New Directions in Statistical Signal Processing: From Systems to Brain

edited by
Simon Haykin
Jose C. Principe
Terrence J. Sejnowski
John McWhirter

The MIT Press
Cambridge, Massachusetts
London, England

1 Empirical Statistics and Stochastic Models for Visual Signals

David Mumford
Division of Applied Mathematics
Brown University
Providence, RI 02912
David_Mumford@brown.edu

1.1 Introduction

The formulation of the vision problem as a problem in Bayesian inference (Mumford, 1996, 2002; Forsyth and Ponce, 2002) is, by now, well-known and widely accepted in the computer vision community. In fact, the insight that the problem of reconstructing 3D information from a 2D image is ill-posed and needs inference can be traced back to the Arab scientist Ibn Al-Haytham (known to Europe as Alhazan) around the year 1000 (Haytham, c.1000). Inheriting a complete hodge-podge of conflicting theories from the Greeks[1], Al-Haytham for the first time demonstrated that light rays originated only in external physical sources, moved in straight lines, reflecting and refracting, until they hit the eye; and that the resulting signal needed to be and was actively decoded in the brain using a largely unconscious and very rapid inference process based on past visual experiences. In the modern era, the inferences underlying visual perception have been studied by many people, notably H. Helmholtz, E. Brunswik (Brunswik, 1956) and J. J. Gibson.

In mathematical terms, the Bayesian formulation is as follows: let I be the observed image, a 2D array of pixels (black and white or colored or possibly a stereoscopic pair of such images). Here we are assuming a static image[2]. Let w stand for variables which describe the external scene generating the image. Such variables should include depth and surface orientation information (Marr's $2\frac{1}{2}$ D sketch), location and boundaries of the principle objects in view, their surface

1. The chief mistake of the Greeks was their persistent belief that the eye must *emit* some sort of ray in order to do something equivalent to touching the visible surfaces.
2. This is certainly biologically unrealistic. Life requires rapid analysis of changing scenes. But this article, like much of vision research, simplifies its analysis by ignoring time

albedos, location of light sources and labelling of object categories and possibly object identities. Then two stochastic models, learned from past experience, are required: a *prior model* $p(w)$ specifying what scenes are likely in the world we live in and an *imaging model* $p(I|w)$ specifying what images should look like, given the scene. Then by Bayes's rule:

Bayes' rule

$$p(w|I) = \frac{p(I|w)p(w)}{p(I)} \propto p(I|w)p(w).$$

Bayesian inference consists is fixing the observed value of I and inferring that w equals that value which maximizes $p(w|I)$ or equivalently maximizes $p(I|w)p(w)$. This is a fine general framework, but to implement or even test it requires (a) a theory of stochastic models of a very comprehensive sort which can express all the complex but variable patterns which the variables w and I obey, (b) a method of learning from experience the many parameters which such theories always contain and (c) a method of computing the maximum of $p(w|I)$.

This chapter will be concerned only with problem (a). Many critiques of vision algorithms have failed to allow for the fact that these are three separate problems: if (b) or (c) are badly implemented, the resulting problems do not imply that the theory in (a) is bad. For example, very slow algorithms of type (c) may reasonably be used to test ideas of type (a). Progress in understanding vision does not require all these problems to be solved at once. Therefore, it seems to me legitimate to isolate problems of type (a).

In the rest of this chapter, I will review some of the progress in constructing these models. Specifically, I will consider, in Section 1.2, models of the empirical probability distribution $p(I)$ inferred from large databases of natural images. Then, in Section 1.3, I will consider the problem of the first step in 'intermediate' vision: inferring the regions which should be grouped together as single objects or structures, problems which include segmentation and gestalt grouping, the basic grammar of image analysis. Finally in Section 1.4, I look at the problem of priors on 2D shapes and the related problem of what it means for two shapes to be 'similar'. Obviously, all of these are huge topics and I cannot hope to give a comprehensive view of work on any of them. Instead, I shall give my own views of some of the important issues and open problems and outline the work that I know well. As this inevitably emphasizes the work of my associates, I must beg indulgence from those whose work I have omitted.

1.2 Statistics of the image alone

The most direct approach to studying images is to ask whether we can find good models for images without any hidden variables. This means first creating large databases of images I which we believe are reasonably random samples of all possible images of the world we live in. Then we can study this database with all the tools of statistics, computing the responses of various linear and nonlinear filters

and looking at the individual and joint histograms of their values. Nonlinear should be taken in the broadest sense, including order statistics or topological analyses. We then seek to isolate the most important properties these statistics have and to create the simplest stochastic models $p(I)$ which duplicate or approximate these statistics. The models can be further tested by sampling from them and seeing if the resulting artificial images have the same 'look and feel' as natural images. Or if not, what are the simplest properties of natural images that have we failed to capture. Another recent survey of such models is referred to (Lee et al., 2003b).

1.2.1 High kurtosis as the universal clue to discrete structure

The first really striking thing about filter responses is that they always have large kurtosis. It is strange that electrical engineers designing TV sets in the 1950's do not seem to have pointed this out and this fact first appeared in the work of David Field (Field, 1987). By kurtosis, we mean the normalized fourth moment. If x is a random real number, its kurtosis is

kurtosis

$$\kappa(x) = E((x - \bar{x})^4)/E((x - \bar{x})^2)^2.$$

Every normal variable has kurtosis 3; a variable which has no tails (e.g. uniformly distributed on an interval) or is bimodal and small at its mean tends to have kurtosis less than 3; a variable with heavy tails or large peak at its mean tends to have kurtosis larger than 3. The empirical result which is observed for images is that for any linear filter F with zero mean, the values $x = (F * I)(i, j)$ of the filtered image follow a distribution with kurtosis larger than 3. The simplest case of this is the difference of adjacent pixel values, the discrete derivative of the image I. But it has been found (Huang, 2000) to hold even for *random* mean 0 filters supported in an 8×8 window.

This high kurtosis is shown in Figure 1.1, from the thesis of J. Huang (Huang, 2000). This data was extracted from a large database of high resolution, fully calibrated images of cities and country taken in Holland by van Hateren, (van Hateren, 1998). It is important, when studying tails of distributions, to plot the *logarithm* of the probability or frequency, as in this figure, not the raw probability. If you plot probabilities, all tails look alike. But if you plot their logarithms, then a normal distribution becomes a downwards facing parabola (since $\log(e^{-x^2}) = -x^2$), so heavy tails appear clearly as curves which do not point down so fast.

stationary Markov process

It is a well-known fact from probability theory that if X_t is a stationary Markov stochastic process, then the kurtosis of $X_t - X_s$ being greater than 3 means that the process X_t has discrete jumps. In the case of vision, we have samples from an image $I(s, t)$ depending on two variables rather than one and the zero-mean filter is a generalization of the difference $X_t - X_s$. Other signals generated by the world, such as sound or prices, are functions of one variable, time. A nice elementary statement of the link between kurtosis and jumps is given by the following result taken from (Mumford and Desolneux):

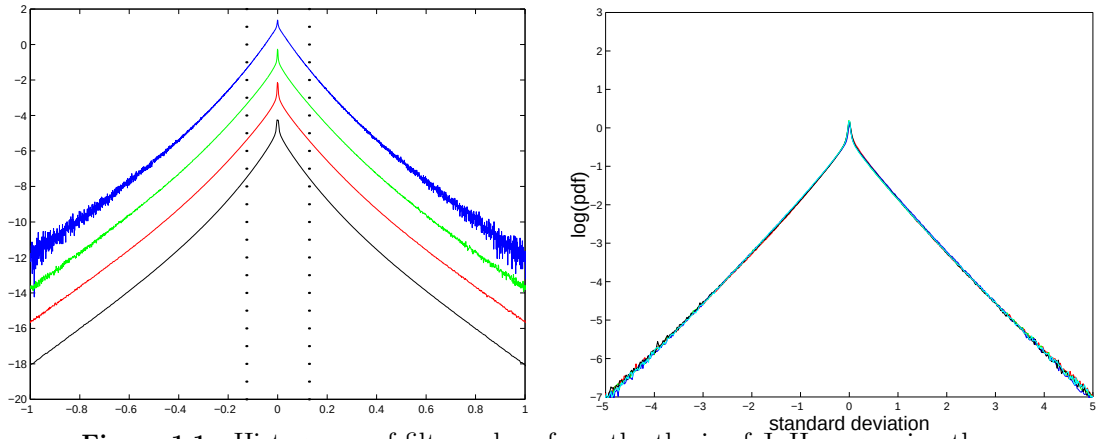

Figure 1.1 Histograms of filter values from the thesis of J. Huang, using the van Hateren database. On the left, the filter is the difference of (a) horizontally adjacent pixels, and (b) of adjacent 2×2, (c) 4×4 and (d) 8×8 blocks; on the right, several *random* mean 0 filters with 8×8 pixel support have been used. The kurtosis of all these filter responses is between 7 and 15. Note that the vertical axis is **log** of the frequency, not frequency. The histograms on the left are displaced vertically for legibility and the dotted lines indicate one standard deviation.

Theorem 1.1

Let x be any real random variable which we normalize to have mean 0 and standard deviation 1. Then there is a constant $c > 0$ depending only on x such that if, for some n, x is the sum:

$$x = y_1 + y_2 + \cdots + y_n$$

where the y_i are independent and identically distributed, then

$$\text{Prob}\left(\max_i |y_i| \geq \sqrt{(\kappa(x) - 3)/2}\right) \geq c.$$

A striking application of this is to the stock market. Let x be the log price change of the opening and closing price of some stock. If we assume price changes are Markov, as many have, and use the experimental fact that price changes have kurtosis greater than 3, then it implies that stock prices cannot be modeled as a continuous function of time. In fact, in my own fit of some stock market data, I found the kurtosis of log price changes to be infinite: the tails of the histogram of log-price changes appeared to be polynomial, like $1/x^\alpha$ with α between 4 and 5.

