To Appear in IEEE Trans. on PAMI, 2003.

.

Statistical Modeling and Conceptualization of Visual Patterns

Song-Chun Zhu

Departments of Statistics and Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
sczhu@stat.ucla.edu

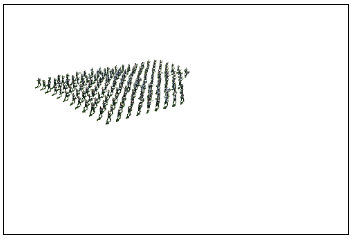
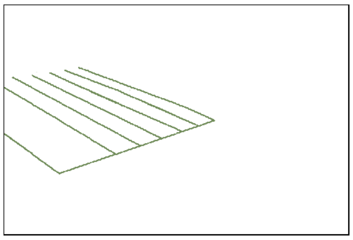
Abstract

Natural images contain an overwhelming number of visual patterns generated by diverse stochastic processes. Defining and modeling these patterns is of fundamental importance for generic vision tasks, such as perceptual organization, segmentation, and recognition. The objective of this epistemological paper is to summarize various threads of research in the literature and to pursue a unified framework for conceptualization, modeling, learning and computing visual patterns. The paper starts with reviewing four research streams: (1) the study of image statistics; (2) the analysis of image components; (3) the grouping of image elements; and (4) the modeling of visual patterns. The models from these research streams are then divided into four categories according to their semantic structures: (1) descriptive models, i.e. Markov random fields (MRF) or Gibbs); (2). variants of descriptive models (causal MRF and "pseudo-descriptive" models); (3) generative models; and (4) discriminative models. The objectives, principles, theories, and typical models are reviewed in each category, and the relationships between the four types of models are studied. Two central themes emerge from the relationship studies. (I) In representation, the integration of descriptive and generative models is the future direction for statistical modeling and should lead to richer and more advanced classes of vision models. (II) To make visual models computationally tractable, discriminative models are used as computational heuristics for inferring generative models. Thus the roles of four types of models are clarified. The paper also addresses the issue of conceptualizing visual patterns and their components (vocabularies) from the perspective of statistical mechanics. Under this unified framework a visual pattern is equalized to a statistical ensemble, and furthermore statistical models for various visual patterns form a "continuous" spectrum in the sense that they belong to a series of nested probability families in the space of attributed graphs.

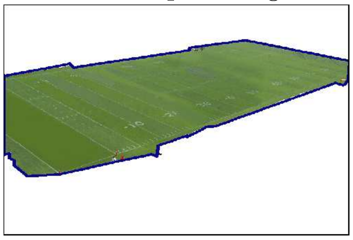
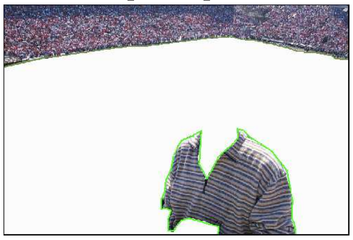
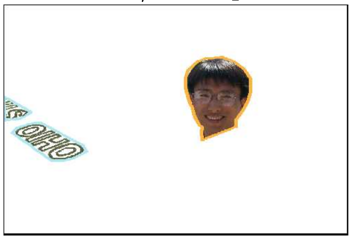
Keywords: perceptual organization, descriptive models, generative models, causal Markov models, discriminative methods, minimax entropy learning, mixed Markov models,

1 Introduction

1.1 Quest for a common framework of visual knowledge representation.

a. An input image b. A point process c. A line/curve process

d. A uniform region e. two texture regions f. objects: face and words

Figure 1: Parsing an image into its constituent patterns. Courtesy of (Tu and Zhu, ECCV 2002).

Natural images consist of an overwhelming number of visual patterns generated by very diverse stochastic processes in nature. The objective of image analysis is to parse generic images into their constituent patterns. For example, Figure 1.a shows an image of a football scene which is parsed into: b. a point process for the music band, c. a line and curve process for the field marks, d. a uniform region for the ground, e. two texture regions for the spectators, and f. two objects – words and human face. Depending on the types of patterns that a task is interested in, the image parsing problem is respectively called (1) *perceptual grouping* for point, line and curve processes, (2) *image segmentation* for region process, and (3) *object recognition* for high level objects. In other words, grouping, segmentation and recognition are sub-tasks of the image parsing problem and thus they ought to be solved in a unified way. This requests a common and mathematically sound framework for representing visual knowledge, and the visual knowledge includes two parts.

(1) Mathematical definitions and models of various visual patterns.

(2) Computational heuristics for effective inference of the visual patterns and models from raw images.

The objective of this epistemological paper is to pursue such a unified framework. More specifically, it should address the following four problems.

