Statistical Principles in Image Modeling

Ying Nian Wu, Jinhui Li, Ziqiang Liu, and Song-Chun Zhu

Department of Statistics
University of California Los Angeles
Los Angeles, CA 90024
(*ywu@stat.ucla.edu*)

Images of natural scenes contain a rich variety of visual patterns. To learn and recognize these patterns from natural images, it is necessary to construct statistical models for these patterns. In this review article we describe three statistical principles for modeling image patterns: the sparse coding principle, the minimax entropy principle, and the meaningful alignment principle. We explain these three principles and their relationships in the context of modeling images as compositions of Gabor wavelets. These three principles correspond to three regimes of composition patterns of Gabor wavelets, and these three regimes are connected by changes in scale or resolution.

KEY WORDS: Gabor wavelet; Meaningful alignment; Minimax entropy; Scale; Sparse coding.

1. INTRODUCTION

1.1 Three Regimes of Image Patterns and Scaling Connection

Images of natural scenes are characterized by a bewildering richness of visual patterns. The goal of computer vision is to learn and recognize these patterns from natural images. To accomplish this goal, it is necessary to construct statistical models for these patterns. The parameters of these models can be learned from training image data. The learned models then can be used to interpret future image data to recognize the objects and patterns. Thus vision is essentially a statistical modeling and inference problem.

But vision also proves to be a highly specialized modeling and inference problem. A special property of vision that distinguishes it from other recognition tasks, such as speech recognition, is that visual objects in natural scenes can appear at a wide range of distance from the camera. At different viewing distances, the objects project different sizes (sometimes even smaller than a single pixel) on the resulting image and thus create different image patterns. To illustrate this concept, Figure 1 displays three images of maple leaves at three different viewing distances (with the scope of the camera remaining fixed). Image (a) is observed at a far viewing distance. In this image maple leaves are so small and densely packed that the individual leaf shapes cannot be recognized. The leaves collectively produce a foliage texture pattern. Image (b) is observed at a medium distance. The leaf sizes are larger than those in image (a), and their shapes can be individually recognized. Image (c) is observed at a very close distance. The overall shape of a single maple leaf is larger than the scope of the camera, so that only an edge pattern is recognized within this image. These three images have different statistical properties. Image (a) is very random, image (c) is very simple, and image (b) is in between. The three images are instances of three regimes of image patterns. Image (a) is from the texture regime, where no clear shapes exist. Image (b) is from the object regime, where the overall shapes of the objects can be recognized. Image (c) is from the geometry regime, which comprises lines, regions, corners, junctions, and so on. From Figure 1, it is clear that these three regimes of image patterns are connected by the variability of viewing distance. This variability is a primary cause for the richness of visual patterns. The effect of varying viewing distance can be equally achieved by varying the camera resolution with zooming-in and zooming-out operations.

1.2 Multiscale Representations for Both Image Data and Their Patterns

The variability of viewing distance requires that the visual system be able to recognize patterns at different scales or resolutions. For a fixed image, it is natural to use this ability to simultaneously analyze the image at multiple scales or resolutions.

Figure 2(a) displays an image at multiple resolutions in a pyramidal structure. The original image is at the bottom layer. On top of that, the image at each layer is a zoomed-out version of the image at the layer beneath. The zooming-out operation can be accomplished by smoothing the current image by convolving it with a Gaussian kernel of a certain standard deviation, and then subsampling the pixels by a certain factor to make the image smaller. This pyramidal structure, called a Gaussian pyramid (Burt and Adelson 1983), can have many layers. Each layer of the Gaussian pyramid is said to be a low-pass version of the original image, because only low-frequency content of the image is preserved. A companion structure is called a Laplacian pyramid. The image at each layer of the Laplacian pyramid is the difference between the images at two adjacent layers of the Gaussian pyramid or, more precisely, the difference between the image at the corresponding layer of the Gaussian pyramid and its smoothed version obtained by convolving it with the Gaussian kernel. Each layer of the Laplacian pyramid is said to be a bandpass version of the original image, because only the content within a certain frequency band is preserved. The Laplacian pyramid is a decomposition of the original image into different frequency bands. (See Sec. 2.2 for explanations of low-pass and bandpass images.)

If we move a window of a certain size (say, 40×40) over the images at multiple layers in the Gaussian or Laplacian pyramid,

TECHNOMETRICS, AUGUST 2007, VOL. 49, NO. 3
DOI 10.1198/004017007000000236

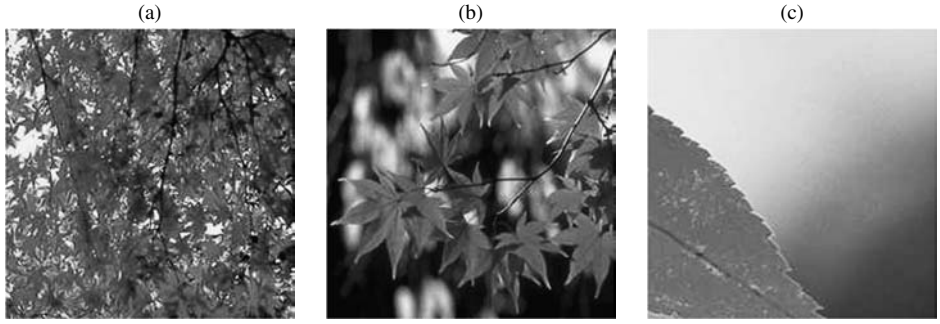

Figure 1. Examples from three regimes of image patterns: (a) texture regime, (b) object regime, and (c) geometry regime. The three regimes are connected by the change in viewing distance or, equivalently, the change in camera resolution.

then we can recognize different image patterns within the window, and these patterns belong to different regimes described in the previous section. We recognize lines and regions at high resolution or small scale, shapes and objects at medium resolution or scale, and textures and clutters at low resolution or large scale. In addition, these patterns are highly organized. The large-scale texture patterns are composed of medium-scale shape patterns, which in turn are composed of small-scale edge and corner patterns. Figure 2(b) provides an illustration, where the image patches at different resolutions are labeled as various patterns such as foliage, leaves, and edges or corners. The vertical and diagonal arrows indicate the compositional relationships, and the horizontal arrows indicate the spatial arrangements. Thus a complete interpretation of an image must consist of the labels of image patches at all of these layers. It also should include the compositional and spatial organizations of the labels.

Figure 2 illustrates the multiscale structures of both the image data and the recognized patterns. The multiscale structure of the image data has been extensively studied in wavelet theory (Mallat 1989; Simoncelli, Freeman, Adelson, and Heeger 1992) and scale-space theory (Witkin 1983; Lindberg 1994); however, the multiscale structure of the recognized patterns has not received as much treatment. Recent attempts to understand this issue have been made by Wu, Zhu, and Guo (2007) and Wang, Bahrami, and Zhu (2005). The compositional relationships have been studied by Geman, Potter, and Chi (2002). The AND–OR grammars for the multilayer organizations of the image patterns have been studied by Zhu and Mumford (2007).

1.3 Bayesian Generative Modeling and Posterior Inference

An elegant approach to recognizing the multiscale patterns from the multiscale image data is to construct a Bayesian generative model and use the posterior distribution to guide the inferential process (Grenander 1993; Grenander and Miller 1994; Geman and Geman 1984; Mumford 1994). Given the multilayer structures of both the image data and the labels of the recognized patterns as shown in Figure 2(b), a natural form of this model is a recursive coarse-to-fine generative process consisting of the following two schemes:

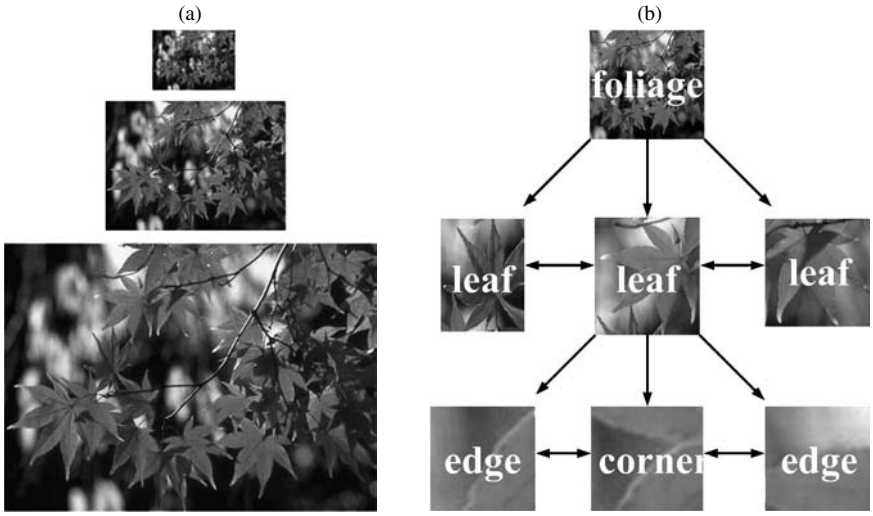

Figure 2. Multiscale representations of data and patterns. (a) A Gaussian pyramid. The image at each layer is a zoomed-out version of the image at the layer beneath. The Laplacian pyramid can be obtained by computing the differences between consecutive layers of the Gaussian pyramid. (b) For the same image, different patterns are recognized at different scales. These patterns are organized in terms of compositional and spatial relationships. The vertical and diagonal arrows represent compositional relationships; the horizontal arrows, spatial relationships.

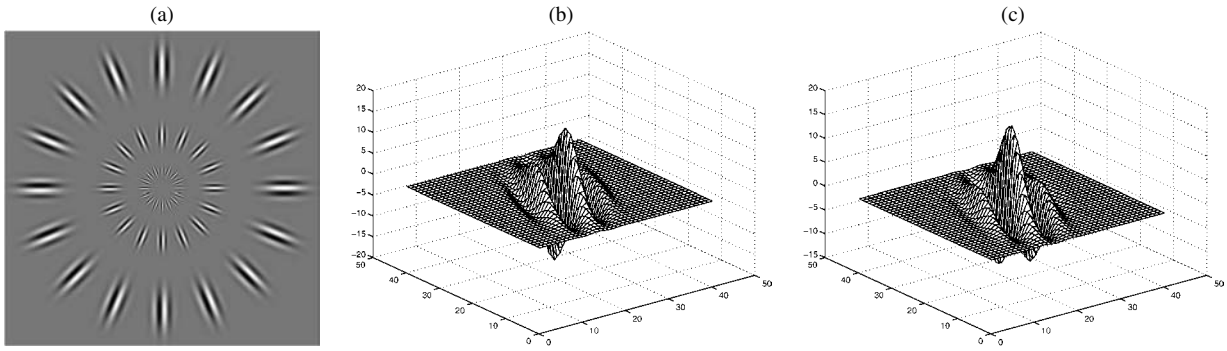

Figure 3. Gabor wavelets. (a) A sample of Gabor wavelets at different locations, scales, and orientations. (b) An example of a Gabor sine wavelet. (c) An example of a Gabor cosine wavelet.

