{I}$ and $\mathbf{I}_1$ become vectors, B is the matrix collecting all the bases $\{B_k\}$, and C is the vector collecting all the $\{c_k\}$.

The uncertainty caused by the overcompleteness can be easily seen via the singular value decomposition of $B = U(D, 0)(V_1, V_0)'$. B is a $N \times K$ matrix, where N is the dimensionality of $\mathbf{I}$, and K is the total number of bases. Because of overcompleteness, $N < K$. U is an $N \times N$ orthogonal matrix. D is N dimensional diagonal matrix of singular values. $V = (V_1, V_0)$ is the $K \times K$ orthogonal matrix, where V_1 is $K \times N$, and V_0 is $K \times (K - N)$. Let $\tilde{C} = (\tilde{C}_1 = V_1'C, \tilde{C}_0 = V_0'C)$, then clearly, $\mathbf{I}_1 = BC = UD\tilde{C}_1$. That is, only $\tilde{C}_1$ can be decided from $\mathbf{I}_1$, while $\tilde{C}_0$ cannot be determined.

For an analysis of entropy,

$$\mathcal{H}(p(C)) = \mathcal{H}(p(\tilde{C})) = \mathcal{H}(p(\tilde{C}_1)) + \mathcal{H}(p(\tilde{C}_0|\tilde{C}_1)).$$

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

Therefore,

Proposition 7 *In the above notation,*

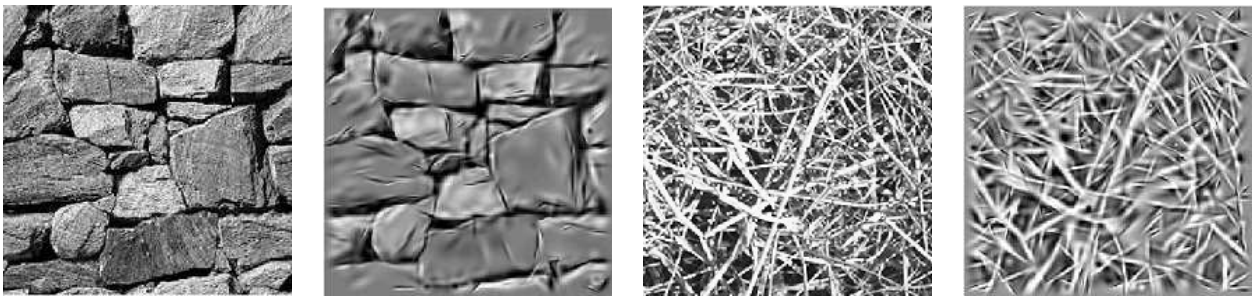
$$\mathcal{H}(p(\mathbf{I}_1)) = \mathcal{H}(p(C)) - \mathcal{H}(p(\tilde{C}_0|\mathbf{I}_1)) + \frac{1}{2}\log|\det(BB')|.$$

If $p(C)$ is very sparse, for instance, the parameter ρ in the mixture model $\rho N(0, \sigma_1^2) + (1-\rho)N(0, \sigma_0^2)$ is very small, then $\mathcal{H}(p(C))$ is small, and thus $\mathcal{H}(p(\mathbf{I}_1))$ is also small. Therefore, 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 $\mathbf{I}_1$. 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, and $\mathcal{H}(p(\tilde{C}_0|\mathbf{I}_1))$, which can be considered imperceptibility for this model, can become large. That is, the uncertainty caused by overcompleteness can be large.

Olshausen and Field (1996) worked with small image patches. For large images, even if we design the bases $\{B_{x,y,s,\theta}\}$ beforehand, it is still a computational challenge to estimate the coefficients by minimizing the posterior distribution of coefficients according to model (11) and (12), or minimizing an objective function (10).

Mallat and Zhang (1993) proposed a greedy algorithm called matching pursuit algorithm for finding a sparse representation of an image $\mathbf{I}$ given an overcomplete set of bases $\{B_k, k = 1, ..., K\}$. In the language of variable selection in linear regression, the matching pursuit algorithm is essentially the forward stepwise regression. We start from an empty set of bases. Each time, we select a base that gives us the largest reduction in the l_2 norm of error. The algorithm stops when the error is smaller than a threshold. Wu, Zhu, and Guo (2002) proposed a Markov chain Monte Carlo version of the matching pursuit algorithm of Mallat and Zhang (1993) that rigorously samples from the posterior distribution of model (11) and (12).

Now let's examine the sparse coding model emiprically by some experiments. In the experiments, we use an overcomplete set of Gabor wavelets as those depicted in Figure (12).



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

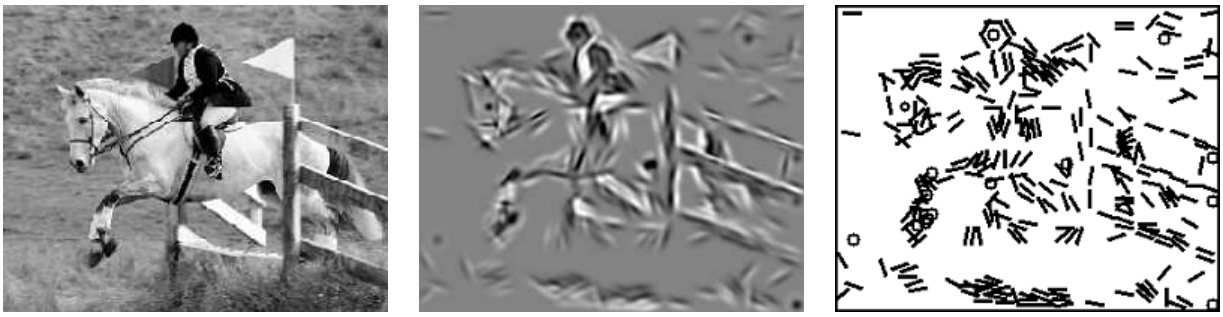
Figure 13: Sparse coding. (a) and (c) are observed images of 128 × 128 pixels. (b) and (d) are respectively the reconstructed images using 300 bases.

At each pixel, there are localized Gabor wavelets of different scales and orientations. So the set of bases are highly overcomplete. We use the matching pursuit algorithm to construct the sparse coding of the observed images.

Figure (13) shows two examples of sparse coding. (a) and (c) are observed images of 128 × 128 pixels, (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 not be sparse any more.

There are two more problems with sparse coding model with independent prior distribution, as illustrated by Figure (14), where (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. As shown by this experiment, one problem of wavelet sparse coding is that the edges can become quite blurry, indicating that wavelet bases are not sharp enough to describe the edges. The other problem is that the bases do not line up very well, indicating that we need stronger prior model for the spatial organization of the local bases, so that they line up into more regular structures.

The wavelet theory started from harmonic analysis, for representing functions in various



(a) (b) (c)

Figure 14: Sparse coding. (a) is the observed 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.

functional classes by localized and self-similar base functions. For functional classes, there is also a concept of entropy, called Kolmogorov ϵ-entropy. See Donoho, Vetterli, Devore, and Daubechies (1998) for a review of this concept as well as the applications of wavelets to image compression. Following the definition of their paper, for a compact functional class Π with norm $\| \|$, let a net $\mathcal{N}_\epsilon = \{f'\}$ be such that

$$\sup_{f \in \Pi} \min_{f' \in \mathcal{N}_\epsilon} \|f - f'\| \leq \epsilon,$$

then the Kolmogorov ϵ-entropy is defined as the logarithm of the minimum cardinality of such a net. Intuitively, it can be imagined as the Shannon entropy of the uniform distribution over the functional class Π that are discretized with precision ϵ. The Kolmogorov entropy has been connected to the decaying rate of the coefficients of orthogonal wavelet transform.