An important question is: how big are the tails of the histograms of image filter statistics. Two models have been proposed for these distributions. The first is the most commonly used model, the 'generalized Laplacian' distributions:

generalized Laplacian distribution

$$p_{\text{laplace}}(x) = \frac{e^{-|x/a|^b}}{Z}, \quad Z = \int e^{-|y/a|^b} dy.$$

Here a is a scale parameter and b controls how large the tails are (larger tails for smaller b). Experimentally, these work well and values of b between 0.5 and 1 are commonly found. However, no rationale for their occurrence seems to have been found. The second are the Bessel distributions (Grenander and Srivastava, 2001; Wainwright and Simoncelli, 2000):

Bessel distribution

$$p_{\text{bessel}}(x) = \widehat{q(\xi)}, \quad q(\xi) = 1/(1 + (a\xi^2))^{b/2}.$$

a is again a scale parameter, b controls the kurtosis (as before, larger kurtosis for smaller b) and the hat means Fourier transform. $p_{\text{bessel}}(x)$ can be evaluated explicitly using Bessel functions. The tails, however, are all asymptotically like those of double exponentials $e^{-|x/a|}$, regardless of b. The key point is these distributions arise as the distributions of *products* $r \cdot x$ of Gaussian random variables x and an independent positive 'scaling' random variable r. For some values of b, the variable r is distributed like $\|\vec{x}\|$ for a Gaussian $\vec{x} \in \mathbb{R}^n$, but in general its square has a Gamma (or Chi-squared) distribution. The great appeal of such a product is that images are also formed as products, especially as products of local illumination, albedo and reflectance factors. This may well be the deep reason for the validity of the Bessel models.

Convincing tests of which model is better have not been made. The difficulty is that they differ most in their tails, where data is necessarily very noisy. The best approach might be to use the Kolmogorov-Smirnov statistic and compare the best fitting models for this statistic of each type.

The world seems to be composed of discrete jumps in time and discrete objects in space. This profound fact about the physical nature of our world is clearly mirrored in the simple statistic — kurtosis.

1.2.2 Scaling properties of images and their implications

scale invariance

After high kurtosis, the next most striking statistical property of images is their approximate scale invariance. The simplest way to define scale invariance precisely is this: imagine we had a database of 64×64 images of the world and that this could be modeled by a probability distribution $p_{64}(I)$ in the Euclidean space $\mathbb{R}^{4096}$ of all such images. Then we can form marginal 32×32 images in two different ways: we either extract the central 32×32 set of pixels from the big image I or we cover the whole 64×64 image by 1024 2×2 blocks of pixels and average each such block to get a 32×32 image (i.e. we 'blow down' I in the crudest way). The assertion that images are samples from a scale invariant distribution is that the two resulting marginal distributions on 32×32 images are the same. This should happen for images of any size and we should also assume that the distribution is stationary, i.e. translating an image gives an equally probable image. The property is illustrated in Figure 1.2.

It is quite remarkable that, to my knowledge, no test of this hypothesis on reasonably large databases has contradicted it. Many histograms of filter responses

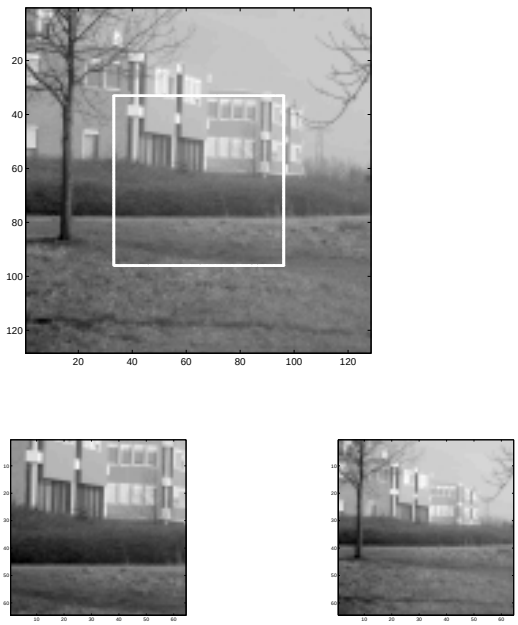

Figure 1.2 Scale invariance defined as a 'fixed point under block renormalization. The top is random $2n \times 2N$ image which produces the two $N \times N$ images on the bottom, one by extracting a subimage, the other by 2×2 block averaging. These two should have the same marginal distributions. (Figure from A. Lee.)

on successively blown down images have been made; order statistics have been looked at; and some topological properties derived from level curves have been studied (Huang and Mumford, 1999; Huang, 2000; Geman and Koloydenko, 1999; Gousseau, 2000). All have shown approximate scale invariance. There seem to be 2 simple facts about the world which combine to make this scale invariance approximately true. The first is that images of the world are taken from random distances: you may photograph your spouse's face from one inch away or from 100 meters away or anything in between. On your retina, except for perspective distortions, his or her image is scaled up or down as you move closer or farther away. The second is that objects tend to have surfaces on which smaller objects cluster: your body has limbs which have digits which have hairs on them, your office has furniture which has books and papers which have writing (a limiting case of very flat objects on its surface) on them, etc. Thus a blow up of a photograph shows roughly not only the same number of salient objects but they occur with roughly the same contrast[3].

The simplest consequence of scale invariance is the law for the decay of power at

3. It is the second idea that helps to explain why aerial photographs also show approximate scale invariance.

high frequencies in the Fourier transform of images (or better, the discrete cosine transform to minimize edge effects). It says that the expected power as a function of frequency should drop off like:

power law

$$\mathbb{E}_I\left(|\hat{I}(\xi,\eta)|^2\right) \approx C/(\xi^2+\eta^2) = C/f^2,$$

where $f = \sqrt{\xi^2+\eta^2}$ is the spatial frequency. This power law was discovered in the 1950's. In the image domain, it is equivalent to saying that the auto-correlation of the image is approximated by a constant minus log of the distance:

$$\mathbb{E}_I\left(\sum_{x,y}(I(x,y)-\bar{I}).(I(x+a,y+b)-\bar{I})\right) \approx C - \log(\sqrt{a^2+b^2}).$$

Note that the models have both infra-red[4] and ultra-violet divergences: the total power diverges for both $f \to 0$ and ∞ and the auto-correlation goes to $\pm\infty$ as $a, b \to 0$ and ∞. Many experiments have been made testing this law over moderate ranges of frequencies and I believe the conclusion to draw is this: for small databases of images, especially databases of special sorts of scenes such as forest scenes or city scenes, different powers are found to fit best. These range from $1/f^3$ to $1/f$ but with both a high concentration near $1/f^2$ and a surprisingly large variance[5] (Huang, 2000; Frenkel et al., 2004). But for *large* databases, the rule seems to hold.

Another striking consequence of the approximate scale-invariance is that images, if they have infinitely high resolution, are not functions at all but must be considered 'generalized functions' (distributions in the sense of Schwartz). This means that as their resolution increases, natural images do not have definite limiting numerical values $I(x,y)$ at almost all points x, y in the image plane. I think of this as the 'mites on your eyelashes' theorem. Biologists tell us that such mites exist and if you had superman's x-ray vision, you not only could see them but by the laws of reflectance, they would have high contrast, just like macroscopic objects. This mathematical implication is proven in (Gidas and Mumford, 2001).

This conclusion is quite controversial: others have proposed other function spaces as the natural home for random images. An early model for images (Mumford and Shah, 1989) proposed that observed images were naturally a sum:

$$I(x,y) = u(x,y) + v(x,y),$$

where u was a piecewise smooth 'cartoon', representing the important content of the image, and v was some L^2 noise. This led to the idea that the natural function space for images, after the removal of noise, was the space of functions of bounded variation, i.e. $\int ||\nabla I|| dxdy < \infty$. However, this approach lumped texture in with

4. The infra-red divergence is readily solved by considering images *mod constants*. If the pixel values are log of the photon energy, this constant is an irrelevant gain factor.
5. Some have found an especially large concentration near $1/f^{1.8}$ or $1/f^{1.9}$, especially for forest scenes (Ruderman and Bialek, 1994).

noise and results in functions u from which all texture and fine detail has been removed. More recent models, therefore, have proposed that:

$$I(x,y) = u(x,y) + v(x,y) + w(x,y),$$

where u is the cartoon, v is the true texture and w is the noise. The idea was put forward by DeVore and Lucier (DeVore and Lucier, 1994) that the true image $u+v$ belongs to a suitable Besov space, spaces of functions $f(x,y)$ for which bounds are put on the L^p norm of $f(x+h,y+k) - f(x,y)$ for (h,k) small. More recently, Carasso has simplified their approach (Carasso, 2004) and hypothesizes that images I, after removal of 'noise' should satisfy:

$$\int |I(x+h,y+k) - I(x,y)|dxdy < C(h^2+k^2)^{\alpha/2},$$

for some α as $(h,k) \to 0$.

However, a decade ago, Rosenfeld argued with me that most of what people discard as 'noise' is nothing but objects too small to be fully resolved by the resolution of the camera and thus blurred beyond recognition or even aliased. I think of this as *clutter*. The real world is made up of objects plus their parts and surface markings *of all sizes* and any camera resolves only so many of these. There is an ideal image of infinite resolution but any camera must use sensors with a positive point spread function. The theorem above says that this ideal image, because it carries all this detail, cannot even be a function. For example, it has more and more high frequency content as the sensors are refined and its total energy diverges in the limit[6], hence it cannot be in L^2.

In Figure 1.3, we illustrate that there is no clear dividing line between objects, texture and noise: depending on the scale at which you view and digitize the ideal image, the same 'thing' may appear as an object, as part of a texture or just a tiny bit of noise. This continuum has been analyzed beautifully recently by (Wu and Zhu, 2004).

random wavelet model

Is there is a simple stochastic model for images which incorporates both high kurtosis and scale-invariance? There is a unique scale-invariant Gaussian model, namely colored white noise whose expected power spectrum conforms to the $1/f^2$ law. But this has kurtosis equal to 3. The simplest model with both properties seems to be that proposed and studied by Gidas and me (Gidas and Mumford, 2001) which we call the *random wavelet* model. In this model, a random image is a countable sum:

$$I(x,y) = \sum_{\alpha} \psi_\alpha(e^{r_\alpha}x - x_\alpha, e^{r_\alpha}y - y_\alpha).$$

Here $(r_\alpha, x_\alpha, y_\alpha)$ is a uniform Poisson process in 3-space and ψ_α are samples from

6. Scale invariance implies that its expected power at spatial frequency (ξ,η) is a constant times $1/(\xi^2+\eta^2)$ and integrating this over (ξ,η) gives ∞.