Scheme 1: The labels generate their image data at the corresponding scale or frequency band; for example, a leaf pattern generates a leaf image patch at the same layer.

Scheme 2: The labels generate those constituent labels at the lower scale; for example, a leaf pattern is decomposed into edge and corner patterns at the lower layer.

These two schemes are closely related and lead to a joint distribution of the labels of the recognized patterns and the bandpass image patches over all of the scales, Pr(labels, image patches). In principle, this model can be learned from training images that have already been labeled. A new image can be interpreted by sampling from or maximizing the posterior distribution Pr(labels|image patches).

In the posterior distribution Pr(labels|image patches), the label of a particular image patch at a certain layer is determined not only by this image patch itself, but also by the labels of other image patches. For instance, the leaf pattern in the center of Figure 2(b) can be recognized by combining the following four sources of information:

Source 1. The image patch of this leaf at the current layer

Source 2. The bottom-up information from the edge and corner patterns of the image patches at the layers below

Source 3. The top-down information from the foliage patterns of the image patches at the layers above

Source 4. The contextual information from the leaf patterns of the adjacent image patches at the same layer.

Although source 1 is the most direct, the other sources also can be important, especially when source 1 contains too much ambiguity, which can often be the case in natural images. The combination of information can be accomplished by Bayesian posterior computation.

A realistic generative model, Pr(labels, image patches), based on the two schemes mentioned earlier, remains beyond our reach. The posterior computation for propagating and combining information also is not well understood. In this article we study statistical principles that may eventually lead to such a generative model. The focus of this article is on Scheme 1—that is, how the patterns generate their image data at the same layer. A recent article by Zhu and Mumford (2007) on the stochastic AND–OR grammars for Scheme 2 explores how the patterns are decomposed into the constituent patterns at lower layers in the compositional hierarchy.

1.4 Hint From Biological Vision

Given the fact that biological vision is far superior to computer vision in terms of learning and recognizing natural image patterns, it is useful to take some clues from biological vision when constructing image models to be used for computer vision. In neuroscience, neuron recording data (Hubel and Wiesel 1962) indicate that visual cells in the primary visual cortex or V1 area (the part of the brain responsible for the initial stage of visual processing) respond to local image structures, such as bars, edges, and gratings at different positions, scales, and orientations.

Specifically, the V1 cells are classified into simple cells and complex cells. The simple cells are modeled by Gabor wavelets (Daugman 1985). Figure 3(a) displays a sample of Gabor wavelets (Lee 1996), which are localized functions at different locations with different orientations and scales. These functions are in the form of sinusoidal waves multiplied by Gaussian functions. Figures 3(b) and 3(c) display the perspective plots of two such functions. A more detailed explanation is provided in Section 2. Responses of simple cells are modeled as projection coefficients of the image onto these functions. These projection coefficients capture image structures at different locations, orientations, and scales. The complex cells are also sensitive to similar local image structures; however, they are more tolerant or invariant of the shifting of locations and scales of the image structures (Lampl, Ferster, Poggis, and Riesenhuber 2004). Such nonlinear behavior can be modeled by combining or pooling the outputs from the simple cells (Adelson and Bergen 1985; Riesenhuber and Poggio 1999).

The set of Gabor wavelets is self-similar and has a multiscale structure that is consistent with the multiscale structures of the image data described in the previous section. Specifically, projecting an image onto a large-scale Gabor wavelet is equivalent to projecting a zoomed-out version of this image onto a smaller-scale Gabor wavelet at the same location and orientation. These Gabor wavelets can be the link between the image data and the corresponding patterns. Specifically, these Gabor wavelets may serve as the building elements in the generative model Pr(image patches|labels), as well as the intermediate step in computing Pr(labels|image patches).

1.5 Three Statistical Principles for Composing Gabor Wavelets

In this article we examine three existing statistical principles for image modeling. These principles shed light on the roles of Gabor wavelets in pattern recognition and provide useful guidance for constructing image models as compositions or poolings of Gabor wavelets.

1.5.1 Sparse Coding Principle. Olshausen and Field (1996) proposed that sparse coding is a strategy used by the primary visual cortex to represent image data. This principle seeks to find a dictionary of linear basis elements, so that any typical natural image can be expressed as a linear combination of a small number of basis elements chosen from the dictionary. The basis elements learned by Olshausen and Field from natural images exhibit close resemblance to the Gabor wavelets.

1.5.2 Minimax Entropy Principle. Zhu, Wu, and Mumford (1997) proposed the minimax entropy principle for modeling textures. This principle seeks to find statistical summaries of the observed image so that the maximum entropy model constrained by these statistical summaries has the minimum entropy. The particular statistical summaries adopted by Zhu et al. are marginal histograms of Gabor wavelet coefficients. The resulting models are in the form of Markov random fields (Besag 1974; Geman and Graffigne 1987). Realistic texture patterns can be generated by sampling from such random fields.

1.5.3 Meaningful Alignment Principle. Moisan, Desolneux, and Morel (2000) proposed the meaningful alignment principle for perceptual grouping. This principle seeks to identify coincidence patterns in the observed image data that otherwise would be extremely rare in a completely random image. Such coincidences are said to be meaningful. One particular example of this is the detection of line segments in image data based on the alignment of local orientations along the potential line segments. The local orientations can be computed as the orientations of the best-tuned Gabor wavelets.

These three principles correspond to the three regimes of image patterns discussed in Section 1.1. The sparse coding principle corresponds to the object regime, where the image can be modeled based on the composition of a small number of wavelet elements at different positions, scales, and orientations. The minimax entropy principle corresponds to the texture regime, where the image is modeled by pooling the wavelet coefficients into marginal histograms while discarding the position information. The meaningful alignment principle corresponds to the geometry regime, where the best-tuned Gabor wavelets are highly aligned in both the spatial and frequency domains.

Although these three principles correspond to three different regimes of patterns with different complexities and scales, they all seek to identify a small number of constraints on the Gabor wavelet coefficients to restrict the observed image to an image ensemble of small volume. As a result, these constraints constitute a simple and informative description of the observed image. The uniform distribution on the constrained image ensemble can be linked to an unconstrained statistical model. This point of view suggests that it is possible to develop a unified statistical model for all these three regimes of image patterns as the composition or pooling of Gabor wavelet coefficients.

The rest of the article is organized as follows. Section 2 reviews Gabor wavelets. Section 3 reviews the three statistical principles, and Section 4 investigates the relationships between these principles. Section 5 concludes with a discussion.

2. GABOR WAVELETS: EDGE AND SPECTRUM

Gabor wavelets are localized in both spatial and frequency domains, and can detect edges and bars in the spatial domain and extract spectra in the frequency domain. In this section we review Gabor wavelets first in the spatial domain, and then in the frequency domain. For simplicity, we assume that both the image and the Gabor wavelets are functions defined on the two-dimensional real domain $\mathbb{R}^2$.

2.1 Edge-Bar Detector

A Gabor wavelet (Daugman 1985) is a sinusoidal wave multiplied by a Gaussian density function. The following function is an example:

$$G(x) = \frac{1}{2\pi\sigma_1\sigma_2}\exp\left\{-\frac{1}{2}\left(\frac{x_1^2}{\sigma_1^2}+\frac{x_2^2}{\sigma_2^2}\right)\right\}e^{ix_1}, \tag{1}$$

where $x = (x_1, x_2) \in \mathbb{R}^2$ and $i = \sqrt{-1}$. This complex-valued function $G(x)$ consists of two parts. The first part is an elongate Gaussian function with $\sigma_1 < \sigma_2$, so the shorter axis is along the x_1 direction and the Gaussian function is oriented along the x_2 direction. The second part is a pair of sine and cosine plane waves propagating along the x_1 axis—that is, the shorter axis of the Gaussian function. Because of the Gaussian function, $G(x)$ is localized in the spatial domain. The Gabor cosine component or the real part of $G(x)$ is an even-symmetric function, and the Gabor sine component or the imaginary part of $G(x)$ is an odd-symmetric function (see Figs. 3(b) and 3(c) for illustrations).

We can translate, rotate, and dilate the function $G(x)$ of (1) to obtain a general form of Gabor wavelets such as those plotted in Figure 3(a): $G_{y,s,\theta}(x) = G(\tilde{x}/s)/s^2$, where $\tilde{x} = (\tilde{x}_1, \tilde{x}_2)$, $\tilde{x}_1 = (x_1 - y_1)\cos\theta - (x_2 - y_2)\sin\theta$, and $\tilde{x}_2 = (x_1 - y_1)\sin\theta + (x_2 - y_2)\cos\theta$. The function $G_{y,s,\theta}$ is centered at y, with orientation θ. The standard deviations of the two-dimensional Gaussian function in $G_{y,s,\theta}(x)$ along its shorter and longer axes are $s\sigma_1$ and $s\sigma_2$. The sine and cosine waves propagate at the frequency $1/s$ along the shorter axis of the Gaussian function.

For an image $\mathbf{I}(x)$, the projection coefficient of $\mathbf{I}$ onto $G_{y,s,\theta}$ is

$$r_{y,s,\theta} = \langle \mathbf{I}, G_{y,s,\theta}\rangle = \int \mathbf{I}(x)G_{y,s,\theta}(x)\,dx, \tag{2}$$

where the integral is over $\mathbb{R}^2$. In engineering literature, the Gabor wavelets are also called Gabor filters, $r_{y,s,\theta}$ obtained by (2) is called the filter response, and the image $\mathbf{J}(y) = r_{y,s,\theta}$ is called the filtered image. The Gabor sine component or the imaginary part of $G_{y,s,\theta}$ is odd-symmetric and sensitive to step-edge structures. If the image $\mathbf{I}$ has a step-edge structure at location y with orientation θ, then the imaginary part of $r_{y,s,\theta}$ will be of large magnitude. The Gabor cosine component of $G_{y,s,\theta}$ is even-symmetric and sensitive to bar structures. If the image $\mathbf{I}$ has a bar structure at location y with orientation θ, then the real part of $r_{y,s,\theta}$ will be of large magnitude.

The Gabor wavelets can be used as edge-bar detectors (Wang and Jenkin 1992; Perona and Malik 1990; Mallat and Zhong 1992; Canny 1986). Let $r_{x,s,\theta} = a_{x,s,\theta} + ib_{x,s,\theta}$. Here $|r_{x,s,\theta}|^2 = a_{x,s,\theta}^2 + b_{x,s,\theta}^2$ is the energy extracted by the Gabor wavelet $G_{x,s,\theta}$ (Adelson and Bergen 1985). For fixed x and s, let $\theta_* =$

Figure 4. The observed image (a) and edge maps at two different scales [(b) and (c)]. The scale of (b) is finer than that of (c).