Olshausen and Field (1996) learned the local image basis empirically from natural images. Candes and Donoho (1999) designed a set of basis they called curvelets based on theoretical considerations under the slogan that "natural images have edges". They study the class of two dimensional functions with discontinuities on smooth curves in $\mathbf{R}^2$, and showed that curvelet basis gives sparse coding for such functions. We would like to point out that the scenario of their revealing investigation is the low entropy or near distance situation.

4.4 Markov random fields and feature statistics

The Markov random fields originated in statistical physics, and they were first introduced to statistics by Besag (1974). Geman and Geman (1984) and many other researchers have used Markov random fields for image processing and modeling. Zhu and Mumford (1997) connected Markov random fields to partial differential equations and variational approaches for image processing (Aubert and Kornprobst, 2002). See the book of Winkler (1995) for a comprehensive treatment to Markov random fields and related topics.

The Markov property of a Markov random field is defined with respect to a neighborhood system, where for each pixel $(x, y) \in \Lambda$, there is a set of neighboring pixels $\partial(x, y) \subset \Lambda$. 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 $\partial = \{\partial(x, y) : (x, y) \in \Lambda\}$, 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 $(x, y) \in \Lambda$,

$$p(\mathbf{I}_{x,y} \mid \mathbf{I}(\Lambda \backslash (x, y))) = p(\mathbf{I}_{x,y} \mid \mathbf{I}(\partial(x, y)). \tag{15}$$

By convention, for $A \subset \Lambda$, we define $\mathbf{I}(A)$ as the intensities of all the pixels in A. $\Lambda \backslash (x, y)$ means all the pixels except (x, y). The Markov property (15) means that the distribution of the pixel intensity only depends on the intensities of neighboring pixels.

According to Hammersley-Clifford (1968) theorem, a Markov random field with respect to the neighborhood system ∂ can be written as the 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 purpose, if A has many pixels, then U_A will be 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 in 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) proposed a modeling strategy to get around this problem: replace the high dimension $\mathbf{I}(A)$ by low dimensional features. Suppose we have a set of feature extractors, $\{\phi_{x,y,k}\}$, so that $\phi_{x,y,k}(\mathbf{I})$ extract local feature at a clique around (x, y). One example of feature extractors is $\phi_{x,y,k} =< \mathbf{I}, B_{x,y,s,\theta} >$, with $k = (s, \theta)$. Then we can model the image $\mathbf{I}$ by the following Gibbs distribution

$$f(\mathbf{I} \mid \beta) = \frac{1}{Z} \exp\{\sum_{k=1}^{K} \sum_{x,y} \beta_{x,y,k}(\phi_{x,y,k}(\mathbf{I}))\}, \tag{16}$$

where $\beta = \{\beta_{x,y,k}()\}$ is a set of a low-dimensional functions of features, and Z is the normalizing constant depending on $\{\beta_k()\}$. This is essentially the model proposed by Zhu, Wu, and Mumford (1997) for modeling textures, where they assume $\phi_{x,y,k} =< \mathbf{I}, B_{x,y,s,\theta} >$.

We can prove the following information-theoretical results.

Proposition 8 *Suppose the true distribution of* $\mathbf{I}$ *is* $p(\mathbf{I})$*, and let* $p_{x,y,k}()$ *be the distribution of* $\phi_{x,y,k}(\mathbf{I})$ *under* $\mathbf{I} \sim p(\mathbf{I})$*. Suppose there exists a* β^**, such that under* $f^*(\mathbf{I}) = f(\mathbf{I} \mid \beta^*)$*, the marginal distribution of* $\phi_{x,y,k}$ *is* $p_{x,y,k}()$ *for all* (x, y, k)*, then* $\mathcal{D}(p||f^*) \leq \mathcal{D}(p||f)$ *for any* $f = f(\mathbf{I} \mid \beta)$*.*

Proof

$$\begin{aligned}
\mathcal{D}(p||f) - \mathcal{D}(p||f^*) &= \log \frac{Z(\beta)}{Z(\beta^*)} \mathrm{E}_p \left\{ \sum_{k=1}^{K} \sum_{x,y} (\beta^*_{x,y,k}(\phi_{x,y,k}(\mathbf{I})) - \beta_{x,y,k}(\phi_{x,y,k}(\mathbf{I}))) \right\} \\
&= \log \frac{Z(\beta)}{Z(\beta^*)} \mathrm{E}_{f^*} \left\{ \sum_{k=1}^{K} \sum_{x,y} (\beta^*_{x,y,k}(\phi_{x,y,k}(\mathbf{I})) - \beta_{x,y,k}(\phi_{x,y,k}(\mathbf{I}))) \right\} \\
&= \mathcal{D}(f^*||f) \geq 0. \text{ QED.}
\end{aligned}$$

Remark: f^* can be considered the best approximation to p among all possible $f(\mathbf{I} \mid \beta)$. It can also be regarded as the "maximum likelihood estimate" , since minimizing $\mathcal{D}(p||f(\mathbf{I}|\beta))$ amounts to maximizing $\mathrm{E}_p[\log f(\mathbf{I} \mid \beta)]$, which can be considered the likelihood function with p serving as the "data".

Proposition 9 *In the above notation, let $q(\mathbf{I})$ be any distribution such that the marginal distribution of $\phi_{x,y,k}(\mathbf{I})$ under $q(\mathbf{I})$ is $p_{x,y,k}()$ for all (x, y, k). Then $\mathcal{H}(f^*) - \mathcal{H}(q) = \mathcal{D}(q||f^*)$.*

Proof

$$\begin{aligned} \mathcal{H}(f^*) - \mathcal{H}(q) &= \mathrm{E}_q[\log q(\mathbf{I})] - \mathrm{E}_{f^*}[\log f^*(\mathbf{I})] \\ &= \mathrm{E}_q[\log q(\mathbf{I})] - \mathrm{E}_q[\log f^*(\mathbf{I})] = \mathcal{D}(q||f^*). \text{ QED.} \end{aligned}$$

The above proposition has the following implications:

1. Maximum entropy. Among all the $q(\mathbf{I})$ that reproduces the marginal distributions $p_{x,y,k}()$, f^* has the maximum entropy. Therefore, f^* can be considered the most unprejudiced fusion of the statistical properties represented by $p_{x,y,k}()$.

2. f^* always approaches the true distribution p from above in terms of entropy. That is, f^* is always more random than p. So Markov random field model may be used to model high entropy patterns.

3. Minimum entropy. If we want to find the best set of features $\{\phi_{x,y,k}\}$, we need to minimize the entropy f^* over all possible sets of features, since that will give the best approximation to the true distribution p in terms of $\mathcal{D}(p\|f)$.