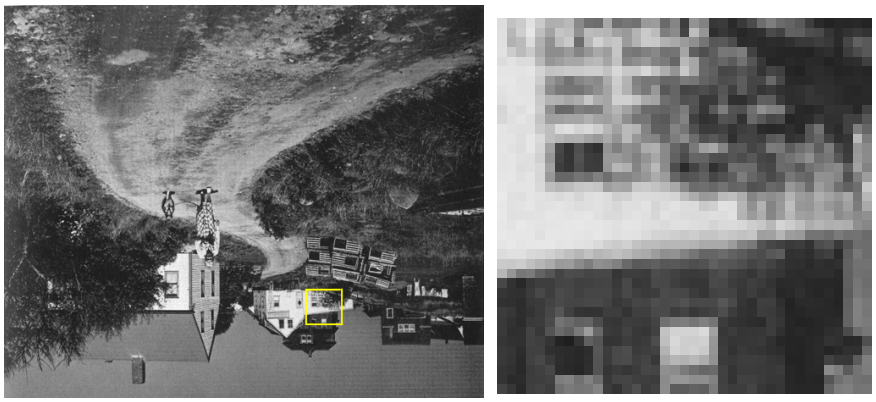

Figure 1.3 This photo is intentionally upside-down, so you can look at it more abstractly. The left photo has a resolution of about 500×500 pixels and the right photo is the yellow 40×40 window shown on the left. Note (a) how the distinct shapes in the road made by the large wet/dry spots gradually merge into dirt texture and (b) the way on the right the bush is pure noise. If the bush had moved relative to the pixels, the pattern would be totally different. There is no clear dividing line between distinct objects, texture and noise. Even worse, some road patches which ought to be texture are *larger* than salient objects like the dog.

the auxiliary 'Levy' process, a distribution on the space of scale and position normalized elementary image constituents, which one may call mother wavelets or textons. These expansions converge almost surely in all the Hilbert-Sobolev spaces $H^{-\epsilon}$. Each component ψ_α represents an elementary consitutent of the image. Typical choices for the ψ's would be Gabor patches, edgelets or curvelets or more complex shapes such as ribbons or simple shapes with corners. We will discuss these in Section 1.2.4 and we will return to the random wavelet model in Section 1.3.3.

1.2.3 Occlusion and the 'dead leaves' model

There is, however, a third basic aspect of image statistics which we have so far not considered: occlusion. Images are 2-dimensional projections of the 3-dimensional world and objects get in front of each other. This means that it is a mathematical simplification to imagine images as *sums* of elementary constituents. In reality, objects are ordered by distance from the lens and they should be combined by the non-linear operation in which nearer surface patches overwrite distant ones. Statistically, this manifests itself in a strongly non-Markovian property of images: suppose an object with a certain color and texture is occluded by a nearer object. Then, on the far side of the nearer object, the more distant object may reappear, hence its color and texture have a larger probability of occurring than in a Markov model.

This process of image construction was studied by the French school of Matheron and Serra based at the École des Mines (Serra, 1983 and 1988). Their

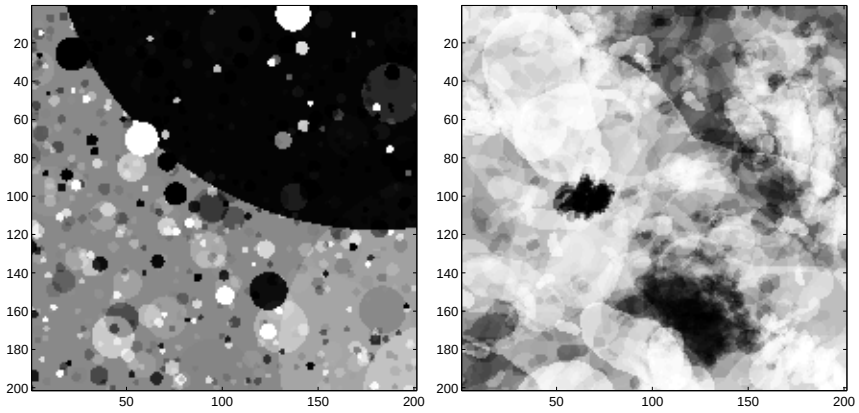

Figure 1.4 Synthetic images illustrating the generic image models from the text. On the left, a sample 'dead leaves model' using disks as primitives; on the right, a 'random wavelet model' whose primitive are short ribbons.

'dead leaves model' is similar to the above random wavelet expansion except that occlusion is used. We imagine that the constituents of the image are tuples $(r_\alpha, x_\alpha, y_\alpha, d_\alpha, D_\alpha, \psi_\alpha)$ where r_α, x_α and y_α are as before, but now d_α is the distance from the lens to the α^{th} image patch and ψ_α is a function only on the set of $(x, y) \in D_\alpha$. We make no a priori condition on the density of the Poisson process from which $(r_\alpha, x_\alpha, y_\alpha, d_\alpha)$ is sampled. The image is then given by:

$$I(x,y) = \psi_{\alpha(x,y)}(e^{r_{\alpha(x,y)}}x - x_{\alpha(x,y)}, e^{r_{\alpha(x,y)}}y - y_{\alpha(x,y)}) \qquad \text{where}$$
$$\alpha(x,y) = \text{argmin}\{d_\alpha \big| (x,y) \in D_\alpha\}$$

This model has been analyzed by A. Lee, J. Huang and myself (Lee et al., 2001) but has more serious infrared and ultraviolet catastrophes than the additive one. One problem is that nearby small objects cause the world to be enveloped in a sort of fog occluding everything in the distance. Another is the probability that one big nearby object occludes everything. In any case, with some cut-offs, Lee's models are approximately scale-invariant and seem to reproduce *all* the standard elementary image statistics better than any other that I know of, e.g. two-point co-occurrence statistics as well as joint wavelet statistics.

I believe a deeper analysis of this category of models entails modeling directly, not the objects in 2D projection, but their statistics in 3D. What is evident then is that objects are not scattered in 3-space following a Poisson process, but rather are agglutinative: smaller objects collect on or near the surface of bigger objects (e.g. houses and trees on the earth, limbs and clothes on people, buttons and collars on clothes, etc.). The simplest mathematical model for this would be a random branching process in which objects had 'children' which were the smaller objects clustering on its surface. We will discuss a 2D version of this in Section 1.3.3.

1.2.4 The phonemes of images

The final component of this direct attack on image statistics is the investigation of its elementary constituents, the ψ above. In analogy with speech, one may call these constituents phonemes (or phones). The original proposals for such building blocks were given by Julesz and Marr. Julesz was interested in what made two textures distinguishable or indistinguishable. He proposed that one should break textures locally into *textons* ((Julesz, 1981) and (Resnikoff, 1989), Chap. 6) and, supported by his psychophysical studies, he proposed that the basic textons were elongated blobs and their endpoints ('terminators'). Marr (Marr, 1982), motivated by the experiments of Hubel and Wiesel on the responses of cat visual cortex neurons, proposed that one should extract from an image its 'primal sketch', consisting of edges, bars and blobs. Linking these proposals with raw image statistics, Olshausen and Fields (Olshausen and Field, 1996) showed that simple learning rules seeking a *sparse* coding of the image, when exposed to small patches from natural images, did indeed develop responses sensitive to edges, bars and blobs. Another school of researchers have taken the elegant mathematical theory of wavelets and sought to find those wavelets which enabled best image compression. This has been pursued especially by Mallat (Mallat, 1999), Simoncelli (Simoncelli, 1999) and Donoho and their collaborators (Candes and Donoho, 2005).

texton

Having large natural image databases and powerful computers, we can ask now for a direct extraction of these or other image constituents from a statistical analysis of the image themselves. Instead of taking psychophysical, neurophysiological or mathematical results as a basis, what happens if we let images speak for themselves. Three groups have done this: Geman-Koloydenko (Geman and Koloydenko, 1999), Lee-Pedersen-Mumford (Lee et al., 2003a) and Malik-Shi (Malik et al., 1999).

The approach of Geman and Koloydenko was based on analyzing all 3×3 image patches using *order statistics*. The same image patches were studied by Lee and myself using their real number values. A very similar study by Lee and Pedersen (Pedersen and Lee, 2002) replaced the 9 pixel values by 9 Gaussian derivative filter responses. In all three cases, a large proportion of such image patches were found to be either low contrast or high contrast cut across by a single edge. This, of course, is not a surprise: but it quantifies the significance of edges in image structure. For example, in the study by Lee, Pedersen and myself, we took the image patches with the top 20% quantile for contrast, then subtracted their mean and divided by their standard deviation, obtaining data points on a 7-dimensional sphere. In this sphere, there is a surface representing the responses to image patches produced by imaging straight edges with various orientations and offsets. Close analysis shows that the data is highly concentrated near this surface, with asymptotic infinite density along the surface itself.

Malik and Shi take small patches and analyze these by a filter bank of 36 wavelet filters. They then apply k-means clustering to find high density points in this point cloud. Again the centers of these clusters resemble the traditional textons and primitives. In addition, they can adapt the set of textons they derive to individual

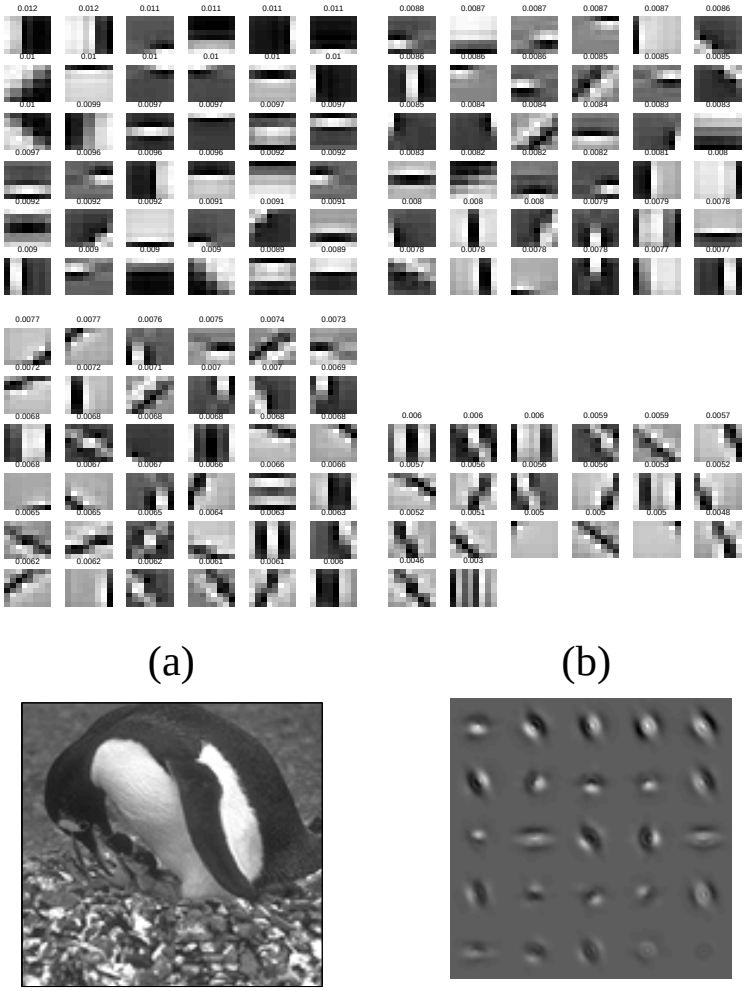

Figure 1.5 Textons derived by k-means clustering applied to 8×8 image patches. On the top, Huang's results for image patches from van Hateren's database; on the bottom, Malik et al's results using single images and filter banks. Note the occasional terminators in Huang's results, as Julesz predicted.

images, obtaining a powerful tool for representing a single image.