$\arg\max_\theta |r_{x,s,\theta}|^2$. We call G_{x,s,θ_*} the best-tuned Gabor wavelet at (x, s). The local orientation is defined as $\theta(x, s) = \theta_*$. The local phase is defined as $\phi(x, s) = \arctan(b_{x,s,\theta_*}/a_{x,s,\theta_*})$. For edge structure, the local phase is $\pi/2$ or $-\pi/2$; for bar structure, the local phase is 0 or π. The local energy is defined as $A(x, s) = |r_{x,s,\theta_*}|^2$. Here x is an edge-bar point at scale s if $A(x, s) \geq A(y, s)$ for all y such that the direction of $y - x$ is perpendicular to $\theta(x, s)$ and $|y - x| < d$ for predefined range d; that is, $A(x, s)$ is a local maximum along the normal direction of x, and we call $G_{x,s,\theta(x,s)}$ a locally best-tuned Gabor wavelet at scale s. Figures 4(b) and 4(c) display the edge-bar points of the observed image in Figure 4(a) at two different scales, where the intensity of an edge-bar point (x, s) is proportional to $A(x, s)^{1/2}$. The scale s for computing the edge map is smaller in Figure 4(b) than in Figure 4(c).

2.2 Spectral Analyzer

The Gabor wavelets also extract power spectra. For an image $\mathbf{I}(x)$, which is a function in L^2, this can be represented as the superposition of sinusoidal waves

$$\mathbf{I}(x) = \frac{1}{4\pi^2} \int \hat{\mathbf{I}}(\omega) e^{i\omega x}\, d\omega, \tag{3}$$

where $x = (x_1, x_2) \in \mathbb{R}^2$, $\omega = (\omega_1, \omega_2) \in \mathbb{R}^2$, $\omega x = x_1\omega_1 + x_2\omega_2$, and the integral is over the two-dimensional frequency domain $\mathbb{R}^2$. For fixed ω, the plane wave $e^{i\omega x}$ in (3) propagates in the spatial domain $\mathbb{R}^2$ along the orientation of the vector $\omega = (\omega_1, \omega_2)$ at frequency $|\omega| = (\omega_1^2 + \omega_2^2)^{1/2}$. The coefficient $\hat{\mathbf{I}}(\omega)$ in (3) can be obtained by the Fourier transform $\hat{\mathbf{I}}(\omega) = \int \mathbf{I}(x) e^{-i\omega x}\, dx$.

Ideally, we can extract certain frequency content of $\mathbf{I}$ by calculating $4\pi^2\mathbf{J} = \int_F \hat{\mathbf{I}}(\omega) e^{i\omega x}\, d\omega$ for a certain set $F \subset \mathbb{R}^2$ of frequencies. If F contains only those frequencies ω such that $|\omega|$ are below a threshold, then $\mathbf{J}$ is called a low-pass image. If F contains only those frequencies ω such that $|\omega|$ are within a certain finite interval, then $\mathbf{J}$ is called a bandpass image. In practice, such images can be approximately extracted by convolving $\mathbf{I}$ with a kernel function g. The convoluted image is $\mathbf{J}(x) = \int \mathbf{I}(y) g(y - x)\, dy$. In the frequency domain, $\hat{\mathbf{J}}(\omega) = \hat{\mathbf{I}}(\omega)\hat{g}(\omega)$; thus $4\pi^2\mathbf{J}(x) = \int \hat{\mathbf{I}}(\omega)\hat{g}(\omega) e^{i\omega x}\, d\omega$. One may consider $\hat{g}(\omega)$ a generalized version of the indicator function for F and design $\hat{g}(\omega)$ to be a function localized in the frequency domain to extract the corresponding frequency content of $\mathbf{I}$.

The Gabor wavelets can serve this purpose. Let $G_{s,\theta}(x) = G_{0,s,\theta}(x)$. [See the previous section for the definition of $G_{y,s,\theta}$; here $y = (0, 0)$.] The Fourier transform of $G_{s,\theta}$ is

$$\hat{G}_{s,\theta}(\omega) = \exp\left\{-\frac{1}{2}\left[\left(s\sigma_1\left(\tilde{\omega}_1 - \frac{1}{s}\right)\right)^2 + (s\sigma_2\tilde{\omega}_2)^2\right]\right\}, \tag{4}$$

where $\tilde{\omega} = (\tilde{\omega}_1, \tilde{\omega}_2)$, with $\tilde{\omega}_1 = \omega_1 \cos\theta + \omega_2 \sin\theta$ and $\tilde{\omega}_2 = -\omega_1 \sin\theta + \omega_2 \cos\theta$. $\hat{G}_{s,0}$ is a Gaussian function centered at $(1/s, 0)$ with standard deviations $1/(s\sigma_1)$ and $1/(s\sigma_2)$ along ω_1 and ω_2, and $\hat{G}_{s,\theta}$ is a Gaussian function obtained by rotating $\hat{G}_{s,0}$ by an angle $-\theta$, and $\hat{G}_{s,\theta}$ is localized in the frequency domain. Figure 5 shows an example of $\hat{G}_{s,\theta}$ in the frequency domain for a sample of (s, θ) corresponding to the sample of Gabor wavelets displayed in Figure 3(a). Each $\hat{G}_{s,\theta}$ is illustrated by an ellipsoid whose two axes are proportional to $1/(s\sigma_1)$ and $1/(s\sigma_2)$. One may intuitively imagine each ellipsoid as the frequency band covered by $\hat{G}_{s,\theta}$, and together these $\hat{G}_{s,\theta}$ can pave or tile the entire frequency domain.

Let $\mathbf{J} = \mathbf{I} * G_{s,\theta}$. $\mathbf{J}(x) = \langle \mathbf{I}(y), G_{x,s,\theta}(-y) \rangle$ can be expressed as the projection of $\mathbf{I}$ onto the Gabor wavelet centered at x. We have that $4\pi^2\mathbf{J}(x) = \int \hat{\mathbf{J}}(\omega) e^{i\omega x}\, d\omega$, where $\hat{\mathbf{J}}(\omega) = \hat{\mathbf{I}}(\omega)\hat{G}_{s,\theta}(\omega)$. Thus $\mathbf{J}$ extracts the frequency content of $\mathbf{I}$ within the frequency band covered by $\hat{G}_{s,\theta}$. According to the Parseval identity, $4\pi^2\|\mathbf{J}\|^2 = 4\pi^2 \int |\mathbf{J}(x)|^2\, dx = \|\hat{\mathbf{J}}\|^2 = \int |\hat{\mathbf{I}}(\omega)|^2 |\hat{G}_{s,\theta}(\omega)|^2\, d\omega$. Thus $\|\mathbf{J}\|^2$ extracts the local average of the power spectrum of $\mathbf{I}$ within the frequency band covered by $\hat{G}_{s,\theta}$. Section 3.2 provides more discussion on this issue.

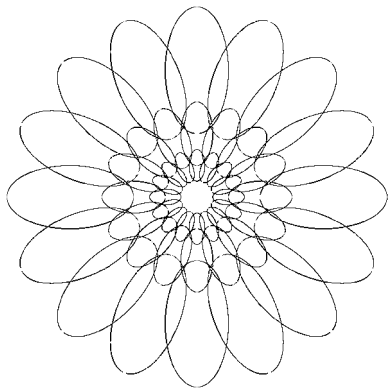

Figure 5. A sample of $\hat{G}_{s,\theta}$ in the frequency domain. Each $\hat{G}_{s,\theta}$ is illustrated by an ellipsoid whose two axes are proportional to the corresponding standard deviations of $\hat{G}_{s,\theta}$.

In this section we explain that Gabor wavelets can be used to extract edge-bar points as well as spectra. However, how to use Gabor wavelets to represent and recognize image patterns remains unclear. In the next section we review three statistical principles that shed light on this issue.

3. THREE STATISTICAL PRINCIPLES

3.1 Sparse Coding Principle: Beyond Edge Detection

The sparse coding principle can be used to justify the Gabor wavelets as linear basis elements for coding natural images. Specifically, let $\{\mathbf{B}_n(x), n = 1, \ldots, N\}$ be a set of linear basis functions. Then we can code $\mathbf{I}$ by

$$\mathbf{I}(x) = \sum_{n=1}^{N} c_n \mathbf{B}_n(x) + \epsilon(x), \tag{5}$$

where c_n are the coefficients and ϵ is the residual image. The dictionary of the basis functions $\{\mathbf{B}_n\}$ in the representation (5) should be designed or learned so that for any typical natural image $\mathbf{I}$, only a small number of coefficients c_n need to be nonzero to code $\mathbf{I}$ with small residual error. The dictionary of the basis elements $\{\mathbf{B}_n\}$ can be overcomplete; that is, the number of basis elements N can be greater than the number of pixels in $\mathbf{I}$.

Using this principle, Olshausen and Field (1996) learned a dictionary of basis elements from a random sample of natural image patches. Specifically, they collected a large number of 12×12 image patches $\{\mathbf{I}_m, m = 1, \ldots, M\}$ from pictures of natural scenes, and learned $\{\mathbf{B}_n, n = 1, \ldots, N\}$ (with $N > 144$) by minimizing the objective function

$$\sum_{m=1}^{M} \left[\left\| \mathbf{I}_m - \sum_{n=1}^{N} c_{m,n} \mathbf{B}_n \right\|^2 + \lambda \sum_{n=1}^{N} S(c_{m,n}) \right] \tag{6}$$

over both $\{\mathbf{B}_n\}$ and $\{c_{m,n}\}$. Here λ is a tuning parameter, and $\sum_n S(c_{m,n})$ measures the sparsity of the coefficients $\{c_{m,n}, n = 1, \ldots, N\}$. The most natural sparsity measure is the number of nonzero elements in $\{c_{m,n}\}$, that is, $S(c) = 0$ if $c = 0$ and $S(c) = 1$ if $c \neq 0$. This is the l_0-norm of the sequence $\{c_{m,n}, n = 1, \ldots, N\}$. For computational convenience, one may choose l_1-norm (Chen, Donoho, and Saunders 1999) or other functions to approximate the l_0-norm, so that gradient-based algorithms can be used for minimization. The basis elements learned by Olshausen and Field (1996) resemble the Gabor wavelets; that is, for each learned $\mathbf{B}_n$, we can approximate it by the sine or cosine component of a Gabor wavelet $G_{x,s,\theta}$ for some (x, s, θ) with small approximation error.

Suppose that the dictionary of basis functions is already given [e.g., a dictionary of Gabor wavelets $\{G_{x,s,\theta}\}$]; then computing the sparse coding of an image $\mathbf{I}$ amounts to selecting a small number of basis functions from the dictionary so that the least squares regression of $\mathbf{I}$ on these selected basis functions has a very small residual error. This is the variable selection problem in linear regression.

A statistical formulation is the generative model (Lewicki and Olshausen 1999)

$$\mathbf{I} = \sum_{n=1}^{N} c_n \mathbf{B}_n + \epsilon, \qquad C = (c_n, n = 1, \ldots, N) \sim p(C), \tag{7}$$

where $p(C)$ is often assumed to have independent components and each c_n is assumed to follow a mixture of a normal distribution and a point mass at 0 (Pece 2002; Olshausen and Millman 2000). This is the Bayesian variable selection model in linear regression (George and McCulloch 1997).