To model observed images, we may assume that the images are locally stationary. Let's still use $\mathbf{I}$ to denote a spatially stationary image patch. Then we can assume that $\beta_{x,y,k}() = \beta_k()$, i.e., the β-function does not depend on (x, y). One can further parametrize $\beta_k()$ by step functions or low order polynomials. If we parametrize $\beta_k()$ by a step function over a set of bins $R_l, l = 1, ..., L$, so that $\beta_k(\phi_{x,y,k}(\mathbf{I})) = \beta_{kl}$ if $\phi_{x,y,k}(\mathbf{I}) \in R_l$, then we can write model (16) as

$$\begin{aligned} f(\mathbf{I} \mid \beta) &= \frac{1}{Z} \exp\{\sum_k \sum_{x,y} \sum_l \beta_{kl} \delta_{\phi_{x,y,k}(\mathbf{I}) \in R_l}\} \\ &= \frac{1}{Z} \exp\{\sum_k \sum_l \beta_{kl} H_{kl}(\mathbf{I})\} = \frac{1}{Z} \exp\{\sum_k \langle \beta_k, H_k(\mathbf{I}) \rangle\}, \end{aligned}$$

where $H_{kl}(\mathbf{I}) = \sum_{x,y} \delta_{\phi_{x,y,k}(\mathbf{I}) \in R_l}$, i.e., the number of $\phi_{x,y,k}(\mathbf{I})$ falling into bin R_l, and $H_k = (H_{kl}, \forall l)$ is the marginal histogram of $\{\phi_{x,y,k}(\mathbf{I}), \forall x, y\}$.

Clearly, the above model is an exponential family model, where the spatial feature statistics $H_k(\mathbf{I})$ are the sufficient statistics. As a well-known fact about exponential family model, if we want to find the maximum likelihood estimate of β from the observed image $\mathbf{I}_{\text{obs}}$, we only need to solve the following estimation equation

$$\mathrm{E}_\beta[H_k(\mathbf{I})] = \hat{H}_k = H_k(\mathbf{I}_{\text{obs}}), \forall k, \tag{17}$$

that is, we need to match the model and the data in terms of the spatial statistics. Or in other words, the fitted model is decided by $H_k(\mathbf{I}_{\text{obs}})$. $\mathrm{E}_\beta[H_k(\mathbf{I})]$ is called the mean parameter of the model, and β is called natural parameter.

There is something much deeper than that, and it is produced by a most fundamental insight in statistics physics: the equivalence of ensembles. Borrowing this insight, Wu, Zhu, and Liu (2000) considered the following ensemble, which is called micro-canonical ensemble in statistical physics (Chandler, 1987):

$$\Pi = \{\mathbf{I} : H_k(\mathbf{I}) = \hat{H}_k, \forall k\}, \tag{18}$$

where $\hat{H}_k$ can be estimated from observed image. This is a deterministic concept of equivalent class, where all the images in that ensemble produce the same spatial statistics.

One can show that under the uniform distribution over Π, if the image domain $\Lambda \to Z^2$, then the image intensities defined on any fixed local lattice Λ_0 follows

$$p(\mathbf{I}_{\Lambda_0} \mid \mathbf{I}_{\partial\Lambda_0}) = \frac{1}{Z} \exp\{\sum_k \sum_{x,y \in \Lambda_0} \beta_k(\phi_{x,y,k}(\mathbf{I}))\}, \tag{19}$$

where $\partial\Lambda_0$ are the neighboring pixels of Λ_0 so that pixels in $\partial\Lambda_0$ and pixels in Λ_0 can be covered by the same filters. β can be solved from equation (17).

One may imagine that we put all the large images in the micro-canonical ensemble (18) on top of each other. Then we only look at these images through the window Λ_0, we will see a lot of image patches. The frequency distribution of these image patches is given by (19).

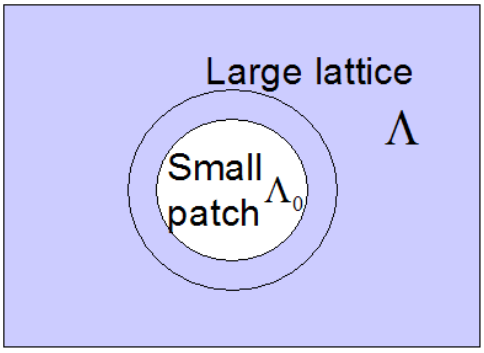


Figure 15: The deterministic concept of micro-canonical ensemble defined on $\Lambda \to \mathbf{Z}^2$ produces the probabilistic concept of Markov random field on a fixed patch Λ_0, according to the equivalence of ensemble in statistical physics.

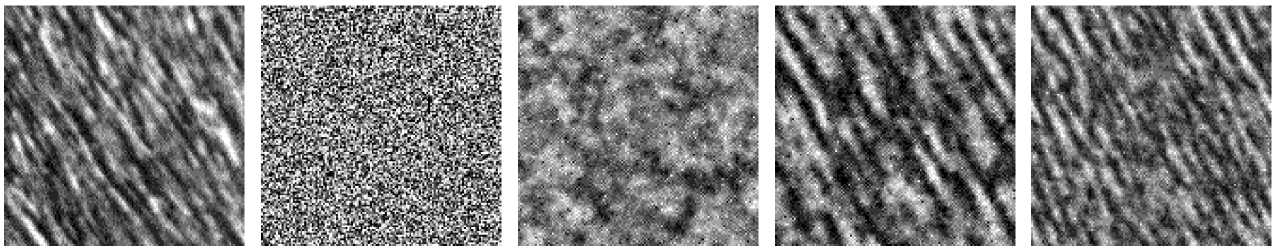
Conversely, as $\Lambda \to Z^2$, the Markov random field model (16) is equivalent to the uniform distribution over the micro-canonical ensemble (18), and one can show that the entropy of the Markov random field model is $\log |\Pi|$, where $|\Pi|$ is the volume of Π. $\log |\Pi|$ is also called combinatorial entropy by Kolmogorov (see Rissanen, 1989), and it appears to be related to Kolmogorov ϵ-entropy. So the entropy can be considered a measure of the dimensionality of Π. In terms of this micro-canonical ensemble and combinatorial entropy,

1. Maximum entropy means that we should put the uniform distribution over the micro-canonical ensemble.

2. Minimum entropy means that we should choose the set of local feature extractors $\{\phi_{x,y,k}\}$, so that the corresponding micro-canonical ensemble has the smallest volume.

Zhu, Wu, and Mumford (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 in the fitted model. Figure (16) displays an example of filter pursuit procedure on homogeneous texture. With $K = 0$ filters, the sampled image is white noise. With $K = 7$ filters, the sampled image in (e) is perceptually equivalent to the input image. This method is similar to the projection pursuit density estimation (Friedman, 1987). The difference is that we have an

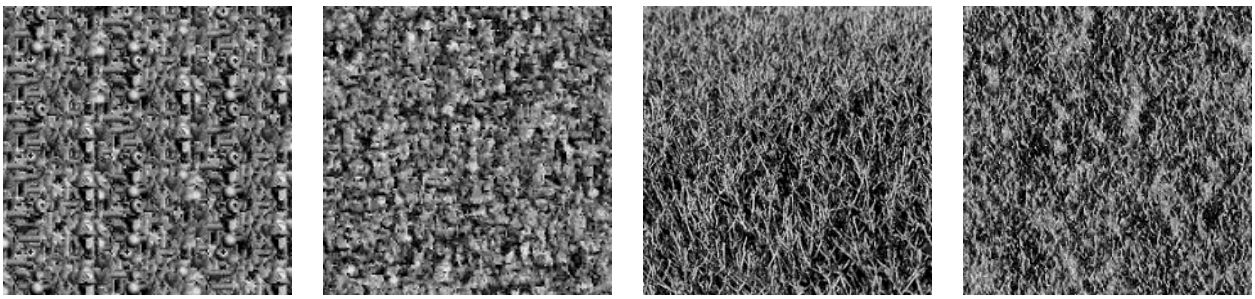
explicit maximum entropy model in the form of exponential family Markov random field, and the linear filters overlap with each other.