A definitive analysis of images deriving directly the correct vocabulary of basic image constituents has not been made but the outlines of the answer are now clear.

1.3 Grouping of image structures

In the analysis of signals of any kind, the most basic 'hidden variables' are the labels for parts of the signal which should be grouped together, either because they are homogeneous parts in some sense or because the components of this part occur together with high frequency. This grouping process in speech leads to words and in language leads to the elements of grammar — phrases, clauses and sentences. On the most basic statistical level, it seeks to group parts of the signal whose probability of occurring together is significantly greater than it would be if they were independent: see Section 1.3.3 for this formalism. The factors causing grouping were the central object of study for the so-called 'Gestalt' School of Psychology. This school flourished in Germany and later in Italy in the first half of the 20th century and included M. Wertheimer, K. Koffka, W. Metzger, E. Brunswik, G. Kanizsa and many others. Their catalog of features which promoted grouping included:

- color and proximity,
- alignment, parallelism and symmetry,
- closedness and convexity

Kanizsa was well aware of the analogy with linguistic grammar, titling his last book *Grammatica del Vedere* (Kanizsa, 1980). But they had no quantitative measures for the strength of these grouping principles, as they well knew. This is similar to the situation for traditional theories of human language grammar — a good story to explain what words are to be grouped together in phrases but no numbers. The challenge we now face is to create theories of *stochastic grammars* which can express why one grouping is chosen is chosen in preference to another. It is a striking fact that, faced either with a sentence or a scene of the world, human observers choose the same groupings with great consistency. This is in contrast with computers which, given only the grouping rules, find thousands of strange parses of both sentences and images.

1.3.1 The most basic grouping: Segmentation and texture

grouping

The simplest grouping rules are those of similar color (or brightness) and proximity. These two rules have been used to attack the segmentation problem. The most naive but direct approach to image segmentation is based on the assumption that images break up into regions on which their intensity values are relatively constant and across whose boundaries it changes discontinuously. A mathematical version of this approach, which gives an explicit measure for comparing different proposed

segmentations, is the energy functional proposed by Shah and myself (Mumford and Shah, 1989). It is based on a model $I = u + v$ where u is a simplified cartoon of the image and v is 'noise':

$$E(I,u,\Gamma) = C_1 \int_D (I-u)^2 + C_2 \int_{D-\Gamma} \|\nabla u\|^2 + C_3 \cdot \text{length}(\Gamma) \qquad \text{where}$$

$$\begin{aligned} D &= \text{ domain of } I \\ \Gamma &= \text{ boundaries of regions which are grouped together and} \\ C_i &= \text{ parameters to be learnt.} \end{aligned}$$

In this model, pixels in $D-\Gamma$ have been grouped together by stringing together pairs of nearby similarly colored pixels. Different segmentations correspond to choosing different u and Γ and the one with lower energy is preferred. Using the Gibbs statistical mechanics approach, this energy can be thought of as a probability: heuristically, we set $p(I,u,\Gamma) = e^{-E(I,u,\Gamma)/T}/Z$. where T and Z are constants. Taking this point of view, the first term in E is equivalent to assuming $v = I - u$ is a sample from white noise. Moreover, if Γ is fixed, then the second term in E makes u a sample from the scale-invariant Gaussian distribution on functions, suitably adapted to the smaller domain $D - \Gamma$. It is hard to interpret the third term even heuristically, although Brownian motion $((x(t), y(t))$ is heuristically a sample from the prior $e^{-\int (x'(t)^2 + y'(t)^2)dt}$, which, if we adopt arc length parametrization, becomes $e^{-\text{length}(\Gamma)}$. If we stay in the discrete pixel setting, the Gibbs model corresponding to E makes good mathematical sense; it is a variant of the Ising model of statistical mechanics (Geman and Geman, 1984; Blake and Zisserman, 1987).

The most obvious weakness in this model is its failure to group similarly textured regions together. Textural segmentation is an example of the hierarchical application of gestalt rules: first the individual textons are grouped by having similar colors, orientations, lengths and aspect ratios. Then these groupings of textons are further grouped into extended textured regions with homogeneous or slowly varying 'texture'. *Ad hoc* adaptations of the above energy approach to textural grouping (Geman and Graffigne, 1986; Lee et al., 1992; Hofmann et al., 1998) have been based on choosing some filter bank the similarity of whose responses are taken as a surrogate for the first low-level texton grouping. One of the problems of this approach is that textures are often not characterized so much by an *average* of all filter responses as by the *very large response* of one particular filter, especially by the outliers occurring when this filter precisely matches a texton (Zhu et al., 1997). A careful and very illuminating statistical analysis of the importance of color, textural and edge features on grouping, based on human segmented images, was given by Malik's group (Foulkes et al., 2003).

1.3.2 Extended lines and occlusion

The most striking demonstrations of gestalt laws of grouping come from occlusion phenomena, when edges disappear behind an object and reappear. A typical

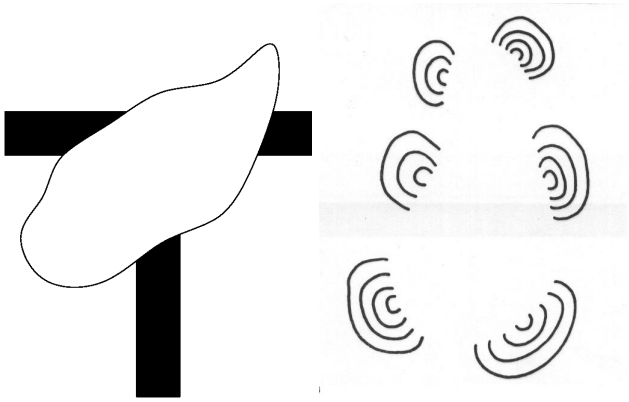

Figure 1.6 Two examples of gestalt grouping laws: on the left, the black bars are continued under the white blob to form the letter 'T', on the right, the semi-circles are continued underneath a foreground 'pear' which must completed by contours with zero contrast.

example is shown in Figure 1.6. The most famous example is the 'Kanizsa triangle' where, to further complicate matters, the foreground triangle has the same color as the background with only black circles of intermediate depth being visible. The grouping laws lead one to infer the presence of the occluding triangle and the completion of the three partially occluded black circles. An amusing variant, the 'Kanizsa pear', is shown in the same figure.

These effects are not merely psychophysical curiosities. Virtually every image of the natural world has major edges which are occluded one or more times by foreground objects. Correctly grouping these edges goes a long way to finding the correct parse of an image.

A good deal of modeling has gone into the grouping of disconnected edges into extended edges and the evaluation of competing groupings by energy values or probabilities. Pioneering work was done by Elder and Zucker (Parent and Zucker, 1989) and Shashua and Ullman (Shashua and Ullman, 1988). Nitzberg, Shiota and I proposed a model for this (Nitzberg et al., 1992) which was a small extension of the Mumford-Shah model. The new energy involves explicitly the overlapping regions R_α in the image given by the 3D objects in the scene, both the visible and the occluded parts of these objects. Therefore, finding its minimum involves inferring the occluded parts of the visible objects as well as the boundaries of their visible parts. (These are literally 'hidden variables'.) Moreover, we need the depth order of the objects, which are nearer, which further away. The cartoon u of the image is

now assumed piecewise constant with value u_α on the region R_α. Then:

$$E(I, \{u_\alpha\}, \{R_\alpha\}) = \sum_\alpha C_1 \int_{R'_\alpha} (I - u_\alpha)^2 + \int_{\partial R_\alpha} \left(C_2 \kappa^2_{\partial R_\alpha} + C_3\right) ds$$

$$R'_\alpha = \left(R_\alpha - \bigcup_{\text{nearer } R_\beta} R_\alpha \cap R_\beta \right) = \text{ visible part of } R_\alpha$$

$$\kappa_{\partial R_\alpha} = \text{ curvature of } \partial R_\alpha.$$

This energy allows one to quantify the application of gestalt rules for inferring occluded objects and predicts correctly, for example, the objects present in the Kanizsa triangle. The minima of this E will infer specific types of hidden contours, namely contours which come from the purely geometric variational problem of minimizing a sum of squared curvature and arc length along an unknown curve. This variational problem was first formulated by Euler, who called the resulting curves *elastica*.

To make a stochastic model out of this, we need a stochastic model for the edges occurring in natural images. There are two parts to this: one is modeling the local nature of edges in images and the other is modeling the way they group into extended curves.

Several very simple ideas for modeling curves locally, based on Brownian motion, were proposed in (Mumford, 1992). Brownian paths themselves are too jagged to be suitable, but one can assume the curves are C^1 and that their orientation $\theta(s)$, as a function of arc length, is Brownian. Geometrically, this is like saying their curvature is white noise. Another alternative is to take 2D projections of 3D curves whose direction of motion, given by a map from arc length to points on the unit sphere, is Brownian. Such curves have more corners and cusps, where the 3D path heads towards or away from the camera. Yet another option is generate parameterized curves whose velocity $(x'(t), y'(t))$ is given by two Ornstein-Uhlenbeck processes (Brownian functions with a restoring force pulling them to 0). These paths have nearly straight segments when the velocity happens to get large.

A key probability distribution in any such theory is $p(x, y, \theta)$, the probability density that if an image contour passes through $(0, 0)$ with horizontal tangent, then this contour will also pass through (x, y) with orientation θ. This function has been estimated from image databases by (Geisler et al., 2001), but I don't know of any comparison of their results with mathematical models.