An efficient algorithm for selecting the basis elements to code a given image is the matching-pursuit algorithm (Mallat and Zhang 1993), which is a stepwise forward-regression algorithm. Assume that all of the basis elements are normalized to have unit L^2 norm. The algorithm starts from the empty set of basis elements. At the kth step, let $\{\mathbf{B}_1, \ldots, \mathbf{B}_k\}$ be the set of elements selected, with coefficients $\{c_1, \ldots, c_k\}$. Let $\epsilon = \mathbf{I} - (c_1\mathbf{B}_1 + \cdots + c_k\mathbf{B}_k)$. Then we choose an element $\mathbf{B}_{k+1}$ from the dictionary so that the inner product $\langle \epsilon, \mathbf{B}_{k+1} \rangle$ is maximum among all of the elements in the dictionary. Then we add $\mathbf{B}_{k+1}$ to the current set of elements and record its coefficient, $c_{k+1} = \langle \epsilon, \mathbf{B}_{k+1} \rangle$. We repeat this process until $\|\epsilon\|^2$ is less than a prefixed threshold. Chen et al. (1999) have provided a more sophisticated computational method.

Figure 6 illustrates the sparse coding representation using the matching-pursuit algorithm. Figure 6(a) is an observed 185×250 image. Figure 6(b) is the image reconstructed by $c_1\mathbf{B}_1 + \cdots + c_K\mathbf{B}_K$ using the matching-pursuit algorithm. Here $K = 343$, and the 343 elements are selected from a dictionary similar to those Gabor wavelets illustrated in Figure 3 (Young 1987; Wu, Zhu, and Guo 2002). Thus there are $185 \times 250/343 \approx 135$ folds of dimension reduction in this sparse coding. Part (c) displays a symbolic representation of the selected 343 basis elements, with each element represented by a bar of the same location, orientation, and length. The intensity of the bar is proportional to the magnitude of the corresponding coefficient.

From Figure 6(c), it is evident that the selected basis elements target the edge-bar structures in the image. Recall that in

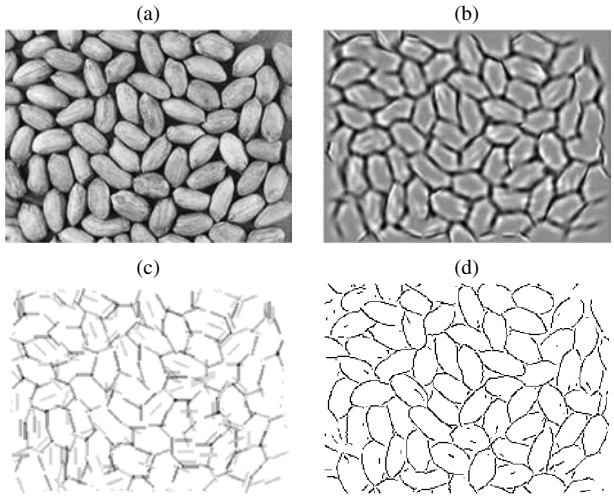

Figure 6. The sparse coding representation using the matching-pursuit algorithm. (a) An observed 185×250 image. (b) The image reconstructed by 343 linear basis elements selected by the matching-pursuit algorithm. (c) A symbolic sketch in which each selected basis element is represented by a bar of the same location, orientation, and length. (d) An edge-bar map created by the edge-bar detector described in Section 2.1. There are 4,099 edge-bar points.

Section 2.1, the edge-bar points are detected by the locally best-tuned Gabor wavelets. One may ask whether the sparse coding principle offers anything more than edge detection. The answer lies in sparsity. Figure 6(d) displays the map of edge-bar points detected at a small scale. This edge map may be more visually pleasing than Figure 6(c); however, there are 4,099 edge points, far more than the number of elements in sparse coding (343). To represent a local shape in Figure 6(a), a few basis elements are sufficient, but a large number of edge points is required for this task. From a modeling perspective, it is much easier to model a low-dimensional structure than a high-dimensional structure. Such models have been studied by Wu et al. (2002), Guo, Zhu, and Wu (2003a,b), Wang and Zhu (2004), and Zhu, Guo, Wang, and Xu (2005). We discuss object modeling in Section 4.1. (See also Candes and Donoho 1999 and Huo and Chen 2005 for curvelet and beamlet systems for sparse coding.)

3.2 Minimax Entropy Principle: Beyond Spectral Analysis

The minimax entropy principle was adopted by Zhu et al. (1997) for modeling textures. For convenience, we assume in this section that $\mathbf{I}$ is a stationary texture image defined on a finite lattice D with $|D|$ pixels; for instance, $D = \{1, \ldots, M\} \times \{1, \ldots, N\}$, so $|D| = M \times N$. A popular choice of texture statistics is the marginal histograms of wavelet coefficients (Heeger and Bergen 1995). Let $\{G_k, k = 1, \ldots, K\}$ be a set of kernel functions—for example, the Gabor sine and cosine functions $G_{s,\theta}$ defined in Section 2.2. Let H_k be the marginal histogram of the convoluted image $\mathbf{I} * G_k$. Specifically, we divide the range of $[\mathbf{I} * G_k](x)$ into T bins $\Delta_1, \ldots, \Delta_T$, so that

$$H_{k,t}(\mathbf{I}) = \sum_x \delta\big([\mathbf{I} * G_k](x) \in \Delta_t\big), \qquad t = 1, \ldots, T, \tag{8}$$

where $\delta(\cdot)$ is the indicator function. Let $h_{k,t}(\mathbf{I}) = H_{k,t}(\mathbf{I})/|D|$ be the normalized marginal histogram of $\mathbf{I} * G_k$. For simplicity, we write $H_k = (H_{k,t}, t = 1, \ldots, T)$ and $h_k = (h_{k,t}, t = 1, \ldots, T)$.

Wu, Zhu, and Liu (2000) defined the following image ensemble: $\Omega = \{\mathbf{I} : h_k(\mathbf{I}) = h_k(\mathbf{I}_{\text{obs}}), k = 1, \ldots, K\}$. If the set of marginal histograms $\{h_k\}$ captures the texture information in $\mathbf{I}_{\text{obs}}$, then all the images in the ensemble Ω should share the same texture pattern. Thus we can model the image $\mathbf{I}_{\text{obs}}$ as a random sample from the uniform distribution over Ω. Figure 7 shows an example of this. Figure 7(a) is the observed image $\mathbf{I}_{\text{obs}}$, whereas Figure 7(b) is an image randomly sampled from Ω, where $\{G_k\}$ is a set of Gabor kernel functions $\{G_{s,\theta}\}$. Although the sampled image in Figure 7(b) is a different image than the observed one in Figure 7(a), the two images share identical texture patterns judged by human visual perception (see Zhu, Liu, and Wu 2000 for details).

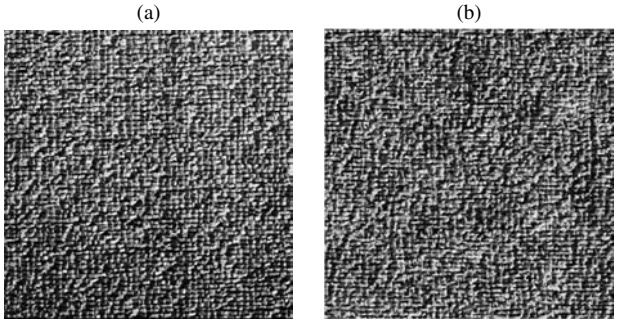

Figure 7. Observed image $\mathbf{I}_{\text{obs}}$ (a) and a random image sampled from the image ensemble $\Omega = \{\mathbf{I} : h_k(\mathbf{I}) = h_k(\mathbf{I}_{\text{obs}}), k = 1, \ldots, K\}$ (b).

In terms of the image ensemble Ω, the meaning of the minimax entropy principle can be stated as follows:

1. For a fixed set of kernel functions $\{G_k\}$, the maximum entropy model is the uniform distribution over Ω.
2. To select the set of kernel functions $\{G_k, k = 1, \ldots, K\}$ (with prespecified K) from a dictionary (e.g., a dictionary of Gabor functions), we want to select the set $\{G_k, k = 1, \ldots, K\}$ to minimize the volume of Ω, (i.e., $|\Omega|$).

For the uniform distribution over Ω, its entropy is $\log|\Omega|$, which is also the negative log-likelihood. By minimizing the volume of Ω, we are minimizing the entropy of the corresponding uniform distribution or maximizing its log-likelihood.

Although intuitively simple, the constrained image ensemble Ω or its entropy $\log|\Omega|$ is computationally difficult to handle. As pointed out by Wu et al. (2000), if the image size $|D|$ is large, then the uniform distribution over the constrained ensemble Ω can be approximated by an unconstrained statistical model,

$$p(\mathbf{I}; \Lambda) = \frac{1}{Z(\Lambda)} \exp\left\{\sum_{k=1}^{K} \langle \lambda_k, H_k(\mathbf{I})\rangle\right\}, \tag{9}$$

where $\Lambda = (\lambda_k, k = 1, \ldots, K)$, $\lambda_k = (\lambda_{k,t}, t = 1, \ldots, T)$, and $Z(\Lambda)$ is the normalizing constant to make $p(\mathbf{I}; \Lambda)$ integrate to 1. Here Λ can be estimated by maximizing the log-likelihood $\log p(\mathbf{I}_{\text{obs}}; \Lambda)$, which amounts to solving the estimating equation $\mathrm{E}_\Lambda[H_k(\mathbf{I})] = H_k(\mathbf{I}_{\text{obs}})$ for $k = 1, \ldots, K$. The entropy $\log|\Omega|$ can be approximated by the negative log-likelihood of the fitted model $\log p(\mathbf{I}_{\text{obs}}; \Lambda)$. Thus the minimax entropy problem is equivalent to selecting K kernel functions $\{G_k\}$ and computing the corresponding $\{\lambda_k\}$ to maximize the log-likelihood $\log p(\mathbf{I}_{\text{obs}}; \Lambda)$.

An intuitive explanation for the approximation to the uniform distribution over Ω by a statistical model (9) is that under $p(\mathbf{I})$, $h_{k,t}(\mathbf{I}) = H_{k,t}(\mathbf{I})/|D| = \sum_x \delta([\mathbf{I} * G_k](x) \in \Delta_t)/|D| \to \Pr([\mathbf{I} * G_k](x) \in \Delta_t)$ as $|D| \to \infty$, if ergodicity holds. Here ergodicity means that the spatial averages go to the corresponding fixed expectations; that is, the random fluctuations in the spatial averages diminish in the limit. Because $p(\mathbf{I})$ assigns equal probabilities to images with fixed values of $\{h_k(\mathbf{I})\}$, $p(\mathbf{I})$ goes to a uniform distribution over an ensemble of images $\mathbf{I}$ with fixed $\{h_k(\mathbf{I})\}$.