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

Figure 16: Filter pursuit: adding one filter at a time to reduce the entropy. The filters pool the statistical information (histograms) to yield a texture impression.

The following are some experiments with $\phi_{x,y,k}(\mathbf{I}) = \langle \mathbf{I}, B_{x,y,s,\theta} \rangle$, with $k = (s, \theta)$. These experiemnts show that the filter statistics are quite effective in representing stochastic textures. Figure (17) 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 (18) by matching feature statistics. See Heeger and Bergen (1995), Srivastava, Grenander, and Liu (2002), Portilla and Simoncelli (2000) for more discussions on feature statistics.



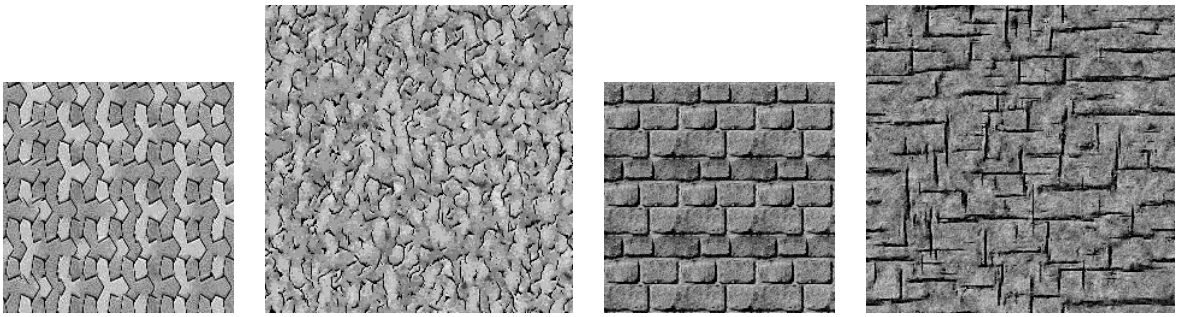
(a) (b) (c) (d)

Figure 17: Feature statistics. (a) and (c) are observed images. (b) and (d) are "reconstructed" by matching feature statistics.

We need to stress that, under the Markov random field model or equivalently the micro-

canonical ensemble, the filter responses $\phi_{x,y,k}(\mathbf{I}) = \langle \mathbf{I}, B_{x,y,s,\theta} \rangle$ are *not* independent of each other, because the number of bases $B_{x,y,s,\theta}$ far exceeds the number of pixels. Although only marginal distributions are specified, the dependencies among adjacent responses from the same filter can be accounted for implicitly by the distributions of the responses from other filters. Sometimes, the long range patterns can emerge by matching statistics of local features.

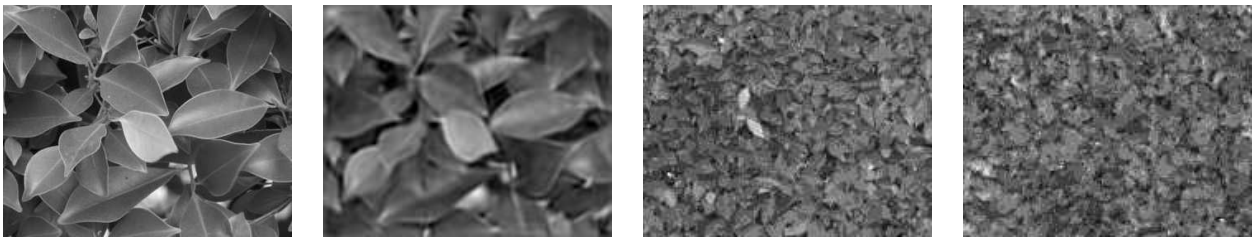
But still, since the model only specifies the marginal distributions of filter responses, it cannot represent large regular structures very well. See Figure (18) for two examples with line structures. In order to model regular structures, we need to represent these structures explicitly. Moreover, we also need to model the spatial organizations of these structure.



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

Figure 18: Feature statistics. (a) and (c) are observed images. (b) and (d) are "reconstructed" by matching feature statistics.

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 (11) and (12) (with $\epsilon = 0$) and the Markov random field model (16) (with $\phi_{x,y,k}(\mathbf{I}) = \langle \mathbf{I}, B_{x,y,s,\theta} \rangle$) reduce to the independent component analysis model (Bell and Sejnowski, 1997).



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

Figure 19: From sparse coding to feature statistics. (a) Observed near-distance image. (b) Reconstructed by sparse coding with 1,000 bases. (c) Observed far-distance image. (d) "Reconstructed" by matching feature statistics.