1. Conceptualization of visual patterns. What is a quantitative definition for a visual pattern? For example, what is a "texture"? what is a "human face"? The concept of a pattern is an abstraction of some properties decided by certain "vision purposes". These properties are feature statistics computing from either raw signals or some hidden descriptions inferred from raw signals. In both ways, a visual pattern is equalized to a set of observable signals governed by a statistical model –which we call an ensemble. In other words, each instance in the set is assigned a probability. For homogeneous patterns, such as texture, on large lattice this probability is a uniform distribution and the visual pattern is an equivalence class of images that satisfy certain descriptions. The paper should review some theoretical background in statistical mechanics and typical physics ensembles, from which a consistent framework is derived for defining various visual patterns.

2. Statistical modeling of visual patterns. First of all, why are the statistical models much needed in vision? In other words, what are the origin of these models? Some argued that probabilities are involved because of noise and distortions in images. This is truly a misunderstanding! With high quality digital cameras, there is rarely noise or distortions in images anymore. Probabilities are associated with the definitions of patterns and are even derived from deterministic definitions. In fact, the statistical models are intrinsic representation of visual knowledge and image regularities. Secondly, what are the mathematical space for patterns and models? Patterns are represented by attributed graphs, and thus models are defined in the space of attributed graphs. The paper should review two classes of models. One is descriptive model which are Markov random fields (or Gibbs) and its variants (including causal Markov models). The other is generative models which engage hidden variables for generating images in a top-down manner. It is shown that the two classes of models should be integrated. In the literature, a generative model often has a trivial descriptive component and a descriptive model usually has a trivial generative component. As a result of this integration, the models for various visual patterns, ranging from textures to geometric shapes, should form a "continuous spectrum" in the sense that they are from a series of nested probability families in this space.

3. Learning a visual vocabulary. What is the hierarchy of visual descriptions for general visual patterns? Can this vocabulary of visual description be defined quantitatively and learned from the ensemble of natural images? Compared with the large vocabulary in speech and language (such as phonemes, words, phrases, and sentences), and the rich structures in physics (such as electrons, atoms, molecules, and polymers), the current visual vocabulary is far from being enough for visual pattern representation. This paper reviews some progress in learning image bases and textons as visual dictionaries. These dictionaries are associated with generative models as parameters and are learned from natural images through model fitting.

4. Computational tractability. Besides the representational knowledge (definitions, models, and vocabularies), there is also computational knowledge. The latter are computational heuristics for effective inference of visual patterns, i.e. inferring hidden variables from raw images. These heuristics are the discriminative models which are approximations to posterior probability or ratios of posterior probabilities. The approximative posteriors are computed through local image features, in contrast to the real posterior computed by Bayes rule following generative models. Then it is natural to ask what are the intrinsic relationships between representational and computational models? Generally speaking, the generative models are expressed as top-down probabilities and the hidden variables have to be inferred from posterior probabilities following the Bayes rule, by Markov chain Monte Carlo techniques in general, such as the Metropolis-Hastings method. In contrast, the discriminative models approximate the posterior in a bottom-up and speedy fashion. These discriminative probabilities are used as proposal probabilities that drive the Markov chain search for fast convergence and mixing.

The questions raised above have motivated long threads of research from many disciplines, for example, applied mathematics, statistics, computer vision, image coding, psychology, and computational neurosciences. Recently a uniform mathematical framework emerges from the interactions between the research streams and experimentally a large number of visual patterns can be modeled realistically. This inspires the author to write an *epistemology* paper to summarize the progress in the field. The objective of the paper is to facilitate communications between different fields and provides a road map for the pursuit of a common mathematical theory for visual pattern representation and computation.

1.2 Plan of the paper

The paper consists of the following five parts.

Part 1. Literature survey. The paper starts with a survey of the literature in section (2) to set the background. We divide the literature in four research streams. (1) the study of natural image statistics, (2) the analysis of natural image components, (3) the grouping of natural image elements, and (4) the modeling of visual patterns. These streams develop four types of models, (i). descriptive model (Markov random fields or Gibbs), (ii) variants of descriptive models (causal MRF and "pseudo-descriptive" models), (iii) generative models, and (iv) discriminative model. The relationships of the models will be studied.

Part 2. A common framework for learning models. Section (3) presents a common maximum likelihood formulation for modeling visual patterns. Then it leads to the choice of two families of the probability models: descriptive models (and its variants) and generative models. Then the paper presents the descriptive and generative models in parallel.

Part 3. Descriptive models and its variants. This includes two sections. The paper firstly presents in Section (4) the basic assumptions and the minimax entropy principles for learning descriptive models and seven typical examples from low level image pattern to high level human face patterns in the literature. Secondly in Section (8), the paper discusses a few variants to the descriptive models, including causal Markov models and the pseudo-descriptive models.