The model (9) generalizes early versions of Markov random-field models (Besag 1974; Geman and Graffigne 1987). Another way to express (9) is $p(\mathbf{I}) \propto \prod_{k,x} f_k([\mathbf{I} * G_k](x))$, where f_k is a step function over the bins of the histogram H_k such that $f_k(r) \propto \exp\{\lambda_{k,t}\}$ if $r \in \Delta_t$. This is similar to the form of projection pursuit density estimation (Friedman 1987) except that for each f_k, there is the product over all pixels x.

Zhu et al. (1997) proposed a filter pursuit procedure to add one filter or kernel function G_k at a time to approximately maximize the log-likelihood $\log p(\mathbf{I}_{\text{obs}}; \Lambda)$ or minimize the entropy $\log|\Omega|$. Figure 8 displays an example of filter pursuit procedure; (a) is the observed image, and (b)–(e) are images sampled

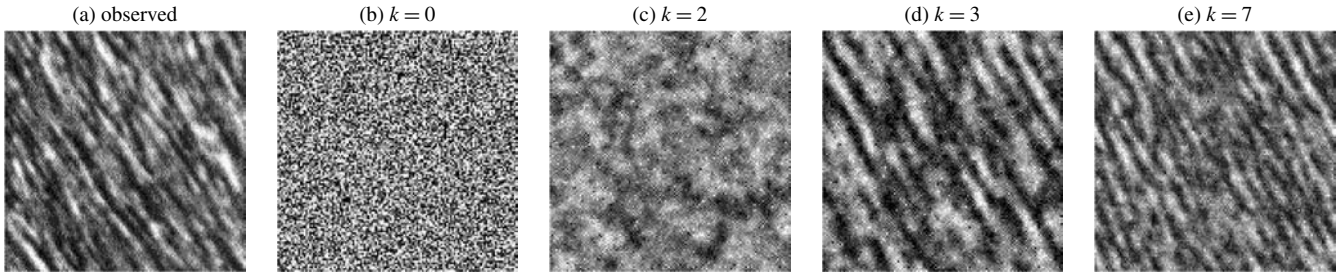

Figure 8. Filter pursuit: Adding one filter at a time to reduce the entropy. (a) The observed image. (b)–(e) Images sampled from the fitted models.

from the fitted models $p(\mathbf{I}; \Lambda)$ corresponding to an increasing sequence of the filter set $\{G_1, \ldots, G_k\}$ for $k = 0, \ldots, 7$. Specifically, for each k, after fitting the model $p(\mathbf{I}; \Lambda)$ to the observed image for filters $\{G_1, \ldots, G_k\}$, we sample an image from the fitted model $p(\mathbf{I}; \Lambda)$. With $k = 0$ filter, the sampled image is white noise. With $k = 7$ filters, the sampled image in (e) is perceptually equivalent to the input image. Della Pietra, Della Pietra, and Lafferty (1997) described earlier work on introducing features into the exponential models.

One may ask what is in the marginal histograms defined in (8) that is beyond the spectrum information extracted by the Gabor kernel functions. Without loss of generality, let $\mathbf{J} = \mathbf{I} * G$, where G is the Gabor function defined in Section 2.1. Suppose that we normalize the sine and cosine parts of G so that both have mean 0 and unit L^2 norm. Let us consider the marginal distribution of $|\mathbf{J}(x)|^2$ under the following three hypotheses:

- Hypothesis 0: $\mathbf{I}(x)$ is a white noise process, that is, $\mathbf{I}(x) \sim \mathrm{N}(0, \sigma^2)$ independently.
- Hypothesis 1: $\mathbf{I}(x)$ is a plane wave $\mathbf{I}(x) = \cos(x_1 + \phi)$ propagating along the x_1 direction at the unit frequency, which is the frequency of the wave component of G.
- Hypothesis 2: $\mathbf{I}(x)$ is a step-edge function $\mathbf{I}(x) = \delta(x_1 > 0)$, where δ is the indicator function.

Under hypothesis 0, the real and imaginary parts of $\mathbf{J}(x)$ follow independent normal distribution $\mathrm{N}(0, \sigma^2)$, so $|\mathbf{J}(x)|^2$ follows $\sigma^2 \chi^2_2$, or the exponential distribution with $\mathrm{E}[|\mathbf{J}(x)|^2] = \sigma^2$ and $\mathrm{var}[|\mathbf{J}(x)|^2] = \sigma^4$. This conclusion still holds as long as $\mathbf{I}(x)$ is a "locally white" stationary Gaussian process, in the sense that the spectrum of the process is a constant within the frequency band covered by G. This is because $\mathbf{J}$ does not contain frequency content of $\mathbf{I}$ outside this frequency band. Therefore, for the marginal distribution of $|\mathbf{J}(x)|^2$, $\mathrm{E}[|\mathbf{J}(x)|^2]$ captures the average spectrum over the frequency band of G, and $\mathrm{var}[|\mathbf{J}(x)|^2] = \mathrm{E}^2[|\mathbf{J}(x)|^2]$ if the process is locally white.

Under hypothesis 1, $\mathbf{I}(x) = [e^{i\phi} e^{ix_1} + e^{-i\phi} e^{-ix_1}]/2$; that is, it has two frequency components at $(1, 0)$ and $(-1, 0)$. $\hat{G}(\omega)$ is centered at $(1, 0)$ with standard deviations $1/\sigma_1$ and $1/\sigma_2$. Assume that $1/\sigma_1$ is sufficiently small so that $\hat{G}$ does not cover $(-1, 0)$. Then $\mathbf{J}(x) = \hat{G}(1, 0) e^{i\phi} e^{ix_1}$. Thus $|\mathbf{J}(x)|^2 = \hat{G}(1, 0)^2$, which is a constant, and $\mathrm{var}[|\mathbf{J}(x)|^2] = 0 < \mathrm{E}^2[|\mathbf{J}(x)|^2]$. Therefore, a small marginal variance of $\mathbf{J}(x)$ indicates the possible presence of a periodic component within the frequency band covered by $\hat{G}$.

Under hypothesis 2, it is clear that $|\mathbf{J}(x)|^2 > 0$ only if $|x_1|$ is small compared with σ_1, and $|\mathbf{J}(x)|^2$ achieves its maximum at the edge point $x_1 = 0$. If $|x_1/\sigma_1|$ is large, so that the corresponding Gabor wavelet is away from the edge, then $\mathbf{J}(x) = 0$. The marginal histogram of $|\mathbf{J}(x)|^2$ has a high probability to be 0 and a low probability to be significantly greater than 0. For such a distribution, $\mathrm{var}[|\mathbf{J}(x)|^2] > \mathrm{E}^2[|\mathbf{J}(x)|^2]$ for a random point x.

We may consider hypothesis 0 a null hypothesis of locally white Gaussian process. Hypothesis 1 is an alternative periodic hypothesis, and hypothesis 2 is an alternative edge hypothesis. The two alternative hypotheses are extremes of two directions of departure from the null hypothesis. Dunn and Higgins (1995) derived a class of marginal distributions for image patterns along the direction toward hypothesis 1, and Srivastava, Grenander, and Liu (2002) derived a class of marginal distributions for image patterns along the direction toward hypothesis 2. Biological vision recognizes patterns along both directions. There are V1 cells that are sensitive to edge and bar patterns (Hubel and Wiesel 1962). There is also a small portion of V1 cells that respond to periodic grating patterns but not to edge patterns (Petkov and Kruizinga 1997). Both types of cells can be modeled by combining Gabor wavelets.

For a set of $\{G_{s,\theta}\}$ that paves the frequency domain (as illustrated in Fig. 5), let $\mathbf{J}_{s,\theta} = \mathbf{I} * G_{s,\theta}$. Then the set of marginal averages $\{\mathrm{E}[|\mathbf{J}_{s,\theta}(x)|^2]\}$ captures the spectrum of $\mathbf{I}$ over the frequency bands covered by these $\{G_{s,\theta}\}$. If we constrain the set of $\{\mathrm{E}[|\mathbf{J}_{s,\theta}(x)|^2]\}$, then the corresponding image ensemble is approximately a Gaussian process with a smooth spectrum. If we further constrain $\{\mathrm{var}[|\mathbf{J}_{s,\theta}(x)|^2]\}$, then we add new information about the two directions of departures from the smooth-spectrum Gaussian process. Spectrum and non-Gaussian statistics of natural images also have been studied by Ruderman and Bialek (1994), Simoncelli and Olshausen (2001), and Srivastava, Lee, Simoncelli, and Zhu (2003). Portilla and Simoncelli (2002) described the use of joint statistics for texture modeling.

For an image patch, the issue of whether to summarize it by marginal histograms or to represent it by sparse coding arises. Figure 9 displays an example of this situation. Here Figure 9(a) is the observed 256×256 image, Figure 9(b) is a random image from the ensemble constrained by the marginal histograms of Gabor wavelets, and Figure 9(c) is the image reconstructed by the sparse coding model using 1,265 wavelet elements. Although Figure 9(b) captures some edge information, the marginal histograms are implicit in the sense that by pooling the marginal histograms, we discard all of the position information. The sparse coding model captures the local edges at different positions explicitly and preserves the local shapes. It would be interesting to study the transition between the implicit marginal

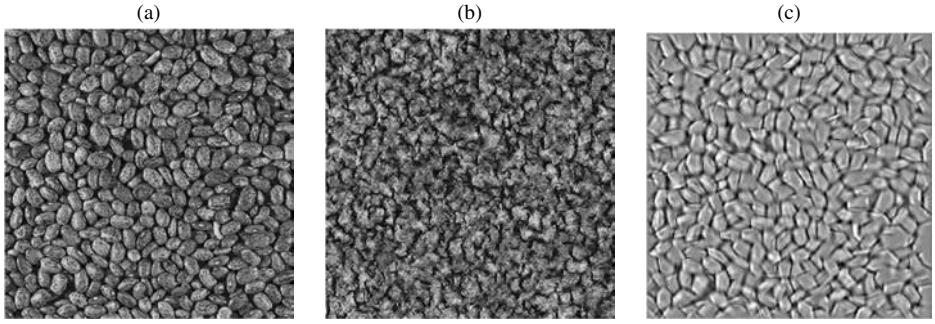

Figure 9. (a) The observed 256 × 256 image. (b) A random sample from the image ensemble constrained by the marginal histograms of Gabor wavelets. (c) Reconstructed image by the sparse coding model using 1,265 wavelet elements.

histograms and the explicit sparse coding. Guo et al. (2003a,b, 2007) have attempted to explain this issue.

3.3 Meaningful Alignment Principle: Beyond Marginal Properties

The sparsity and marginal histograms studied in the previous two sections are marginal properties of the wavelet coefficients. They do not capture the joint properties of wavelet coefficients. The meaningful alignment principle studied by Moisan et al. (2000) is intended to capture these joint properties.