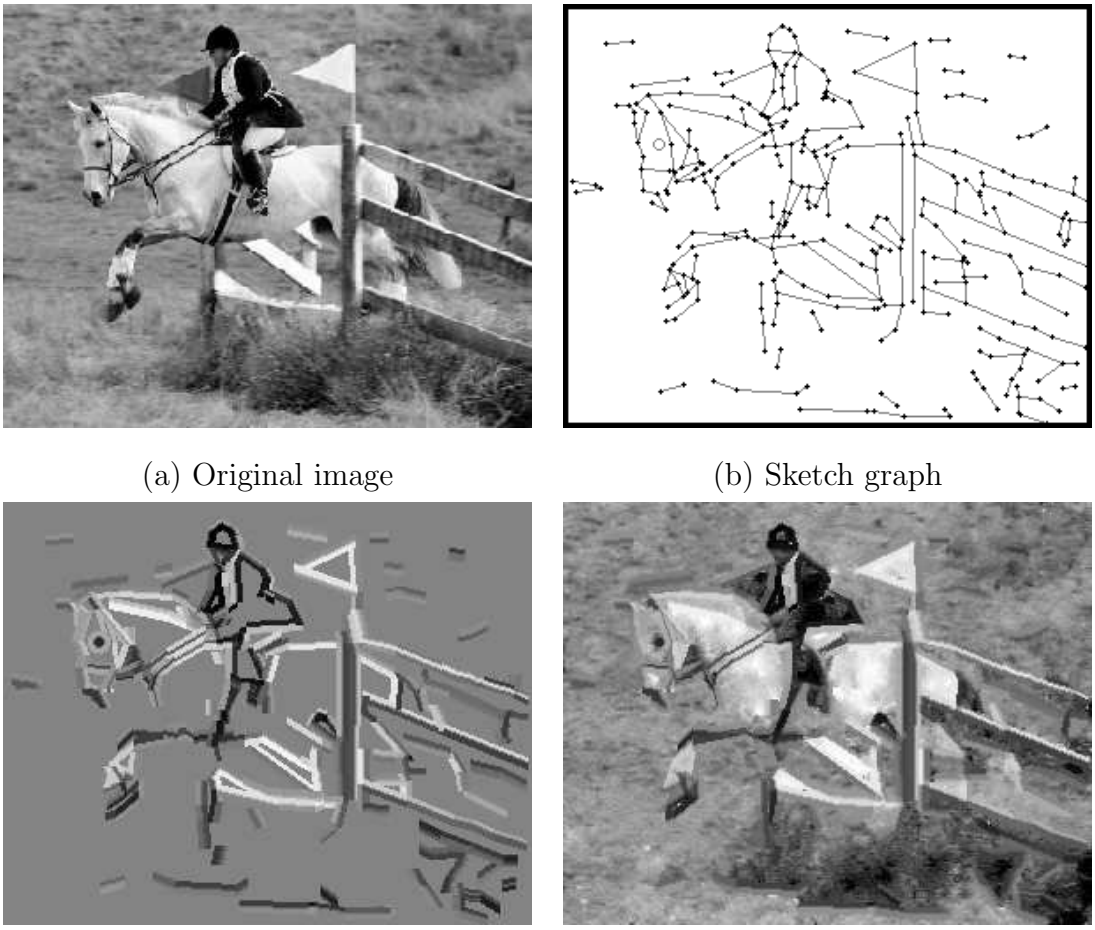
5 Full-Zoom Primal Sketch Model

5.1 Integrating two regimes of models

Because of information scaling, image data may have different entropy rates, and the underlying geometric structures that produce the image data may have different perceptibilities. Our examination of the wavelet sparse coding and Markov random fields indicates that the wavelet sparse coding model is appropriate for low entropy regime, and the Markov random fields or feature statistics are appropriate for high entropy regime. For instance, Figure (19) displays results for two images of leaves. (a) is the observed 300×200 image of leaves at near distance. (b) is the image reconstructed by the matching pursuit algorithm using 1,000 Gabor wavelets. (c) is the observed image at far distance. (d) is obtained by matching the histograms of filter responses from a set of Gabor wavelets. (d) is not an exact reconstruction of (a), but it captures the texture appearance of (c).

Therefore, we may combine the two regimes of models into a full-zoom primal sketch model to account for the whole range of scale and entropy.

However, as shown by Figure (14) as well as Figure (19.b), the wavelet bases are not sparse enough for coding geometric structures such as edges and bars, neither do they align into lines, curves, junctions, and corners. For such geometric structures, we need more



(a) Original image (b) Sketch graph

(c) Sketchable part (d) Synthesized image

Figure 20: Full zoom primal sketch model. (a) The observed image. (b) The "sketchable" part is described by a geometric sketch graph. (c) The sketch part of the image. (d) Fill in the "non-sketchable" part by matching feature statistics.

sophisticated local parametric models with explicit geometric parameters such as length, width, scale, orientation for edges and bars. These local models should also have photometric parameters in order to generate the image intensities. We call these local parametric models the sketch primitives. These primitives line up into lines and curves, and their joints and intersections form corners and junctions etc.

See Figure (20) for an example. Figure (20.a) is the observed image, the same as the observed image in Figure (14). It is represented by a small number of sketch primitives, which form a sketch graph, see Figure (20.b). The nodes are end points, corners, junctions. The nodes are connected by edges and bars. These sketch primitives generate what we call the "sketchable" part of the image, see Figure (20.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 gives us most of the information in the image data.

Since these local primitive models are much more flexible and sophisticated than the wavelet sparse coding model, they are also much more computationally expensive to fit than the wavelet model. So we still need to start from local wavelet sparse coding to detect the geometric structures, and then fit the sparser non-linear primitive models to sketch the image.

Then how about the "non-sketchable" part of the image? The fact that no clear geometric structures are identified in this part of the image indicates that the underlying constituent elements are of very small scales or at far distances. They are not perceptible, and there can be a large number of them. So instead of coding them deterministically, we can only summarize them statistically if we want a parsimonious description of this part of the image.

Then what statistics should we use for non-sketchable part of the image? The answer seems to come from the effort of sketching the image. In order to detect the sketchable part of the image, we have to apply some local detectors, such as local wavelet sparse coding, everywhere in the image. Even if these detectors fail to report any structures that deserve further model fitting by sketch primitives, these detectors must have performed some

computations and extracted some features from local image intensities. These features must be quite informative about the local image intensity patterns, especially when the local image intensities are close to the borderline of being sketchable. Computationally, these extracted features should not be thrown away if the local image intensities are not sketchable. Instead, we should pool these features within a local area into statistics for describing the non-sketchable part of the image. Since no clear structures are identified in this part of the image, we may only need to find the marginal statistics of the extract features, without worrying about the spatial organizations of these features.

So our proposal is that the non-sketchable statistics summarize the features extracted by the failed sketch detectors. Or in other words, the effort of sketching the image is like putting the image through a mesh. What is left on the mesh is the sketch graph of the image. The part of the image that falls through the mesh is recycled into local marginal statistics of failed sketches.

In our work, we still use the marginal histograms of filter responses as the non-sketchable statistics. These statistics imply a Markov random field model. See Figure (20.d) for a synthesized image which is obtained by filling in the blank part of Figure (20.c) by matching the non-sketchable statistics of the observed images. Mathematically, the non-sketchable statistics play the role of regularizing the *interpolation* of the sketchable part of the image. Graphically, filling in the non-sketchable part of the image is a matter of *inpainting*. From this perspective, we may consider our modeling approach as "sketching and inpainting". See Chan and Shen (2001) for more details on the issue of inpainting.

5.2 The sketch primitives and sketchability

A most prominent sketch primitive is an oriented and elongate structure such as an edge or a bar:

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

where $h()$ is a one-dimensional profile function, and S is an oriented rectangle set of pixels along direction θ. The low entropy is achieved by the fact that the two-dimensional image

patch $\mathbf{I}(S)$ is represented by one-dimensional profile $h()$ along a direction θ. Moreover, the profile $h()$ can be further modeled by some parameteric functions. For edges, Elder and Zucker (1998) proposed the following profile. We start from a step edge $h_0(x) = 1/2$ for $x \leq 0$, and $h_0(x) = -1/2$ for $x > 0$. We let $h(x) = a + bh_0(x) * g(s)$, where $g(s)$ is a Gaussian kernel with bandwidth s. Here the parameter of such an edge primitive includes $(x_0, y_0, l, w, s, \theta, a, b)$ with location (x_0, y_0), length l, width w, scale s, orientation θ, local intensity level a, edge magnitude 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.

For a bar, it is a composition of two scaled edges. The junctions and corners are composition of edges and bars.

For an image $\mathbf{I}$ defined on lattice Λ, let Λ_{sk} be the sketchable part of the lattice. Let $\{\Phi_i(x, y \mid \alpha_i), i = 1, ..., n\}$ be the set of sketch primitives that describes $\mathbf{I}(\Lambda_{\text{sk}})$, and let S_i be the pixels covered by $\Phi_i(x, y \mid \alpha_i)$. Then $\Lambda_{\text{sk}} = S_1 \cup ... \cup S_n$, and

$$\mathbf{I}(x, y) = \Phi_1(x, y|\alpha_1) + ...\Phi_n(x, y|\alpha_n) + \epsilon(x, y), \quad (x, y) \in \Lambda_{\text{sk}},$$

and $\epsilon(x, y) \sim N(0, \tau_i^2)$ for $(x, y) \in S_i$.

As a provisional working model, we temporally assume that for $(x, y) \in \Lambda \backslash \Lambda_{\text{sk}}$, $\mathbf{I}(x, y) \sim N(\mu_{x,y}, \sigma_{x,y}^2)$ independently, where $\mu_{x,y}$ and $\sigma_{x,y}^2$ are slowly varying. This part of model is going to be replaced after we sketch the image.