Subsequently, Zhu (Zhu, 1999) and Ren and Malik (Ren and Malik, 2002) directly analyzed edges and their curvature in hand segmented images. Zhu found a high kurtosis empirical distribution much like filter responses: a peak at 0 showing the prevalence of straight edges and large tails indicating the prevalence of corners. He built a stochastic model for polygonal approximations to these curves using an exponential model of the form:

$$p(\Gamma) \propto e^{-\int_\Gamma \psi_1(\kappa(s)) + \psi_2(\kappa'(s)) ds},$$

where κ is the curvature of Γ and the ψ_i are unknown functions chosen so that the model yields the same distribution of κ, κ' as that found in the data. Finding continuum limits of his models under weak convergence is an unsolved problem. Ren and Malik's models go beyond the previous strictly local ones. They are k^{th}-order Markov models in which the orientation θ_{k+1} of a curve at a sample point P_{k+1} is a sample from a joint probability distribution of the orientations θ_k^α of both the curve and smoothed versions of itself at other scales α, all at the previous point P_k.

A completely different issue is finding probabilities that 2 edges should be joined, e.g. if Γ_1, Γ_2 are two curves ending at points P_1, P_2, how likely is it that in the real world there is a curve Γ_h joining P_1 and P_2 and creating a single curve $\Gamma_1 \cup \Gamma_h \cup \Gamma_2$. This link might be hidden in the image because of either occlusion, noise or low contrast (anyone with experience with real images will not be surprised at how often this happens). Jacobs, Williams, Geiger and others have developed algorithms of this sort based on elastica and related ideas (Williams and Jacobs, 1997; Geiger et al., 1998). Elder (Elder and Goldberg, 2002) and Geisler (Geisler et al., 2001) have carried out psychophysical experiments to determine the effects of proximity, orientation difference and edge contrast on human judgements of edge completions.

One of the subtle points here (as Ren and Malik make explicit) is that this probability does not depend only on the endpoints P_i and the tangent lines to the Γ_i at these points. So, for instance, if Γ_1 is straight for a certain distance before its endpoint P_1, then the longer this straight segment is, the more likely it is that any continuation it has will also be straight. An elegant analysis of the situation purely for straight edges has been given by Desolneux, Moisan and Morel (Desolneux et al., 2003). It is based on what they call 'maximally meaningful alignments', which come from computing the probabilities of accidental alignments and no other prior assumptions. The most compelling analysis of the problem, to my mind, is that in the thesis of Jonas August (August, 2001). He starts with a prior on a countable set of true curves, assumed to be part of the image. Then he assumes a noisy version of this is observed and seeks the maximally probable reconstruction of the whole set of true curves. An example of his algorithms is shown in Figure 1.7. Another algorithm for global completion of all image contours has been given recently by Malik's group (Ren et al., 2005).

1.3.3 Mathematical formalisms for visual grammars

The 'higher level' gestalt rules for grouping based on parallelism, symmetry, closedness and convexity are even harder to make precise. In this subsection, I want to describe a general approach to these questions.

So far, we have described grammars loosely as recursive groupings of parts of a signal, where the signal can be a string of phonemes or an image of pixels. The mathematical structure which these groupings define is a *tree*: each subset of the domain of the image which is grouped together defines a node in this tree and, whenever one such group contains another, we join the nodes by an edge. In the

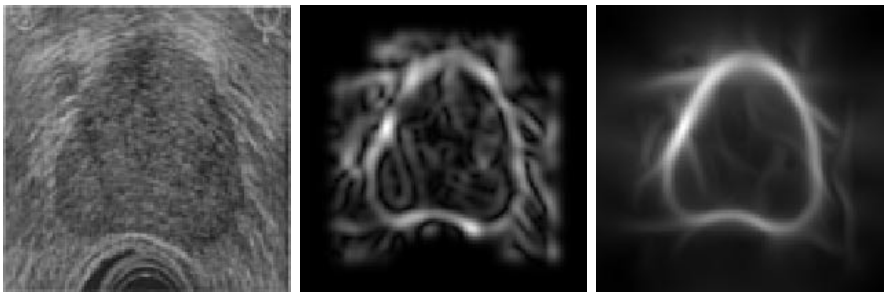

Figure 1.7 An experiment finding the prostate in a MRI scan from August (August, 2002). On the left, the raw scan; in the middle, edge filter responses; on the right, the computed posterior of August's *curve indicator random field*, (which actually lives in (x, y, θ) space, hence the boundary of the prostate is actually separated from the background noise).

case of sentences in human languages, this tree is called the parse tree. In the case of images, it is similar to the 'image pyramid' made up of the pixels of the image plus successively 'blown-down' images 2^n times smaller. However, unlike the image pyramid, its nodes only stand for natural groupings, so its structure is adaptively determined by the image itself.

To go deeper into the formalism of grammar, the next step is to label these groupings. In language, typical labels are 'noun phrase', 'prepositional clause' , etc. In images, labels might be 'edgelet', 'extended edge', ribbon', 'T-junction' or even 'the letter A'. Then the grouping laws are usually formulated as *productions*:

$$\begin{aligned} \text{noun phrase} &\longrightarrow \text{determiner} + \text{noun} \\ \text{extended edge} &\longrightarrow \text{edgelet} + \text{extended edge} \end{aligned}$$

where the group is on the left and its constituents are shown on the right. The second rule creates a long edge by adding a small piece, an edgelet, to one end. But now the issue of *agreement* surfaces: one can say 'a book' and 'some books' but not 'a books' or 'some book'. The determiner and the noun must agree in number. Likewise, to group an edge with a new edgelet requires that the edgelet connect properly to the edge: where one ends, the other must begin. So we need to endow our labeled groupings with a list of attributes which must agree for the grouping to be possible. So long as we can do this, we have created a context-free grammar. Context-freeness means that the possibility of the larger grouping depends only on the labels and attributes of the constituents and nothing else. An exanple of the parse of the letter 'A' is shown in Figure 1.8.

We make the above into a probability model in a top-down generative fashion by assigning to probabilities to each production. For any given label and attributes, the sum (or integral) of the probabilities of all possible productions it can yield should be 1. This is called a PCFG (probabilistic context-free grammar) by linguists. It is the same as what probabilists call a random branching tree (except that grammars

probabilistic context-free grammar

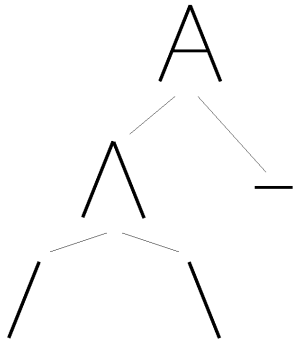

Figure 1.8 The parse tree for the letter 'A' which labels the top node; the lower nodes ight be labelled 'edge' and 'corner'. Note that in grouping the 2 sides, the edge has an attribute giving its length and approximate equality of the lengths of the sides must hold; and in the final grouping, the bar of the 'A' must meet the two sides in approximately equal angles. These are probabilistic constraints involving specific attributes of the constituents, which must be included in B_ℓ.

are usually assumed to almost surely yield *finite* parse trees).

A more general formalism for defining random trees with random data attached to their nodes has been given by Artur Fridman (Fridman, 2003). He calls his models 'Mixed Markov Models' because some of the nodes carry *address variables* whose value is the index of another node. Thus in each sample from the model, this node adds a new edge to the graph. His models include PCFG's as a special case.

Random trees can fit be fit naturally into the random wavelet model (or the dead leaves model) described above. To see this, we consider each 4-tuple $\{x_\alpha, y_\alpha, r_\alpha, \psi_\alpha\}$ in the model not merely as generating one elementary constituent of the image, but as the root of a whole random branching tree. The child nodes it generates should add parts to a now compound object expanding the original simple image constituent ψ_α. For example the root might be an elongated blob representing the trunk of a person and the tree it generates would add the limbs, clothes, face, hands, etc. to the person. Or the root might be a uniform patch and the tree would add a whole set of textons to it, making it into a textured patch. So long as the rate of growth of the random branching tree is not too high, we still get a scale-invariant model.

Two groups have implemented image analysis programs based on computing such trees. One is the multi-scale segmentation algorithm of Galun, Sharon, Basri and Brandt (Galun et al., 2003) which produces very impressive segmentation results. The method follows Brandt's adaptive tree-growing algorithm called 'algebraic multi-grid'. In their code, texture and its component textons play the same role as objects and their component parts: each component is identified at its natural scale and grouped further at a higher level in a similar way. Their code is fully scale-invariant except at the lowest pixel level. It would be very interesting to fit their scheme into the Bayesian framework.

The other algorithm is an integrated bottom-up and top-down image parsing

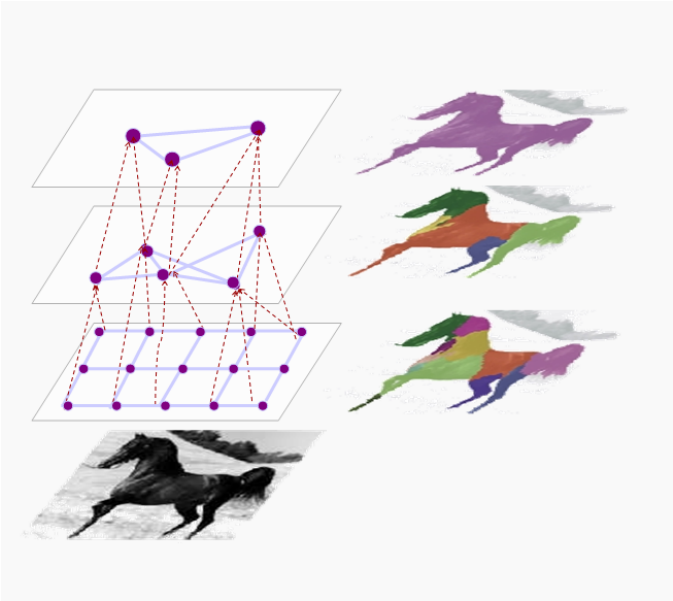

Figure 1.9 A simplification of the parse tree inferred by the segmentation algorithm of Galun, Sharon, Basri and Brandt. The image is at the bottom and part of its tree is shown above it. On the right are shown some of the regions in the image grouped by successive levels of the algorithm.

program from Zhu's lab (Tu et al., 2005). The output of their code is a tree with semantically labelled objects at the top, followed by parts and texture patches in the middle with the pixels at the bottom. This program is based on a full stochastic model.

A basic problem with this formalism is that it is not sufficiently expressive: the grammars of nature appear to be context-sensitive. This is often illustrated by contrasting languages that have sentences of the form *abcddcba*, which can be generated recursively by a small set of productions as in

$$s \rightarrow asa \rightarrow absba \rightarrow abcscba \rightarrow abcddcba,$$

versus languages which have sentences of the form *abcdabcd*, with two complex repeating structures, which cannot be generated by simple productions. Obviously, images with two identical faces are analogs of this last sentence. Establishing symmetry requires you to reopen the grouped package and examine everything in it to see if it is repeated! Unless you imagine each label given a huge number of attributes, this cannot be done in a context-free setting.