One particular type of joint property studied by Moisan et al. (2000) is the alignment of local orientations on line segments in natural images. For a pixel x, its orientation $\theta(x, s)$ at scale s is defined in Section 2.1. The alignment event is defined in terms of the coincidence of the orientations of a set of pixels. Specifically, consider any two pixels a and b, and let $\theta(a, b)$ be the orientation of the line segment that connects a and b. We can sample a number of equally spaced pixels $a = x_1, x_2, \ldots, x_l = b$ on this line segment (a, b) and compute their orientations $\theta_1, \theta_2, \ldots, \theta_l$. Here θ_i is said to be aligned with $\theta(a, b)$ if $|\theta_i - \theta(a, b)| \le \alpha$, where α is a prespecified threshold. We then compute $S = \sum_{i=1}^{l} \delta(|\theta_i - \theta| \le \alpha)$, that is, the number of sampled pixels whose orientations are aligned with $\theta(a, b)$.

Then we want to determine whether S is sufficiently large for (a, b) to be a meaningful line segment. For this purpose, we may choose a threshold $k(l) = \min\{k : \Pr(S \ge k) \le \epsilon/|D|^2\}$, where $\Pr(A)$ is the probability of an event A under the hypothesis that the image is white noise and $|D|$ is the total number of pixels, so that $|D|^2$ bounds the total number of potential line segments (a, b). If $S \ge k(l)$, then (a, b) is said to be an ϵ-meaningful line segment. In a completely random image, the expected number of falsely declared ϵ-meaningful line segments is less that ϵ. For computational convenience, we can sample the equally spaced pixels $a = x_1, x_2, \ldots, x_l = b$ so that in the random image, $\theta_1, \theta_2, \ldots, \theta_l$ are independent. Then S follows a binomial distribution, and relevant probabilities can be calculated or approximated in closed form.

Generally speaking, an ϵ-meaningful event A is such that in a white noise image, the expected number of such events is less than ϵ. This means that $\Pr(A)$ is small under the complete-randomness hypothesis. The meaningful alignment principle is closely related to the hypothesis testing (or, more precisely, multiple hypothesis testing) problem in statistics.

4. CONNECTIONS AMONG THREE STATISTICAL PRINCIPLES

The three statistical principles reviewed in the previous section can be connected by a common perspective, where the goal is to identify a small number of most powerful constraints to restrict the observed image into an image ensemble of a minimum volume. Also see the recent article by Wu et al. (2007), which unifies the three statistical principles in a theoretical framework combining hypothesis testing and statistical modeling.

4.1 Pooling Projections

First, we consider the sparse coding principle. Given the dictionary of basis elements, this principle seeks to identify a small number of basis elements $\{\mathbf{B}_1, \ldots, \mathbf{B}_K\}$ from the dictionary to code the observed image $\mathbf{I}$ with small error. Geometrically, this means that $\mathbf{I}$ projects most of its squared norm onto the subspace spanned by $\{\mathbf{B}_1, \ldots, \mathbf{B}_K\}$. Thus we can view sparse coding as constraining the projection of $\mathbf{I}$ onto the subspace spanned by $\{\mathbf{B}_1, \ldots, \mathbf{B}_K\}$.

Specifically, for any image $\mathbf{I}$ defined on a finite grid D of $|D|$ pixels, we may vectorize $\mathbf{I}$ to make it a $|D|$-dimensional vector. Similarly, we also can vectorize $\mathbf{B}_1, \ldots, \mathbf{B}_K$ into $|D|$-dimensional vectors. Let $r_k = \langle \mathbf{I}, \mathbf{B}_k \rangle$, and let $R = (r_1, \ldots, r_K)'$ be the K-dimensional vector of projection coefficients. Let $\mathbf{B} = (\mathbf{B}_1, \ldots, \mathbf{B}_K)$ be the $(|D| \times K)$-dimensional matrix whose kth column is $\mathbf{B}_k$. Then $R = \mathbf{B}'\mathbf{I}$. Let $\mathbf{H} = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$ be the projection matrix. Then $\mathbf{HI} = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}R$ is the projection of $\mathbf{I}$ onto the subspace spanned by $\{\mathbf{B}_1, \ldots, \mathbf{B}_K\}$, and $\|\mathbf{HI}\|^2 = R'(\mathbf{B}'\mathbf{B})^{-1}R$.

Let $\bar{\mathbf{B}}$ be an $(|D| \times (|D| - K))$-dimensional matrix whose columns are orthonormal vectors that are also orthogonal to all of the $\mathbf{B}_k$'s. Let $\bar{R} = \bar{\mathbf{B}}'\mathbf{I}$. Then $\|\bar{R}\|^2 = \|\mathbf{I}\|^2 - \|\mathbf{HI}\|^2 = \|\mathbf{I}\|^2 - R'(\mathbf{B}'\mathbf{B})^{-1}R$.

Consider the image ensemble

$$\Omega = \left\{\mathbf{I} : \|\mathbf{I}\|^2 = |D|\sigma^2 = \|\mathbf{I}_{\text{obs}}\|^2, \mathbf{B}'\mathbf{I} = R = \mathbf{B}'\mathbf{I}_{\text{obs}}\right\}. \quad (10)$$

This ensemble is a sphere viewed in the $(|D| - K)$-dimensional space spanned by $\bar{\mathbf{B}}$,

$$\Omega = \{\bar{R} : \|\bar{R}\|^2 = |D|\sigma^2 - R'(\mathbf{B}'\mathbf{B})^{-1}R\}. \quad (11)$$

Let

$$\gamma^2 = \frac{|D|\sigma^2 - R'(\mathbf{B}'\mathbf{B})^{-1}R}{|D| - K}. \quad (12)$$

Then, using the formula for computing the volume of high-dimensional sphere and Stirling's formula, as $|D| \to \infty$,

$$\frac{1}{|D| - K} \log |\Omega| \to \log(\sqrt{2\pi e}\gamma), \tag{13}$$

where $|\Omega|$ is the volume of Ω.

Thus, to find the K basis functions $\mathbf{B}$ to constrain $\mathbf{I}_{\text{obs}}$ within a minimum-volume Ω, we should minimize the radius of the sphere Ω or, equivalently, the least squares error $\|\mathbf{I}_{\text{obs}}\|^2 - \|\mathbf{H}\mathbf{I}_{\text{obs}}\|^2$. This is clearly the goal of the sparse coding model presented in Section 3.1.

Now we can connect the sparse coding model and the texture model. In the sparse coding model, we restrict a small number of projection coefficients, as well as the marginal variance. In the texture model, we restrict the marginal histograms of the projection coefficients pooled over positions. Both models seek to restrict the observed image into an image ensemble of small volume. The two models differ in their respective constraints. In the sparse coding model, the constraints are on individual projection coefficients of Gabor wavelets of particular positions, scales, and orientations. In the texture models, the constraints are on the marginal histograms of projection coefficients of particular scales and orientations but pooled over positions; that is, the position information is discarded.

Just as the constrained texture ensemble can be approximated by an unconstrained Markov random field, the constrained sphere also can be approximated by an unconstrained statistical model, which is nothing but a white noise model $\bar{R} \sim \text{N}(0, \gamma^2 I_{|D|-K})$, where $I_{|D|-K}$ is the $(|D| - K)$-dimensional identity matrix. This is because under the white noise model, $\|\bar{R}\|^2/(|D| - K) \to \gamma^2$ as $|D| \to \infty$ according to the law of large numbers, and the white noise model assigns equal probabilities to all of the $\bar{R}$'s with fixed $\|\bar{R}\|^2$. The log-volume of the sphere Ω in (13) is the same as the negative log-likelihood of the white noise model.

This leads to a statistical model for image patterns in the object regime, where we can select the basis elements $\{\mathbf{B}_1, \ldots, \mathbf{B}_K\}$ and estimate the joint distribution $f(r_1, \ldots, r_K)$ of the projection coefficients, by learning from a sample of training image patches of the same class of objects, for example, a sample of 40×40 image patches of cars, where the scales and locations of the objects in these image patches are roughly fixed and the objects are roughly aligned.

Specifically, consider the change of variable

$$\begin{pmatrix} R \\ \bar{R} \end{pmatrix} = (\mathbf{B}, \bar{\mathbf{B}})'\mathbf{I}. \tag{14}$$

We can model $(R, \bar{R})$ by $f(R, \bar{R}) = f(R)f(\bar{R}|R)$, where $f(\bar{R}|R) \sim \text{N}(0, \gamma^2 I_{|D|-K})$, with γ^2 computed according to (12). Then the probability distribution of $\mathbf{I}$ is $p(\mathbf{I}) = f(R, \bar{R}) \det(\mathbf{B}'\mathbf{B})^{1/2}$, where $\det(\mathbf{B}'\mathbf{B})^{1/2}$ is the Jocobian of the change of variable (14). The log-likelihood for an image patch $\mathbf{I}$ is

$$\begin{aligned} \log p(\mathbf{I}) &= \log\left[f(R, \bar{R}) \det(\mathbf{B}'\mathbf{B})^{1/2}\right] \\ &= \log f(R) - \frac{1}{2}(|D| - K) \log(2\pi e\gamma^2) \\ &\quad + \frac{1}{2} \log \det(\mathbf{B}'\mathbf{B}). \end{aligned} \tag{15}$$

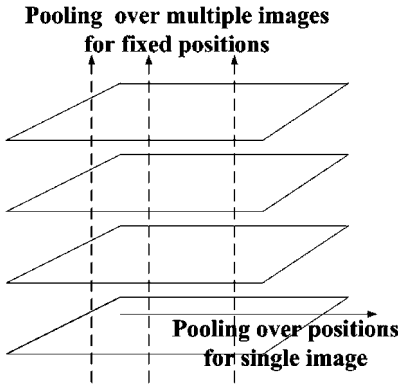

Figure 10. Difference between the object model and the texture model. In the object model, projection coefficients of certain basis elements (subject to local shifting) are pooled over multiple training images. In the texture model, projection coefficients of basis elements are pooled over positions in a single image.

The goal is to select $\mathbf{B} = (\mathbf{B}_1, \ldots, \mathbf{B}_K)$ from the dictionary and estimate $f(R)$ from the training images by maximizing the log-likelihood (15) over independent training images. In other words, for each class of objects, we want to model it as a composition of a small number of basis elements. We may allow these basis elements to change their positions, scales, and orientations slightly to account for local deformations. Tolerance of such deformations has been observed in the complex cells in V1 (Lampl et al. 2004; Riesenhuber and Poggio 1999).

The foregoing model is different from the more commonly assumed form in sparse coding, $\mathbf{I} = \mathbf{B}C + \epsilon$, $C \sim p(C)$, $\epsilon \sim \text{N}(0, I_{|D|})$ (Lewicki and Olshausen 1999; Pece 2002; Olshausen and Millman 2000; Zhu et al. 2005), which models the reconstruction coefficients, C, instead of the projection coefficients, R. Modeling C is technically difficult, because C is a latent variable that needs to be inferred, and $p(\mathbf{I})$ must be obtained by integrating out C. In contrast, R can be obtained deterministically through $R = \mathbf{B}'\mathbf{I}$, and the distribution of $p(\mathbf{I})$ can be obtained explicitly by the change of variable. The model bears some similarity to independent component analysis (Bell and Sejnowski 1997) and projection pursuit density estimation (Friedman 1987), but has a different mathematical form and has the capability of modeling the composition and deformation of the basis elements.