The sketch primitives $\{\Phi_i(x, y \mid \alpha_i), i = 1, ..., n\}$ form a sketch graph G, where G collects the geometric aspects of the primitives. The prior model for G is of the following form

$$p(G) = \frac{1}{Z} \exp\{-\sum_j \eta_j F_j(G)\},$$

where $F_j(G)$ are some features computed from the graph G, including the number of sketch primitives and the number of end points. The parameter η_j are chosen to favor graphs with extended and connected primitives, so that the number of primitives and the number of end points are small.

In order to fit the model, we can minimize the following sketchability criterion

$$\sum_i \frac{1}{2}|S_i| \log \frac{\|\mathbf{I}(S_i) - \Phi(x, y|\alpha_i)\|^2}{|S_i|\text{Var}(\mathbf{I}(S_i))} + \sum_j \eta_j F_j(G). \tag{20}$$

This is the penalized likelihood for image $\mathbf{I}$, with $\hat{\tau}_i^2 = \|\mathbf{I}(S_i) - \Phi(x, y|\alpha_i)\|^2/|S_i|$, and for $(x, y) \in S_i$, $\hat{\mu}_{x,y} = \text{E}(\mathbf{I}(S))$, $\hat{\sigma}^2_{x,y} = \text{Var}(\mathbf{I}(S_i))$, i.e., the empirical marginal mean and variance of $\mathbf{I}(S_i)$. One may also use a more strict Bayesian method for model fitting. We consider this a minor point.

The sketchability criterion (20) can also be interpret as description length of the sketchable part of the image. The first term is the coding cost for residual error (relative to white noise model). The second term can be interpreted as the cost for coding the primitives.

As we discussed above, we still need to use wavelet bases as a precursor to minimize (20). We first compute $r_{x,y,s,\theta} = \langle \mathbf{I}, B_{x,y,s,\theta} \rangle$ for all (x, y, s, θ). Then we identify the local maxima, where $r_{x,y,s,\theta}$ is a local maximum if it is greater than other $r_{x,y,s,\theta}$ within a predefined local neighborhood in the space of (x, y, s, θ). These local maxima can produces a sparse coding of the image using the corresponding $B_{x,y,s,\theta}$. Moreover, they give us important information as to where to fit the sketch primitives. These local maxima behave similarly to Canny (1986) edge detector.

We then use a sketching pursuit process to identify a sketch graph. The process starts to fit a single primitive at the global maximum of $r_{x,y,s,\theta}$. Then we use a set of moves to reduce the sketchability criterion (20). This set of moves includes: 1) extend a primitive, or shrink a primitive; 2) add a new primitive, or remove an existing primitive; 3) join two primitives to form a corner, or break a corner; 4) extend one primitive to touch another primitive to form a junction, or break a junction. More technical details are reported in Guo, Zhu, and Wu (2004). The computation is quiet efficient, taking less than 30 seconds for an 200×200 image.

5.3 The non-sketchable statistics

In the ideal case, we can fit the sketch model $\mathbf{I}(x, y) = \Phi(x, y|\alpha) + \epsilon$ at every location (x_0, y_0) of the image lattice, and let $\phi_{x_0,y_0}(\mathbf{I}(S)) = (\hat{\alpha}, \hat{\tau}^2)$, i.e., the local model fitting result. For those non-sketchable (x_0, y_0), the model does not fit very well, that is, the sketch $\Phi(x, y|\hat{\alpha})$ does not give an accurate account of local image intensities around (x_0, y_0). For those failed sketches, there is not way for them to form any conspicuous sketch graph structures. As a result, we can pool the marginal distributions of the model fitting results to form the feature statistics, by throwing out all the position information. That is, the ideal non-sketchable statistics can be the marginal statistics of failed model fitting results.

Since we cannot afford to fit local model everywhere, we may just pool the marginal distributions of local sketch detectors $\phi_{x,y,k}(\mathbf{I})$. Let $N_{x,y} \subset \Lambda_{\text{nsk}} = \Lambda \backslash \Lambda_{\text{sk}}$ be a non-sketchable local image patch around pixel (x, y). We summarize $\mathbf{I}(N_{x,y})$ by

$$f_{x,y,k}(z) = \frac{1}{|N_{x,y}|} \sum_{(x',y') \in N_{x,y}} \delta_{\phi_{x',y',k}(\mathbf{I})}(z), \forall k,$$

i.e., the local empirical distribution (or histogram if we discretize the range of z) of features $\{\phi_{x',y',k}(\mathbf{I})\}$ for all the (x', y') around (x, y). We can further group pixels (x, y) into several regions based on $\{f_{x,y,k}(z)\}$, so that each region is stationary in $\{f_{x,y,k}(z)\}$.

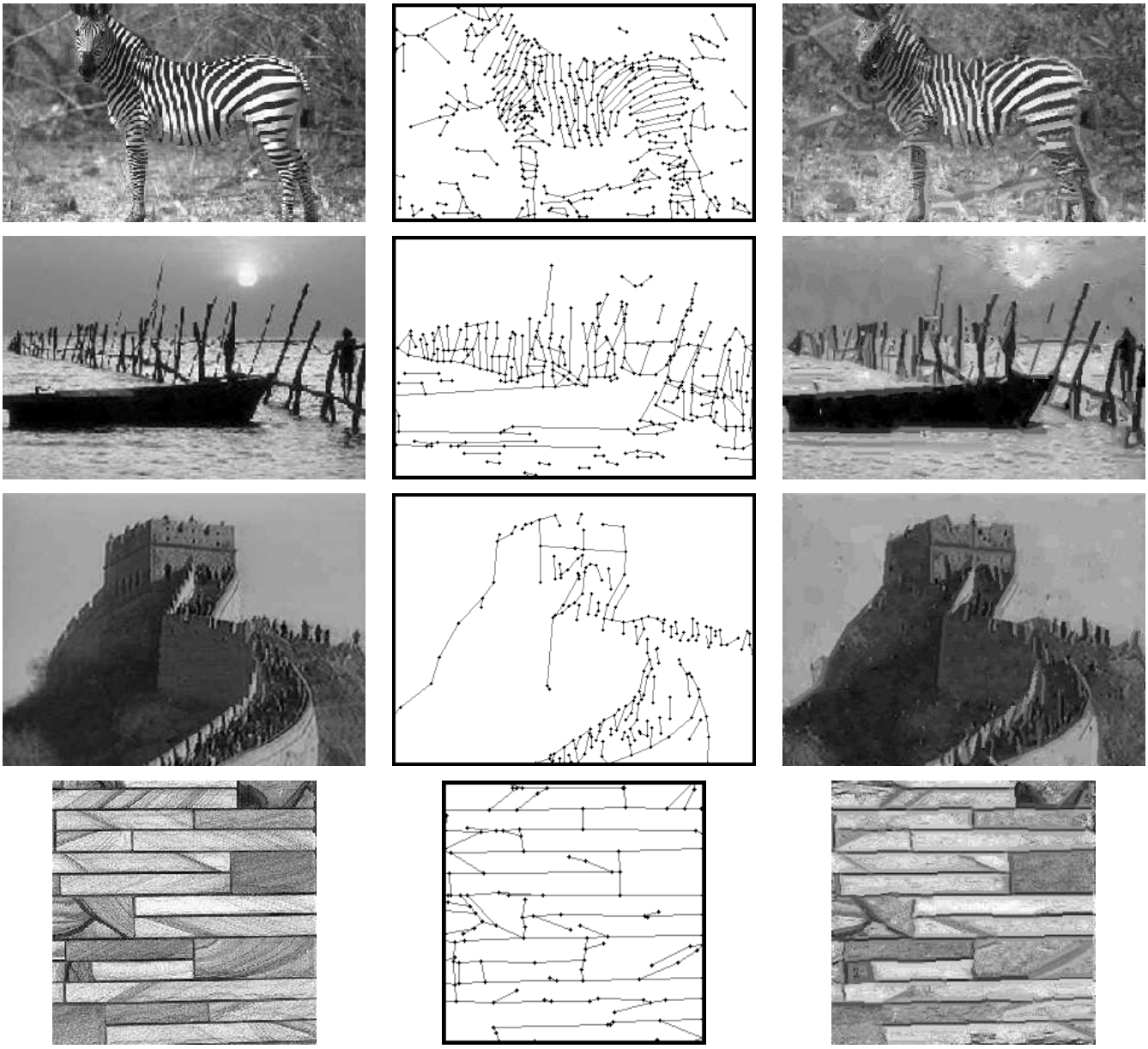
According to the equivalence of ensembles, the above statistics imply a Markov random field, which interpolates the $\mathbf{I}(\Lambda_{\text{sk}})$ in the form of the conditional distribution of $\mathbf{I}(\Lambda_{\text{nsk}})$ given $\mathbf{I}(\Lambda_{\text{sk}})$,