In general, 2-dimensional geometry creates complex interactions between groupings and the strength of higher order groupings seem to always depend on multiple aspects of the each piece. Take the example of a square. Ingredients of the square are a) the two groupings of parallel edges, each made up of a pair of parallel sides of equal length and b) the grouping of edgelets adjacent to each vertex into a 'right-

angle' group. The point is is that the pixels involved in these smaller groupings partially intersect. In PCFG's, each group should expand to disjoint sets of primitives or to one set contained in another. The case of the square is best described with the idea of *graph unification*, in which a grouping rule unifies parts of the graph of parts under each consitutent.

S. Geman and his collaborators (Bienenstock et al., 1998; Geman et al., 2002) have proposed a general framework for developing such probabilistic context-sensitive grammars. He proposes that for grouping rule ℓ, in which groups $y_1, y_2, \cdots, y_k$ are to be unified into a larger group x, there is a binding function $B_\ell(y_1, y_2, \cdots, y_k)$ which singles out those attributes of the constituents which affect the probability of making the k-tuple of y's into an x. For example, to put 2 edgelets together, we need to ask if the endpoint of the first is near the beginning of the second and whether their directions are close. The closer are these points and directions, the more likely it is that the two edgelets should be grouped. The basic hypothesis is that the likelihood ratio $p(x, y_1, \cdots, y_k)/\prod_i p(y_i)$ depends only on $B_\ell(y_1, \cdots, y_k)$. In their theory, they analyze how to compute this function from data.

This general framework needs to be investigated in many examples to further constrain it. An interesting example is the recent work of Ullman and collaborators (Ullman et al., 2002) on face recognition, built up through the recognition of parts: this would seem to fit into this framework. But, overall, the absence of mathematical theories which incorporate all the gestalt rules at once seems to me the biggest gap in our understanding of images.

1.4 Probability measures on the space of shapes

The most characteristic new pattern found in visual signals, but not in 1-dimensional signals, are *shapes*, 2-dimensional regions in the domain of the image. In auditory signals, one has *intervals* on which the sound has a particular spectrum, for instance, corresponding to some specific type of source (for phonemes, some specific configuration of the mouth, lips and tongue). But an interval is nothing but a beginning point and an endpoint. In contrast, a subset of a 2-dimensional region is much more interesting and conveys information by itself. Thus people often recognize objects by their shape alone and have a rich vocabulary of different categories of shapes often based on prototypes ('heart' shaped, 'egg' shaped, 'star' shaped, etc.). In creating stochastic models for images, we must face the issue of constructing probability measures on the space of all possible shapes. An even more basic problem is to construct metrics on the space of shapes, measures for the dissimilarity of 2 shapes. It is striking how people find it quite natural to be asked if some new object has a shape similar to some old object or category of objects. They act as though they carried a clearcut psychophysical metric in their heads, although, when tested, their similarity judgements show a huge amount of context sensitivity.

1.4.1 The space of shapes and some basic metrics on it

What do we mean by the space of shapes? The idea is simply to define this space as the set of 2-dimensional shapes, where a shape is taken to mean an open[7] subset $S \subset \mathbb{R}^2$ with smooth[8] boundary[9]. We let $\mathcal{S}$ denote this set of shapes. The mathematician's approach is to ask: what structure can we give to $\mathcal{S}$ to endow it with a geometry? In particular, we want to define a) local coordinates on $\mathcal{S}$, so that it is a manifold, b) a metric on $\mathcal{S}$ and c) probability measures on $\mathcal{S}$. Having probability measures will allow us to put shapes into our theory as hidden variables and extend the Bayesian inference machinery to include inferring shape variables from images.

manifold

$\mathcal{S}$ itself is not a vector space: one cannot add and subtract 2 shapes in a way satisfying the usual laws of vectors. Put another way, there is no obvious way to put global coordinates on $\mathcal{S}$, that is to create a bijection between points of $\mathcal{S}$ and points in some vector space. One can, e.g. describe shapes by their Fourier coefficients, but the Fourier coefficients coming from shapes will be very special sequences of numbers. What we can do, however, is put a *local linear structure* on the space of shapes. This is illustrated in Figure 1.10. Starting from one shape S, we erect normal lines at each point of the boundary Γ of S. Then nearby shapes will have boundaries which intersect each normal line in a unique point. Suppose $\psi(s) \in \mathbb{R}^2$ is arc length parametrization of Γ. Then the unit normal vector is given by $\vec{n}(s) = \psi'^{\perp}(s)$ and each nearby curve is parameterized uniquely in the form:

$$\psi_a(s) = \psi(s) + a(s) \cdot \vec{n}(s), \qquad \text{for some function } a(s).$$

All smooth functions $a(s)$ which are sufficiently small can be used, so we have created a bijection between an open set of functions a, that is an open set in a vector space, and a neighborhood of $\Gamma \in \mathcal{S}$. These bijections are called *charts* and on overlaps of such charts, one can convert the a's used to describe the curves in one chart into the functions in the other chart: this means we have a manifold. For details, see (Michor and Mumford, 2005). Of course, the function $a(s)$ lies in an infinite-dimensional vector space, so $\mathcal{S}$ is an infinite-dimensional manifold. But that is no deterrent to its having its own intrinsic geometry.

tangent space

Being a manifold means $\mathcal{S}$ has a tangent space at each point $S \in \mathcal{S}$. This tangent space consists in the infinitesimal deformations of S, i.e. those coming from infinitesimal $\epsilon a(s)$. Dropping the ϵ, the infinitesimal deformations may be thought of simply as normal vector fields to Γ, that is the vector fields $a(s).\vec{n}(s)$. We denote this tangent space as $T_{S,\mathcal{S}}$.

How about metrics? In analysis, there are many metrics on spaces of functions

7. A set S of points is *open* if S contains a small disk of points around each point $x \in S$.
8. *Smooth* means that its is a curve that is locally a graph of a function with infinitely many derivatives.
9. In many applications, one may want to include shapes with corners. We simplify the discussion here and assume there are no corners.

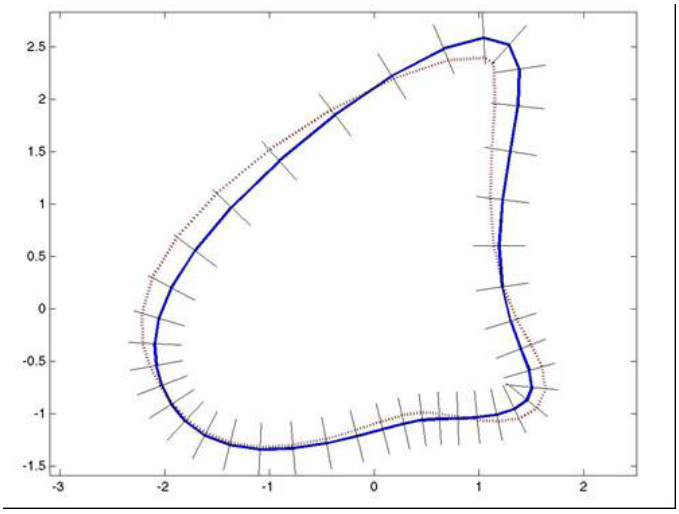

Figure 1.10 The manifold structure on the space of shapes is here illustrated: all curves near the heavy one meet the normal 'hairs' in a unique point, hence are described by a *function*, namely how far this point has been displaced normally.

and they vary in two different ways. One choice is whether you make a worst case analysis or an average analysis of the difference of two functions — or something in between. This means you define the difference of two functions a and b either as the $\sup_x |a(x) - b(x)|$, the integral $\int |a(x) - b(x)|dx$ or as an L^p norm, $(\int |a(x) - b(x)|^p dx)^{1/p}$ (which is in between). The case $p = \infty$ corresponds to the sup, and $p = 1$ to the average. Usually, the three important cases[10] are $p = 1, 2$ or ∞. The other choice is whether to include derivatives of a, b as well as the values of a, b in the formula for the distance and, if so, up to what order k. These distinctions carry over to shapes. The best known measures are the so-called Hausdorff measure

$$d_{\infty,0}(S,T) = \max\left(\sup_{x\in S}\inf_{y\in T}\|x-y\|, \sup_{y\in T}\inf_{x\in S}\|x-y\|\right),$$

for which $p = \infty, k = 0$ and the area metric,

$$d_{1,0}(S,T) = \text{Area}(S - S\cap T) \cup \text{Area}(T - S\cap T),$$

for which $p = 1, k = 0$.

It is important to realize that there is no one *right* metric on $\mathcal{S}$. Depending on the application, different metrics are good. This is illustrated in Figure 1.11, adapted from Kimia. The central bow-tie like shape is similar to all the shapes around it. But different metrics bring out their dissimilarities and similarities in each case. The

10. O. Faugeras et al. Charpiat et al. (2005) however have used p-norm as for $p \gg 1$ in order to 'tame' L^∞ norms.

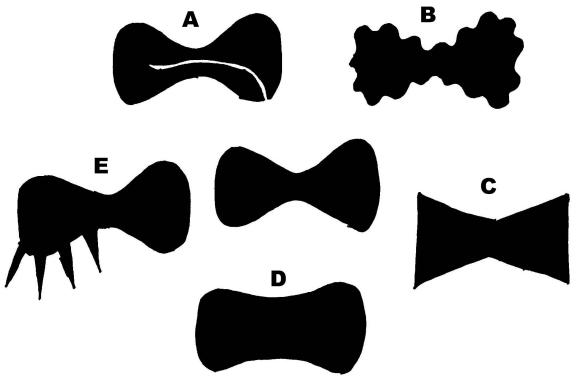

Figure 1.11 Each of the shapes A,B,C,D and E is similar to the central shape, but *in different ways*. Different metrics on the space of shape bring out these distinctions.

Hausdorff metric applied to the outsides of the shapes makes A far from the central shape; any metric using the first derivative (i.e. the orientation of the tangent lines to the boundary) makes B far from the central shape; a sup-type metric with the second derivative (i.e. the curvature of the boundary) makes C far from the central shape, as curvature becomes infinite at corners; D is far from the central shape in the area metric; E is far in all metrics, but the challenge is to find a metric in which it is close to the central shape. E has 'outliers', the spikes, but is identical to the central shape if they can be ignored. To do this needs what are called 'robust' metrics of which the simplest example is $L^{1/2}$ (not a true metric at all).