We conclude this section by comparing the foregoing object model with the texture model. In the object model, we pool the distribution of projections coefficients $f(R)$ for a number of basis elements at certain positions (subject to local shifting) over multiple training image patches. In the texture model, we pool the marginal distributions of projection coefficients over the positions of a single texture image. The difference is illustrated in Figure 10.

4.2 Identifying Alignments

In this section we consider the meaningful alignment principle, which also can be understood as searching for constraints that restrict the observed image into an image ensemble of small volume.

Let Ω_0 be the set of images defined on a grid D with fixed marginal variance $\Omega_0 = \{\mathbf{I} : \|\mathbf{I}\|^2 = |D|\sigma^2\}$. Let $A_1, \ldots, A_K$ be

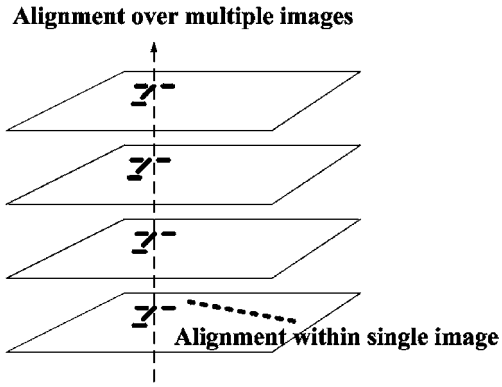

Figure 11. In the object regime, the basis elements (illustrated as small bars) around the vertical arrow are aligned (subject to local shifting) across multiple images. In the geometry regime, those basis elements that are away from the vertical arrow are aligned within a single image.

K statements about the alignment events in the observed image $\mathbf{I}_{\text{obs}} \in \Omega_0$. Mathematically, each A_k is a subset of Ω_0 that contains all of the images satisfying the statement A_k. To find informative $A_1, \ldots, A_K$, we want the volume of $\Omega = A_1 \cap \cdots \cap A_K$ to be as small as possible. Under the uniform distribution over Ω_0, the probability is $p(\Omega) = |\Omega|/|\Omega_0|$. For a large image lattice, the uniform distribution over Ω is equivalent to a Gaussian white noise model; thus $p(\Omega)$ is also the probability of Ω under the white noise hypothesis. Because $|\Omega| = p(\Omega)|\Omega_0|$, searching for alignment events that are very rare under the white noise hypothesis is equivalent to searching for alignment events to constrain the observed image into an image ensemble of small volume.

The alignment patterns can be found in the geometry regime, where the orientations of locally best-tuned Gabor wavelets are highly aligned. The probability for such alignment is small under the white noise hypothesis.

The alignments also can be found in the object regime, except that alignments of the locally best-tuned Gabor wavelets are observed across multiple training image patches of the same class of objects instead of in alignment within a single image. Figure 11 illustrates the difference. The basis elements (shown as small bars) around the vertical arrow can be aligned across multiple images, but within a single image, such as the one at the bottom, those basis elements near the vertical arrow do not have significant alignment, and thus the pattern cannot be detected by meaningful alignment without the knowledge learned from multiple training images. But the alignment pattern of the basis elements that are away from the vertical arrow can be detected from this image without any specifically learned knowledge.

Both the geometry and object regimes can be modeled as compositions of Gabor wavelets. Patterns in the geometry regime are compositions of highly aligned Gabor wavelets, and patterns in the object regime are compositions of Gabor wavelets that are not highly aligned within a single image. In this sense, edge and region patterns are just simple generic object patterns. In comparison, images in the texture regime exhibit no meaningful alignments or a priori known compositions and are summarized by pooling the marginal statistics that discard the position information.

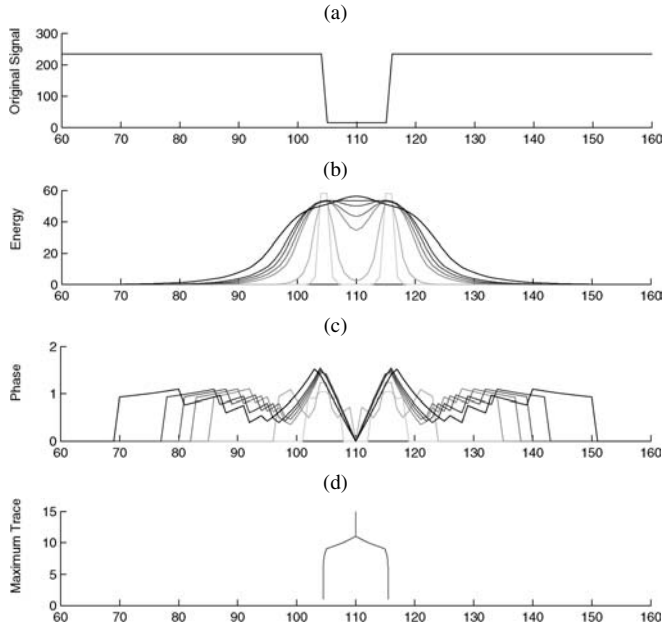

Figure 12. Alignment over scale. (a) A horizontal slice of an image of a vertical bar, $\mathbf{I}(x)$ versus x_1. (b) Energy $A(x, s)$ versus x_1. (c) Phase $\phi(x, s)$ versus x_1. Each curve corresponds to a scale s. (d) Positions of edge-bar points over the scale. The two edge points merge into a bar point at a certain scale.

Another type of alignment in the geometry regime also can be very important: the alignment over scales (Witkin 1983; Lindberg 1993) or frequencies (Morrone, Ross, Burr, and Owens 1986). Such an alignment was used by Kovesi (1999) to detect edges and corners. Figure 12 illustrates the basic idea. Figure 12(a) shows a horizontal slice of an image $\mathbf{I}(x)$ of a vertical bar. The horizontal axis is x_1, and the vertical axis is $\mathbf{I}(x)$. Figures 12(b) and 12(c) display the energy $A(x, s)$ and phase $\phi(x, s)$ on this slice. The definitions of $A(x, s)$ and $\phi(x, s)$ were provided in Section 2.1. In Figures 12(b) and 12(c), each curve corresponds to $A(x, s)$ or $\phi(x, s)$ for a fixed s. It is evident that an edge-bar point $\tilde{x}$ [i.e., a local maximum in energy $A(x, s)$] can exist over a range of scales. Within this range, the phase $\phi(\tilde{x}, s)$ and orientation $\theta(\tilde{x}, s)$ remain constant over different scales s. For an edge point, the energy $A(\tilde{x}, s)$ also remains constant (subject to discretization error) if the Gabor sine and cosine wavelets are normalized to have unit L^1 norm. If the Gabor wavelets are normalized to have unit L^2 norm, then $A(\tilde{x}, s)$ increases over s within the range of alignment. Figure 12(d) traces the positions of the two edge points over scale s. In this plot, the horizontal axis is x_1 and the vertical axis is s. We can see that at a certain scale, the two edge points merge into a bar point. Beyond this scale, the orientation and the phase continue to be aligned at this bar point.

The alignment over scale can be important for modeling the coarse-to-fine generative model shown in Figure 2. For example, an edge pattern at a certain scale can be expanded into larger edge patterns at a finer scale. A bar pattern can be expanded into two edge patterns at a finer scale. More generally, an object pattern may be composed of several Gabor wavelets, each of which may be expanded into an edge or bar pattern. This point of view can be useful for modeling the relationship between schemes 1 and 2 discussed in Section 1.3.

Multiscale hierarchical models of wavelet coefficients have been developed by De Bonet and Viola (1997) for texture synthesis and recognition and by Buccigrossi and Simocelli (1999) for image compression and denoising. What we are proposing here is to add the labels of the recognized patterns in the coarse-to-fine generative model. Needless to say, we remain far away from such a model, and much theoretical and experimental work needs to be done. A preliminary attempt toward such a model has been made by Wang et al. (2005).

5. DISCUSSION

5.1 Classification Rule or Generative Model?

There are two major schools of thoughts on statistical learning in computer vision. One school emphasizes learning generative models (Cootes, Edwards, and Taylor 2001; Doretto, Chiuso, Wu, and Soatto 2003; Isard and Blake 1998; Wang and Zhu 2004) and deriving the likelihood or the posterior distribution Pr(labels|image patches) from the learned models. The other school emphasizes learning Pr(labels|image patches) or, more simply, the classification rules: Label $= f$(image patch) directly (Amit and Geman 1997; Viola and Jones 2004), without learning the generative models. Attempts have been made to combine these two schools of methods (Tu and Zhu 2002; Tu, Chen, Yuille, and Zhu 2005).

Comparing these two schools shows that the second school is more direct than the first and is very powerful for classification tasks in which the variable "label" is a category that takes values in a finite or binary set. But for general vision tasks, the variable "labels" has a much more complex structure than a category; for example, it can have a hierarchical structure with compositional and spatial organizations, as illustrated in Figure 2(b). For such a complex structure, learning Pr(labels|image patches) may not be practical; learning the coarse-to-fine generative model may be simpler and more natural.

5.2 Low-Level and Midlevel Vision?

Vision tasks are roughly classified into low-level, mid-level, and high-level tasks. Low-level tasks involve detecting local image structures (e.g., edges, corners, junctions) or detecting what Julesz called "textons" (Julesz 1981), or what Marr called "symbols" or "tokens" (Marr 1982). Midlevel tasks involve extracting curves and contours from the image or segmenting the image into regions of coherent intensity patterns. High-level tasks involve recognizing objects and their shapes and poses.

Detecting local image structures in low-level tasks can be difficult, because natural images are full of ambiguities if we look only at a local image patch at a fixed resolution. Extracting contours or regions from images also can be difficult, even with the help of regularity terms or generic prior distributions (Kass, Witkin, and Terzopoulos 1987; Mumford and Shah 1989) to enforce smoothness and continuity. The low-level and midlevel tasks are often considered the foundation for the high-level tasks of object recognition, but at the same time, the ambiguities in low-level and midlevel tasks can be resolved only after the specific objects are recognized.

The Bayesian framework discussed in Section 1.3 can be an elegant framework for resolving the mutual dependencies between the tasks at different levels. In this framework, object patterns and edge patterns are labeled simultaneously at different layers. The edge patterns are not treated as being more fundamental than the object patterns. Although an object pattern can be composed of edge patterns at a fine resolution, at a coarse resolution an object pattern occupies an image patch that is no larger than the image patch of an edge pattern at a fine resolution. Prior knowledge of how the object patterns are composed of the edge patterns aids the labeling of both types of patterns. Intuitively, this prior knowledge serves as a constraint to help eliminate the ambiguities in the local image patches at different layers.

ACKNOWLEDGEMENT

The work presented in this article was supported by NSF DMS 0707055, NSF IIS 0413214, ONR C2Fusse.

[Received January 2006. Revised September 2006.]

REFERENCES

Adelson, E. H., and Bergen, J. R. (1985), "Spatiotemporal Energy Models for the Perception of Motion," *Journal Optical Society of America*, Ser. A, 2, 284–299.