$$f(\mathbf{I}(\Lambda_{\text{nsk}})) \mid \mathbf{I}(\Lambda_{\text{sk}}) = \frac{1}{Z} \exp\{\sum_{(x,y) \in \Lambda_{\text{nsk}}} \sum_{k} \beta_{x,y,k}(\phi_{x,y,k}(\mathbf{I}))\}.$$

In our current experiments, we use $\phi_{x,y,k} = \langle \mathbf{I}, B_{x,y,s,\theta} \rangle$, with $k = (s, \theta)$, because these results are the initial computations of the effort of sketching the image. It can also be other feature extractors for detecting sketch primitives.

Figure (21) shows more experiments of full-zoom primal sketch model.

For a theoretical understanding of the pooling of local model fitting results, we have the following proposition.



(a) (b) (c)

Figure 21: Examples of full-zoom primal sketch model. (a) observed image; (b) sketch graph; (c) synthesized image from the fitted model.

Proposition 10 *For a stationary random field f defined on $\mathbf{Z}^2$, let $S \subset \mathbf{Z}^2$ be a local squared patch, and let $\Lambda \subset \mathbf{Z}^2$ be a large squared patch whose length or width is the multiple of that of S. Then $\overline{\mathcal{H}}(f(\mathbf{I}(\Lambda))) \leq \overline{\mathcal{H}}(f(\mathbf{I}(S)))$.*

The local model fitting targets small patches such as $\mathbf{I}(S)$. These local patches overlap with each other. By fusing the local model fitting results together into a micro-canonical ensemble or a Markov random field f, we are able to achieve a smaller entropy rate. In some sense, these local model fittings tighten each other since they cover overlapping patches.

6 Limitations and future work

The experiments with full-zoom primal sketch model suggest that it can capture considerable amount of low-level essence of natural images. However, the current form of the model cannot handle the large number of objects and their parts at the small and non-sketchable scales, such as faces, hand-written characters, many man-made objects. For these objects, we need more sophisticated detection algorithms and representation schemes.

In our current experiments, we still use histograms of filter responses as spatial statistics for non-sketchable part of the image. According to our proposal that the non-sketchable statistics should recycle failed local sketches, we could use more sophisticated local features that are built on filter responses. In our future work, we shall search for more powerful statistics along this line of thinking, and eventually catalog the wide variety of non-sketchable patterns.

A visual object or pattern generates a scale space of images at different viewing distances. So there should also correspond a scale space of models or representations. Since the viewer can move in a visual scene, we should track the change of the topology of the sketch graph, the transition between sketchable primitives and non-sketchable statistics, as well as the change of non-sketchable statistics over scale. In addition, visual patterns, such as a grass ground or brick wall, can appear in perspective. We should model the changes caused by the view perspective as well.

Our prior model on the spatial configuration of the sketch primitives only reflects some simple regularities. We need more sophisticated model to account for local geometry and spatial organization.

In our current work, we only deal with static images. For motion images, there are also issues of complexity and perceptibility. Some motion patterns are very simple, for instance, the motion of ridge body. Some motion patterns are complex, for instance, the motion of fluid. The simple patterns are often trackable, whereas the complex patterns are often not trackable. Downsampling in both spatial and temporal dimensions can increase the complexity and reduce the trackability. We shall extend our method to motion patterns in future work.

In terms of modeling, we would like to raise the following points.

There seem to be two notions of simplicity. One is simple regularity, such as the simulated ivy wall image at very near distance. The other is simple randomness, such as the simulated ivy wall image at very far distance, where the image is white noise. While the notion of sparsity or dimension reduction captures the first notion of simplicity, it does not capture the second notion of simplicity, which is the simplicity of the underlying probability model instead of the simplicity of the data. Deeper thinking is needed to define the precise meaning of simplicity.

Our model is essentially a parametric model. For computer vision, parametric modeling seems the right approach given the specific tasks of vision, our knowledge of the physical world, and the large amount of training data. This is different from many other classification and function estimation situations where no much domain-specific knowledge is available. In these situations, non-parametric classification methods and wavelet analysis can be more appropriate than parametric models.

The goal of vision is to interpret the image data in terms of what is where (that is, recognize objects and geometry) instead of synthesizing images. The result from our model-based analysis is represented by the sketch graph as well as non-sketchable statistics. The synthesized images are used for model checking to see if our model captures the essence of

the observed images to the extend that synthesized images generated by the fitted models are judged realistic by human vision. In many cases, this is still the harshest criterion to pass, especially for the regime where a physical description of the scene is not entirely perceptible. Moreover, the synthesis mode of the model can also be useful for computer graphics.

Acknowledgement

We thank Siavosh Bahrami for his skillful assistance on experiments presented in Section 2. We thank Yizhou Wang, Zhuowen Tu, Alan Yuille, and Stefano Soatto for insightful discussions. A shorter version of the paper has appeared in the Workshop on Generative Model Based Vision 2004, organized by Arthur Pece. The work is supported by NSF IIS-0222967.

References

[1] G. Aubert and P. Kornprobst, *Mathematical Problems in Image Processing*, Springer, 2002.

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

[3] A. Bell, and T.J. Sejnowski, "The 'Independent components' of natural scenes are edge filters", *Vision Research*, 37:3327-3338, 1997.

[4] J. Besag, "Spatial Interaction and the Statistical Analysis of Lattice Systems (with discussion)", *J. Royal Statist. Soc., series B*, vol.36. pp. 192-236, 1974.

[5] E. J. Candes and D. L. Donoho, "Ridgelets: a key to higher-dimen. intermittency?" *Phil. Trans. R. Soc. Lond. A.*, 357, 2495-509, 1999.

[6] J. Canny, "A computational approach to edge detection", *IEEE Trans. PAMI*, 36:961-1005, 1986.

[7] T. F. Chan and J. Shen, "Mathematical models for local nontexture inpaintings." *SIAM J. Appl. Math.*, 62(3):1019–1043, 2001.