1.4.2 Riemannian metrics and probability measures via diffusion

Riemannian metrics

There are great mathematical advantages to using L^2, so-called Riemannian metrics. More precisely, a Riemannian metric is given by defining a quadratic inner product in the tangent space $T_{S,\mathcal{S}}$. In Riemannian settings, the unit balls are nice and round and extremal problems, such as paths of shortest length, are usually well-posed. This means we can expect to have geodesics, optimal deformations of one shape S to a second shape T through a family S_t of intermediate shapes, i.e. we can *morph* S to T in a most efficient way. Having geodesics, we can study the geometry of $\mathcal{S}$, for instance whether its geodesics diverge or converge[11] — which depends on the curvature of $\mathcal{S}$ in the metric. But most important of all, we can define diffusion and use this to get Brownian paths and thus probability measures on $\mathcal{S}$.

A most surprising situation arises here: there are three completely different ways to define Riemannian metrics on $\mathcal{S}$. We need to assign a norm to normal vector

11. This is a key consideration when seeking means to clusters of finite sets of shapes and in seeking 'principle components of such clusters.

fields $a(s)\vec{n}(s)$ along a simple closed plane curve Γ.

local metric

1. In 'infinitesimal' metric, the norm is defined as an integral along Γ. In general, this can be any expression:

$$\|a\|^2 = \int_\Gamma F(a(s), a'(s), a''(s), \cdots, \kappa(s), \kappa'(s), \cdots)ds$$

involving a function F quadratic in a and the derivatives of a whose coefficients can possibly be functions associated to Γ like the curvature and its derivatives. We call these local metrics. We might have $F = a(s)^2$ or $F = (1 + A\kappa^2(s)) \cdot a(s)^2$, where A is a constant; or $F = a(s)^2 + Aa'(s)^2$, etc.. These metrics have been studied by (Michor and Mumford, 2005; Michor and Mumford). Globally, the distance between two shapes is then:

$$d(S_0, S_1) = \inf_{\text{paths } \{S_t\}} \int_0^1 \|\frac{\partial S_t}{\partial t}\|dt,$$

where $\partial S_t/\partial t$ is the normal vector field given by this path.

diffeomorphism

2. In other situations, a morph of one shape to another needs to be considered as part of a morph of the whole plane. For this, the metric should be a quotient of a metric on the group $\mathcal{G}$ of diffeomorphisms of $\mathbb{R}^2$, with some boundary condition, e.g. equal to the identity outside some large region. But an infinitesimal diffeomorphism is just a vector field $\vec{v}$ on $\mathbb{R}^2$ and the induced infinitesimal deformation of Γ is given by $a(s) = (\vec{v} \cdot \vec{n}(s))$. Let $\mathbb{V}$ be the vector space of all vector fields on $\mathbb{R}^2$, zero outside some large region. Then this means that the norm on a is:

$$\|a\|^2 = \inf_{\vec{v}\in\mathbb{V}, (\vec{v}\cdot\vec{n})=a} \int_{\mathbb{R}^2} F(\vec{v}, \vec{v}_x, \vec{v}_y, \cdots)dxdy$$

where we define an inner product on $\mathbb{V}$ using a symmetric positive definite quadratic expression in $\vec{v}$ and its partial derivatives. We might have $F = \|\vec{v}\|^2$ or $F = \|\vec{v}\|^2 + A\|\vec{v}_x\|^2 + A\|\vec{v}_y\|^2$, etc. It is convenient to use integration by parts and write all such F's as $(L\vec{v}, \vec{v})$, where L is a positive definite partial differential operator

Miller's metric

($L = I - A\triangle$ in the second case above). These metrics have been studied by Miller, Younes and their many collaborators (Miller, 2002; Miller and Younes, 2001) and applied extensively to the subject they call 'computational anatomy', that is the analysis of medical scans by deforming them to template anatomies. Globally, the distance between two shapes is then:

$$d_{\text{Miller}}(S, T) = \inf_\phi \int_0^1 \left(\int_{\mathbb{R}^2} F(\frac{\partial \phi}{\partial t} \circ \phi^{-1})dxdy \right)^{1/2} dt, \quad \text{where}$$
$$\phi(t), 0 \le t \le 1 \text{ is a path in } \mathcal{G}, \phi(0) = I, \phi(1)(S) = T$$

Weil-Petersen metric

3. Finally, there is a remarkable and very special metric on $\bar{\mathcal{S}} = \mathcal{S}$ modulo translations and scalings (i.e. one identifies any two shapes which differ by translation plus a scaling). It is derived from complex analysis and known as the Weil-Petersen (or WP) metric. Its importance is that it makes $\bar{\mathcal{S}}$ into a *homogeneous* metric space,

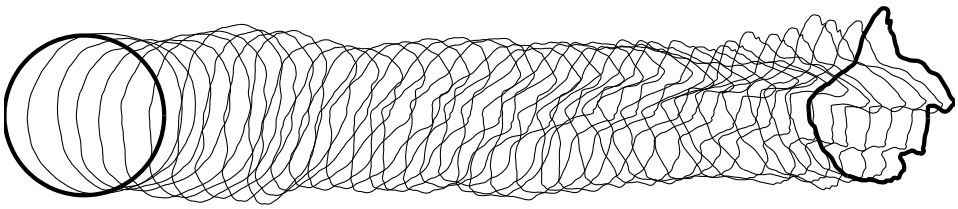

Figure 1.12 A diffusion on the space of shapes in the Riemannian metric of Miller et al. The shapes should be imagined on top of each other, the translation to the right being added in order that each shape can be seen clearly. The diffusion starts at the unit circle.

that is, it has everywhere the same geometry. There is a group of global maps of $\mathcal{S}$ to itself which preserve distances in this metric and which can take any shape S to any other shape T. This is not the case with the previous metrics, hence the WP metric emerges as the analog of the standard Euclidean distance in finite-dimensions. The definition is more elaborate and we do not give it here: see (Mumford and Sharon, 2004). This metric also has negative or zero curvature in all directions and hence finite sets of shapes as well as probability measures on $\bar{\mathcal{G}}$ should always have a well-defined mean (minimizing the sum of squares of distances) in this metric. Finally, this metric is closely related to the medial axis which has been frequently used for shape classification.

The next step in each of these theories is to investigate the heat kernel, the solution of the heat equation starting at a delta function. This important question has not been studied yet. But diffusions in these metrics are easy to simulate. In Figure 1.12 we show three random walks in $\mathcal{S}$ in one of Miller's metrics. The analog of Gaussian distributions are the probability measures gotten by stopping diffusion at a specific point in time. And analogs of the scale mixtures of Gaussians discussed above are obtained by using a so-called 'random stopping time', that choosing the time to halt the diffusion randomly from another probability distribution. It seems clear that one or more of these diffusion measures are natural general purpose priors on the space of shapes.

1.4.3 Finite approximations and some elementary probability measures

A completely different approach is to infer probability measures directly from data. Instead of seeking general purpose priors for stochastic models, one seeks special purpose models for specific object recognition tasks. This has been done by extracting from the data a finite set of *landmark points*, homologous points which can be found on each sample shape. For example, in 3 dimensions, skulls have long been compared by taking measurements of distances between classical landmark points. In 2 dimensions, assuming these points are on the boundary of the shape, the infinite dimensional space $\mathcal{S}$ is replaced by the finite dimensional space of the polygons $\{P_1, \cdots, P_k\} \in \mathbb{R}^{2k}$ formed by these landmarks. But, if we start from

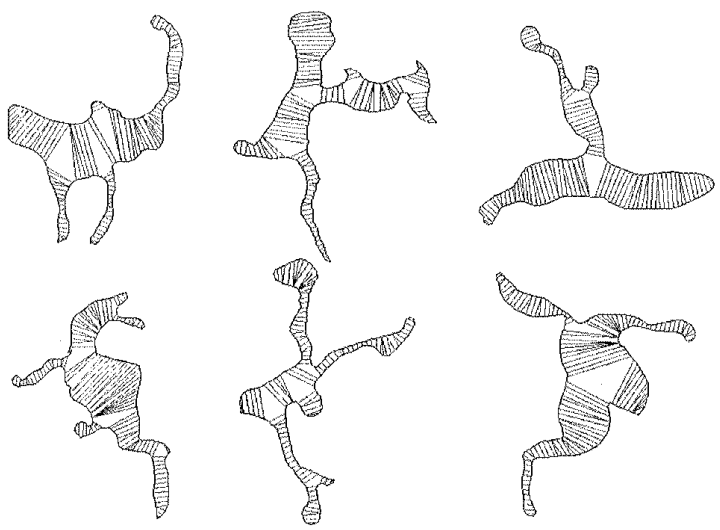

Figure 1.13 Six 'animals' that never existed: they are random samples from the prior of S.C. Zhu trained on real animal silhouettes. The interior lines come from his use of medial axis techniques to generate the shapes.

images, we can allow the landmark points to lie in the interior of the shape also. This approach was introduced a long time ago to study faces. More specifically, it was used by Cootes and Taylor (Cootes et al., 1993) and by Hallinan Hallinan et al. (1999) to fit multi-dimensional Gaussians to the cloud of points in $\mathbb{R}^{2k}$ formed from landmark points on each of a large set of faces. Both groups then apply principle component analysis and find the main directions for face variation.

However, it seems unlikely to me that Gaussians can give a very good fit. I suspect rather that in geometric situations as well, one will encounter the high kurtosis phenomenon, with geometric features often near zero but, more often than for Gaussian variables, very large too. A first attempt to quantify this point of view was made by Zhu Zhu (1999). He took a database of silhouettes of 4-legged animals, computed for each landmark points and their medial axis as well as curvature. Then he fit a general exponential model to a set of 6 scalar variables describing this geometry. The strongest test of whether he has captured some of their essential shape properties is to sample from the model he gets. The results are shown in Figure 1.13. It seems to me that these models are getting much closer to the sort of special purpose prior that is needed in object recognition programs. Whether his models have continuum limits and of what sort is an open question.

There are really three goals for a theory of shapes adapted to the analysis of images. The first is is to understand better the global geometry of $\mathcal{S}$ and which metrics are appropriate in which vision applications. The second is to create the best general purpose priors on this space, which can apply to arbitrary shapes. The third is to mold special purpose priors to all types of shapes which are encountered frequently, to express their specific variability. Some progress has been made on all

three of these but much is left to be done.