Amit, Y., and Geman, D. (1997), "Shape Quantization and Recognition With Randomized Trees," *Neural Computation*, 9, 1545–1588.

Bell, A., and Sejnowski, T. J. (1997), "The 'Independent Components' of Natural Scenes Are Edge Filters," *Vision Research*, 37, 3327–3338.

Besag, J. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 36, 192–236.

Buccigrossi, R. W., and Simoncelli, E. P. (1999), "Image Compression via Joint Statistical Characterization in the Wavelet Domain," *IEEE Transactions on Image Processing*, 8, 1688–1701.

Burt, P., and Adelson, E. H. (1983), "The Laplacian Pyramid as a Compact Image Code," *IEEE Transactions on Communication*, 31, 532–540.

Candes, E. J., and Donoho, D. L. (1999), "Curvelets: A Surprisingly Effective Nonadaptive Representation for Objects With Edges," in *Curves and Surfaces*, ed. L. L. Schumakeretal, Nashville, TN: Vanderbilt University Press.

Canny, J. (1986), "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8, 679–698.

Chen, S., Donoho, D., and Saunders, M. A. (1999), "Atomic Decomposition by Basis Pursuit," *SIAM Journal on Scientific Computing*, 20, 33–61.

Cootes, T. F., Edwards, G. J., and Taylor, C. J. (2001), "Active Appearance Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23, 681–685.

Daugman, J. (1985), "Uncertainty Relation for Resolution in Space, Spatial Frequency, and Orientation Optimized by Two-Dimensional Visual Cortical Filters," *Journal of the Optical Society of America*, 2, 1160–1169.

De Bonet, J. S., and Viola, P. A. (1997), "A Non-Parametric Multi-Scale Statistical Model for Natural Images," *Advances in Neural Information Processing Systems*.

Della Pietra, S., Della Pietra, V., and Lafferty, J. (1997), "Inducing Features of Random Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, 380–393.

Doretto, G., Chiuso, A., Wu, Y. N., and Soatto, S. (2003), "Dynamic Textures," *International Journal of Computer Vision*, 51, 91–109.

Dunn, D., and Higgins, W. E. (1995), "Optimal Gabor Filters for Texture Segmentation," *IEEE Transactions on Image Processing*, 4, 947–964.

Friedman, J. H. (1987), "Exploratory Projection Pursuit," *Journal of the American Statistical Association*, 82, 249.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Geman, S., and Graffigne, C. (1987), "Markov Random Field Image Models and Their Applications to Computer Vision," *Proceedings of the International Congress of Mathematicians*, 1, 1496–1517.

Geman, S., Potter, D. F., and Chi, Z. (2002), "Composition System," *Quarterly of Applied Mathematics*, 60, 707–736.

George, E. I., and McCulloch, R. E. (1997), "Approaches to Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373.

Grenander, U. (1993), *General Pattern Theory*, Oxford, U.K.: Oxford University Press.

Grenander, U., and Miller, M. I. (1994), "Representation of Knowledge in Complex Systems," *Journal of the Royal Statistics Society*, Ser. B, 56, 549–603.

Guo, C., Zhu, S. C., and Wu, Y. N. (2003a), "Visual Learning by Integrating Descriptive and Generative Models," *International Journal of Computer Vision*, 53, 5–29.

——— (2003b), "A Mathematical Theory of Primal Sketch and Sketchability," in *Proceedings of the International Conference of Computer Vision*, pp. 1228–1235.

——— (2007), "Primal Sketch: Integrating Structure and Texture," *Computer Vision and Image Understanding*, 106, 5–19.

Heeger, D. J., and Bergen, J. R. (1995), "Pyramid-Based Texture Analysis/Synthesis," *Computer Graphics Proceedings*, 229–238.

Hubel, D., and Wiesel, T. (1962), "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex," *Journal of Physiology*, 160, 106–154.

Huo, X., and Chen, J. (2005), "JBEAM: Multiscale Curve Coding via Beamlets," *IEEE Transactions on Image Processing*, 14, 1665–1677.

Isard, M., and Blake, A. (1998), "Condensation-Conditional Density Propagation for Visual Tracking," *International Journal of Computer Vision*, 29, 5–28.

Julesz, B. (1981), "Textons, the Elements of Texture Perception and Their Interactions," *Nature*, 290, 91–97.

Kass, M., Witkin, A., and Terzopoulos, D. (1987), "Snakes: Active Contour Models," *International Journal of Computer Vision*, 1, 321–331.

Kovesi, P. (1999), "Image Features From Phase Congruency," *Videre: Journal of Computer Vision Research*, 1.

Lampl, I., Ferster, D., Poggio, T., and Riesenhuber, M. (2004), "Intracellular Measurements of Spatial Integration and the MAX Operation in Complex Cells of the Cat Primary Visual Cortex," *Journal of Neurophysiology*, 92, 2704–2713.

Lee, T. S. (1996), "Image Representation Using 2D Gabor Wavelets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10, 959–971.

Lewicki, M. S., and Olshausen, B. A. (1999), "Probabilistic Framework for the Adaptation and Comparison of Image Codes," *Journal of the Optical Society of America*, 16, 1587–1601.

Lindberg, T. (1993), "Effective Scale: A Natural Unit for Measuring Scale-Space Lifetime," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15, 1068–1074.

——— (1994), *Scale-Space Theory in Computer Vision*, Dordrecht: Kluwer Academic Publishers.

Mallat, S. (1989), "A Theory of Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11, 674–693.

Mallat, S., and Zhang, Z. (1993), "Matching Pursuit in a Time-Frequency Dictionary," *IEEE Transactions on Signal Processing*, 41, 3397–3415.

Mallat, S., and Zhong, Z. (1992), "Characterization of Signals From Multiscale Edges," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14, 710–732.

Marr, D. (1982), *Vision*, Gordonsville, VA: W. H. Freeman.

Moisan, L., Desolneux, A., and Morel, J.-M. (2000), "Meaningful Alignments," *International Journal of Computer Vision*, 40, 7–23.

Morrone, M. C., Ross, J., Burr, D. C., and Owens, R. A. (1986), "Mach Bands Are Phase Dependent," *Nature*, 324, 250–253.

Mumford, D. B. (1994), "Pattern Theory: A Unifying Perspective," in *Proceedings of the First European Congress of Mathematics*, Boston: Birkhauser.

Mumford, D., and Shah, J. (1989), "Optimal Approximations by Piecewise Smooth Functions and Associated Variational Problems," *Communications on Pure and Applied Mathematics*, 42, 577–685.

Olshausen, B. A., and Field, D. J. (1996), "Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images," *Nature*, 381, 607–609.

Olshausen, B. A., and Millman, K. J. (2000), "Learning Sparse Codes With a Mixture-of-Gaussians Prior," *Advances in Neural Information Processing Systems*, 12, 841–847.

Pece, A. (2002), "The Problem of Sparse Image Coding," *Journal of Mathematical Imaging and Vision*, 17, 89–108.

Perona, P., and Malik, J. (1990), "Detecting and Localizing Composite Edges in Images," in *Proceedings of the International Conference of Computer Vision*, pp. 52–57.

Petkov, N., and Kruizinga, P. (1997), "Computational Models of Visual Neurons Specialised in the Detection of Periodic and Aperiodic Oriented Visual Stimuli: Bar and Grating Cells," *Biological Cybernetics*, 76, 83–96.

Portilla, J., and Simoncelli, E. P. (2000), "A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients," *International Journal of Computer Vision*, 40, 49–71.

Riesenhuber, M., and Poggio, T. (1999), "Hierarchical Models of Object Recognition in Cortex," *Nature Neuroscience*, 2, 1019–1025.

Ruderman, D. L., and Bialek, W. (1994), "Statistics of Natural Images: Scaling in the Woods," *Physics Review Letters*, 73, 814–817.

Simoncelli, E. P., Freeman, W. T., Adelson, E. H., and Heeger, D. J. (1992), "Shiftable Multiscale Transforms," *IEEE Transactions on Information Theory*, 38, 587–607.

Simoncelli, E. P., and Olshausen, B. A. (2001), "Natural Image Statistics and Neural Representation," *Annual Review of Neuroscience*, 24, 1193–1216.

Srivastava, A., Grenander, U., and Liu, X. (2002), "Universal Analytical Forms for Modeling Image Probabilities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 1200–1214.

Srivastava, A., Lee, A., Simoncelli, E., and Zhu, S. (2003), "On Advances in Statistical Modeling of Natural Images," *Journal of Mathematical Imaging and Vision*, 18, 17–33.

Tu, Z. W., and Zhu, S. C. (2002), "Image Segmentation by Data Driven Markov chain Monte Carlo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 657–673.

Tu, Z. W., Chen, X. R., Yuille, A. L., and Zhu, S. C. (2005), "Image Parsing: Unifying Segmentation, Detection and Recognition," *International Journal of Computer Vision*, 63, 113–140.

Viola, P. A., and Jones, M. J. (2004), "Robust Real-Time Face Detection," *International Journal of Computer Vision*, 57, 137–154.

Wang, Y. Z., and Zhu, S. C. (2004), "Analysis and Synthesis of Textured Motion: Particles and Waves," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26, 1348–1363.

Wang, Y. Z., Bahrami, S., and Zhu, S. C. (2005), "Perceptual Scale Space and Its Applications," in *Proceedings of the International Conference of Computer Vision*, pp. 58–65.

Wang, Z., and Jenkin, M. (1992), "Using Complex Gabor Filters to Detect and Localize Edges and Bars," *Advances in Machine Vision*, 32, 151–170.

Witkin, A. (1983), "Scale-Space Filtering," in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1019–1021.

Wu, Y. N., Zhu, S. C., and Guo, C. (2002), "Statistical Modeling of Texture Sketch," in *Proceedings of the European Conference of Computer Vision*, pp. 240–254.

——— (2007), "From Information Scaling to Regimes of Statistical Models," *Quarterly of Applied Mathematics*, to appear.

Wu, Y. N., Zhu, S. C., and Liu, X. W. (2000), "Equivalence of Julesz Ensemble and FRAME Models," *International Journal of Computer Vision*, 38, 245–261.

Young, R. A. (1987), "The Gaussian Derivative Model for Spatial Vision: I. Retinal Mechanism," *Spatial Vision*, 2, 273–293.

Zhu, S. C., and Mumford, D. (2007), "Quest for a Stochastic Grammar of Images," *Foundations and Trends in Computer Graphics and Vision*, to appear.

Zhu, S. C., Guo, C. E., Wang, Y. Z., and Xu, Z. J. (2005), "What Are Textons?" *International Journal of Computer Vision*, 62, 121–143.

Zhu, S. C., Liu, X., and Wu, Y. N. (2000), "Exploring Texture Ensembles by Efficient Markov Chain Monte Carlo: Towards a 'Trichromacy' Theory of Texture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 554–569.

Zhu, S. C., Wu, Y. N., and Mumford, D. (1997), "Minimax Entropy Principle and Its Applications in Texture Modeling," *Neural Computation*, 9, 1627–1660.