[8] D. Chandler, *Introduction to Modern Statistical Mechanics*, Oxford University Press, 1987.

[9] Z. Chi, "Stationary self-similar random fields on the integer lattice", *Stoch. Proc. and Appl*, 91:99-113, 2001

[10] H.A. Chipman, E.D. Kolaczyk, and R.E. McCulloch, "Adaptive Bayesian Wavelet Shrinkage." *Journal of the American Statistical Association*, 92, 1413-1421, 1997.

[11] N. Cressie, *Statistics for Spatial Data*, revised edition, Wiley, NY, 1993

[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[13] J. Daugma, "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] D.L. Donoho, M. Vetterli, R.A. DeVore, and I. Daubechie, "Data compression and harmonic analysis", *IEEE Trans. Information Theory*. 6, 2435-2476, 1998.

[15] J.H. Elder and S. W. Zucker, "Local scale control for edge detection and blur estimation." *IEEE Trans. PAMI*, vol. 20, no. 7, 699-716, 1998.

[16] D.J. Field, "What is the goal of sensory coding?", *Neural Computation*. 6, 559-601, 1994. Friedman, J. H. (1987)

[17] J. H. Friedman, "Exploratory Projection Pursuit." *J. Amer. Statist. Assoc.*, 82, 249, 1987.

[18] S. Geman and D. Geman. "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images", *IEEE Trans. PAMI*, 6:721-741, 1984.

[19] D. Geman and A. Koloydenko, "Invariant statistics and coding of natural microimages", *1st IEEE Workshop on Stat. and Comp. Theories of Vision*, Fort Collins, Co. 1999.

[20] U. Grenander, *General Pattern Theory*, Oxford Univ Press, 1993.

[21] E.I. George and R.E. McCulloch, "Approaches to Bayesian variable selection", *Statistica Sinica*, 7:339-373, 1997.

[22] U. Grenander and M. I. Miller, "Representation of knowledge in complex systems", *J. R. Stat. Soc., B*, vol 56, 549-603, 1994.

[23] C. E. Guo, S. C. Zhu and Y. N. Wu, "Integrating Structures and Textures", Technical Report, 2004.

[24] J. Hammersley and P. Clifford, *Markov Fields on Finite Graphs and Lattices*, Preprint, UC. Berkeley, 1968.

[25] M. Hansen and B. Yu (2000). Wavelet thresholding via MDL for natural images. IEEE Trans. Inform. Theory (Special Issue on Information Theoretic Imaging). vol. 46, 1778-1788.

[26] D. J. Heeger and J. R. Bergen, "Pyramid Based Texture Analysis/Synthesis", *Computer Graphics Proc.*, pp. 229-238, 1995.

[27] D. Huber and T. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex", *J. of Physiology*, 160, 1962.

[28] O. Johnson, "An information theoretical central limit theorem for finitely susceptible FKG systems", technical report, 2004.

[29] E.D. Kolaczyk, Bayesian Multi-Scale Models for Poisson Processes. Journal of the American Statistical Association, 94, 920-933, 1999.

[30] T. S. Lee, "Image Representation Using 2D Gabor Wavelets", *IEEE Trans. PAMI*, 10, 959-971, 1996.

[31] T. Lee, "An Introduction to Coding Theory and the Two-Part Minimum Description Length Principle", *International Statistical Review* 69, 169-183, 2001.

[32] 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.

[33] T. Lindeberg, *Scale-Space Theory in Computer Vision*, Kluwer Academic Publishers, Netherlands, 1994.

[34] A. Lee, D. Mumford, and J. Huang, "Occlusion Models for Natural Images: A Statistical Study of a Scale-Invariant Dead Leaves Model," *Int'l Journal of Computer Vision*, Vol. 41(1/2), pp. 35-59, 2001.

[35] J. Malik and R. Perona, "Preattentive Texture Discrimination with Early Vision Mechanisms", *J. Opt. Soc. AM* , Vol. 7, No. 5, pp. 923-932, 1990.

[36] S. Mallat, "A Theory of Multiresolution Signal Decomposition: the Wavelet Representation", *IEEE Trans. PAMI*, 11(7):674-693, 1989.

[37] S. Mallat and Z. Zhang, "Matching Pursuit in a Time-Frequency Dictionary", *IEEE Sig. Proc.*, 41, 3397-415, 1993.

[38] B.B. Mandelbrot, *The fractal Geometry of Nature*, S.F. CA. Freeman, 1982

[39] D. Marr, *Vision,* W. H. Freeman and Company, 1982.

[40] S. G. Matheron, *Random Sets and Integral Geometry*, John Wiley and Sons, 1975.

[41] D. B. Mumford, "Pattern Theory: a Unifying Perspective", *Proc. of 1st European Congress of Mathematics*, Birkhauser-Boston, 1994.

[42] D. Mumford and B. Gidas, "Stochastic models for generic images", *Quarterly of Applied Math*, 59(1), 85-111, 2001.

[43] B. A. Olshausen and D. J. Field, "Emergence of Simple-cell Receptive Field Properties by Learning a Sparse Code for Natural Images," *Nature*, Vol. 381, pp. 607-609, 1996.

[44] B.A. Olshausen and K.J. Millman, "Learning Sparse Codes with a Mixture-of-Gaussians Prior", *Advances in Neural Information Processing Systems*, 12, Ed. by S.A. Solla, T.K. Leen, and K.R. Muller, MIT Press, pp. 841-847, 2000.

[45] A. Pece, "The Problem of Sparse Image Coding," *Journal of Mathematical Imaging and Vision*, vol. 17(2), pp. 89-108, 2002.

[46] J. Portilla and E.P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients". *Int'l Journal of Computer Vision*, 40(1):49-71, October, 2000.

[47] J. Rissanen, *Stochastoc complexity in statistical inquiry*, Singapore: World Scientific, 1989.

[48] D. L. Ruderman, "Origins of scaling in natural images", *Vision Research*, 37(23), 3385-3395, 1997.

[49] D.L. Ruderman and W. Bialek, "Statistics of Natural images: Scaling in the Woods", *Phy. Rev. Lett*, 73, 1994.

[50] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation". *Annual Review of Neuroscience*, 24:1193-1216, May 2001.

[51] A. Srivastava, U. Grenander, and X. Liu, "Universal Analytical Forms for Modeling Image Probabilities", *IEEE Pattern Analysis and Machine Intelligence* 24(9), 1200-1214, September, 2002.

[52] 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

[53] D. Stoyan, W. Kendall, and J. Mecke. *Stochastic Geometry and Its Applications*. John Wiley & Sons, 1987.

[54] G. Winkler, *Image Analysis, Random Fields, and Dynamic Monte Carlo Methods*, Springer, 1995.

[55] Y. Wu, S. C. Zhu, and C. Guo, "Statistical modeling of texture sketch". *Proc. of European Conference of Computer Vision.* 2002.

[56] Y. N. Wu, S. C. Zhu, and X. W. Liu, "Equivalence of Julesz Ensemble and FRAME Models", *Int'l Journal of Computer Vision*, 38(3):245–261, 2000.

[57] S.C. Zhu and D.B. Mumford, "Prior learning and Gibbs reaction-diffusion", *IEEE Trans. on PAMI*, vol.19, no.11, pp1236-1250, 1997.

[58] 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.