1.5 Summary

Solving the problem of vision requires solving three subproblems: finding the right classes of stochastic models to express accurately the variability of visual patterns in nature, finding ways to learn the details of these models from data and finding ways to reason rapidly using Bayesian inference on these models. This chapter has addressed the first. Here a great deal of progress has been made but it must be said that much remains to be done. My own belief is that good theories of groupings are the biggest gap. Although not discussed in this article, let me add that great progress has been made on the second and third problem with a large number of ideas, e.g. the EM-algorithm, much faster Monte Carlo algorithms, maximum entropy (MaxEnt) methods to fit exponential models, Bayesian belief propagation, particle filtering, graph-theoretic techniques.

References

J. August. *The Curve Indicator Random Field*. PhD thesis, Yale Computer Science, 2001.

J. August. Volterra filtering of noisy images. In *Proceedings, European Conference on Computer Vision*, 2002.

E. Bienenstock, S. Geman, and D.Potter. Compositionality, MDL priors and object recognition. In Jordan Mozer and Petsche, editors, *Advances in Neural Information Processing, vol.9*. MIT Press, 1998.

A. Blake and A. Zisserman. *Visual Reconstruction*. MIT Press, Cambridge, MA, 1987.

E. Brunswik. *Perception and the representative design of psychological experiments*. University of California Press, 1956.

E. Candes and D. Donoho. Continuous curvelet transform I and II. *Applied and Computational Harmonic Analysis*, 2005. To appear.

A. Carasso. Singular integrals, image smoothness and the recovery of texture in image deblurring. *SIAM Journal Applied Mathematics*, 64:1749–1774, 2004.

G. Charpiat, O. Faugeras, and R. Keriven. Approximations of shape metrics. *Journal of Foundations Of Computational Mathematics*, 2005. To appear.

T. Cootes, C. Taylor, A. Lanitis, D. Cooper, and J. Graham. Building and using flexible models. In *Proceedings 4th ICCV*, 1993.

A. Desolneux, L. Moisan, and J.-M. Morel. Maximal meaningful events and applications to image analysis. *Annals of Statistics*, 31:1822–1851, 2003.

R. A. DeVore and B. J. Lucier. Classifying the smoothness of images. In *Proceedings IEEE Int. Conf. Image Processing*, pages 6–10. IEEE press, 1994.

J. Elder and R. Goldberg. Ecological statistics of gestalt laws for the perceptual organization of contours. *Journal of Vision*, 2:324–353, 2002.

D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal Optical Society America, A*, 4:2379–94, 1987.

D. Forsyth and J. Ponce. *Computer Vision*. Prentice Hall, Upper Saddle River, NJ, 2002.

C. Foulkes, D. Martin, and J. Malik. Learning affinity functions for image segmentation. In *Proceedings IEEE Conference CVPR*, 2003.

G. Frenkel, E. Katzav, M. Schwartz, and N. Sochen. Distribution of anomalous exponents of natural images. *Physical Review Letters*, 2004. Submitted.

A. Fridman. Mixed markov models. *Proceedings National Academy Science*, 100: 8092–8096, 2003.

M. Galun, E. Sharon, R. Basri, and A. Brandt. Texture segmentation by multiscale aggregation of filter responses and shape elements. In *Proceedings ICCV, Nice*, pages 716–723, 2003.

D. Geiger, H-K. Pao, and N. Rubin. Salient and multiple illusory surfaces. In *Proceedings of the IEEE CVPR*, 1998.

A. Geisler, J. Perry, B. Super, and D. Gallogly. Edge co-occurence in natural images predicts contour grouping performance. *Vision Research*, 41:711–724, 2001.

D. Geman and A. Koloydenko. Invariant statistics and coding of natural microimages. In *Proceedings IEEE Workshop on Statistical and Computational Theories of Vision*, 1999.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions and bayesian restoration of images. *IEEE Transactions PAMI*, 6, 1984.

S. Geman and C. Graffigne. Markov random field image models and their applications to computer vision. In *Proceedings of the International Congress of Math.*, pages 1496–1517, 1986.

S. Geman, D. Potter, and Z. Chi. Composition systems. *Quarterly of Applied Mathematics*, 60:707–736, 2002.

B. Gidas and D. Mumford. Stochastic models for generic images. *Quarterly of Applied Mathematics*, 59:85–111, 2001.

Y. Gousseau. Morphological statistics of natural images. In *SPIE 4119, Wavelet applications in Signal and Image Processing*, volume 8, pages 208–214, 2000.

U. Grenander and A. Srivastava. Probability models for clutter in natural images. *IEEE Transactions PAMI*, 23:424–429, 2001.

P. Hallinan, G. Gordon, A. Yuille, P. Giblin, and D. Mumford. *Two- and Three-dimensional Patterns of the Face*. AKPeters, Wellesley, MA, 1999.

Ibn Al Haytham. *Kitab al-Manazir*. c.1000. In Arabic; translation by A.Sabra, *The Optics of Ibn Al-Haytham*, London: The Warburg Institute, 1989.

T. Hofmann, J. Puzicha, and J. M. Buhmann. Unsupervised texture segmentation in a deterministic annealing framework. *IEEE Transactions PAMI*, 20, 1998.

J. Huang. *Statistics of Natural Images and Models*. PhD thesis, Division Applied Mathematics, Brown, 2000. Available at:
`www.dam.brown.edu/people/mumford/Papers/Huangthesis.pdf`.

J. Huang and D. Mumford. Statistics of natural images and models. In *Proceedings IEEE Conference CVPR*, 1999.

B. Julesz. Textons, the elements of texture perception. *Nature*, 290:91–97, 1981.

G. Kanizsa. *Grammatica del vedere*. Società Editrice il Mulino, 1980. French

translation, *La Grammaire du Voir*, Diderot, 1997.

A. Lee, D. Mumford, and J. Huang. Occlusion models for natural images. *International Journal of Computer Vision*, 41:35–59, 2001.

A. Lee, K. Pedersen, and D. Mumford. The nonlinear statistics of high-contrast patches in natural images. *International Journal of Computer Vision*, 54:83–103, 2003a.

A. Lee, E. Simoncelli, A. Srivastava, and S.-C. Zhu. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, 18: 17–33, 2003b.

T. S. Lee, D. Mumford, and A. Yuille. Texture segmentation by minimizing vector-valued energy functionals. In *Proceedings European Conference Computer Vision*, volume 588 of *Lecture Notes in Computer Science*, pages 165–173, 1992.

J. Malik, S. Belongie, J. Shi, and T. Leung. Textons, contours and regions. In *Proceedings ICCV*, volume 7, pages 918–925, 1999.

S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

D. Marr. *Vision*. W.H.Freeman, 1982.

P. Michor and D. Mumford. Vanishing geodesic distance on spaces of submanifolds and diffeomorphisms. *Journal of the American Mathematical Society*. Submitted.

P. Michor and D. Mumford. Riemannian geometries on spaces of plane curves. *European Journal of Mathematics*, 2005. To appear.

M. Miller. On the metrics of euler-lagrange equations of computational anatomy. In *Annual Reviews of Biomedical Engineering*, volume 4, pages 375–405. Annual Reviews Press, 2002.

M. Miller and L. Younes. Group actions, homeomorphisms, and matching. *International Journal of Computer Vision*, 41:61–84, 2001.

D. Mumford. Elastica and computer vision. In C. Bajaj, editor, *Alg. Geom. and its Applications*. Springer, Berlin, 1992.

D. Mumford. Pattern theory, a unifying perspective. In D. Knill and W. Richards, editors, *Perception as Bayesian Inference*. Cambridge University Press, Cambridge, UK, 1996.

D. Mumford. Pattern theory: the mathematics of perception. In *Proc. Int. Congress Math.*, Beijing, China, 2002. Higher Education Press, Beijing.

D. Mumford and A. Desolneux. *Pattern Theory through Examples*. In preparation. Drafts at: `www.dam.brown.edu/people/mumford/Papers/IHP`.

D. Mumford and J. Shah. Optimal approximations of piuecewise smooth functions and associated variational problems. *Communications in Pure and Applied Mathematics*, 42:577–685, 1989.

D. Mumford and E. Sharon. 2d shape analysis using conformal mappings. In *Proceedings IEEE Conference CVPR*, 2004.

M. Nitzberg, D. Mumford, and T. Shiota. *Filtering, Segmentation and Depth*,

volume 662 of *Springer Lecture Notes in Computer Science*. Springer-Verlag, 1992.

B. Olshausen and D. Field. Emergence of simple cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

P. Parent and S. Zucker. Trace inference, curvature consistency and curve detection. *IEEE Transactions PAMI*, 11:823–839, 1989.

K. Pedersen and A. Lee. Toward a full probability model of edges in natural images. In *Springer Lecture Notes Computer Science*, volume 2350, page 328. Springer-Verlag, 2002.

X. Ren, C. Fowlkes, and J. Malik. Mid-level cues improve boundary detection. In *IEEE Conference CVPR*, 2005.

X. Ren and J. Malik. A probabilistic multi-scale model for contour completion based on image statistics. In *Proceedings European Conf.erence Computer Vision*, 2002.

H. Resnikoff. *The Illusion of Reality*. Springer-Verlag, 1989.

D. Ruderman and W. Bialek. Statistics of natural images: Scaling in the woods. *Physical Review Letters*, 73, 1994.

J. Serra. *Image Analysis and Mathematical Morphology I and II*. Academic Press, 1983 and 1988.

A. Shashua and S. Ullman. Structural saliency. In *Proceedings ICCV*, pages 321–327, 1988.

E. Simoncelli. Modeling the joint statistics of images in the wavelet domain. In *Proceedings 44th annual meeting SPIE*, pages 188–195, 1999.

Z. Tu, X. Chen, A. Yuille, and S-C. Zhu. Image parsing. *International Journal of Computer Vision*, 2005.

S. Ullman, M. Vidal-Naquet, and E. Sali. Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, 5:1–6, 2002.

H. van Hateren. The image database at: `hlab.phys.rug.nl/imlib`., 1998.

M. Wainwright and E. Simoncelli. Scale mixtures of gaussians and the statistics of natural images. In *Advances in Neural Information Processing Systems, 12*. MIT Press, 2000.

L. Williams and D. Jacobs. Stochastic completion fields. *Neural Computation*, 9: 837–858, 1997.

Y. Wu and S.-C. Zhu. From information scaling of natural images to regimes of statistical models. *Journal American Statistical Association*, 2004. Submitted.

S.C. Zhu. Embedding gestalt laws in markov random fields. *IEEE Transactions PAMI*, 21:1170–1187, 1999.

S.C. Zhu, Y. Wu, and D. Mumford. Minimax entropy principle and its application to texture modeling. *Neural Computation*, 9:1627–1660, 1997.