Part 4. Generative models. In parallel, Section (6) presents the basic assumptions, methods, and five typical examples for learning generative models.

Part 5. Conceptualization of visual patterns. This includes two sections. Firstly in Section (5.2) it addresses the issue of conceptualization from the perspective of descriptive models. It presents the statistical physics foundation of descriptive models and three types of ensembles: the micro-canonical, canonical, and grand-canonical ensembles. Then it conceptualizes a visual pattern to an ensemble of physical states. In Section (7) the paper revisits the conceptualization of patterns from the perspectives of generative models, and states that the visual vocabulary can be learned as parameters in the generative models.

Part 5. Discriminative models. Then the paper turns to computational issues in section (9). It reviews how discriminative models can be used for inferring hidden structures in generative models, and presents maximum mutual information principle for selecting informative features for discriminations.

Finally Section (10) concludes the paper by raising some challenging issues in model selection and the balance between descriptive and generative models.

2 Literature survey — a global picture

In this section, we briefly review four research streams and summarize four types of probabilistic models to represent a global picture of the field.

2.1 Four research streams

Stream 1: the study of natural image statistics.

Any generic vision systems, biologic or machine, must account for image regularities. Thus it is of fundamental importance to study the statistical properties of natural images. Most of the early work studied natural image statistics from the perspective of image coding and redundancy reduction, and often use them to predict/explain the neuron responses.

In history, Attneave (1954), Barlow (1961), and Gibson (1966) were among the earliest who argued for the ecologic influence on vision perception. Kersten (1987) did perhaps the first experiment measuring the conditional entropy of the intensity at a pixel given the intensities of its neighboring pixels, in a spirit similar to Shannon's (1948) experiment of measuring the entropy of English words. Clearly the strong correlation of intensities between adjacent pixels results in low entropy. Further study of the intensity correlation in natural images leads to an interesting re-discovery of a $1/f$ power law by Field (1987).[1] By doing a Fourier transform on natural images, the amplitude of the Fourier coefficients at frequency f (averaged over orientations) fall off in a $1/f$-curve (see Figure 4.a). The power may not be exactly $1/f$ and vary in different image ensembles (Ruderman, 1994). This inspired a large body of work in biologic vision and computational neurosciences, which study the correlations of not only pixel intensities but responses of various filters at adjacent locations. These work also expand from grey level static images to color and motion images (see Atick et al 1992, Simoncelli et al 2001 for more references).

[1] The spectra power-law was firstly reported in (Deriugin, 1957) in studying television signals, and re-discovered by (Cohen and Carlson et al, 1975-78) in photographic analysis, and then (Burton and Moorhead, 1987) in optics study. It was Fields' work that brought it to attention of the broad vision communities.

Meanwhile, the study on natural image statistics extends from correlations to histograms of filter responses, for example, Gabor filters.[2] This leads to two interesting observations. Firstly the histograms of Gabor type filter responses on natural images have high kurtosis (Field, 1994). This reveals that natural images have high order (non-Gaussian) structures. Secondly, it was reported independently by (Ruderman, 1994) and (Zhu and Mumford 1996-97) that the histograms of gradient filtered images are consistent over a range of scales (see Fig. 5). The scale invariance experiment is repeated by several teams (Chi and S.Geman 1998, Grenander and Srivastava, 2001). Further studies along this direction include investigations on joint histograms and low dimensional manifolds in high dimensional spaces. For example, the density on a 7-D unit sphere for all 3×3 pixel patches of natural images (Lee, Huang, and Mumford, 2000, Koloydenko and D.Geman 2000). Going beyond pixel statistics, some most recent work measured the statistics of object shapes (Zhu, 1999), contours (Geisler et al. 2001), and the size of regions and objects in natural images (Alvarez, Grousseau, and Morel 1998).

Stream 2: the analysis of natural image components

The high kurtosis in image statistics observed in stream 1 is only a marginal evidence for hidden structures in natural scenes. A direct way for discovering structures and reducing image redundancy is to transform an image into a superposition of image components. For example, Fourier transform, wavelet transforms (Mallat 1989, Coifman and Wickerhauser 92), and various image pyramids (Simoncelli et al 1992) for generic images, and principal component analysis for some specific ensembles of images.

The transforms from image pixels to bases achieve two desirable properties. The first is variable decoupling. The coefficients of these bases are less correlated or become independent in ideal cases. The second is dimension reduction. The number of bases for approximately reconstructing an image is often much smaller than the number of pixels.

If one treats an image as a continuous function, then a mathematical tool for decomposing images is *harmonic analysis* (see Meyer 1985, 1988, Donoho etc 1998). Harmonic analysis is concerned with decomposing various classes of functions (i.e. mathematic spaces) by different bases. Further development along this vein includes the wedgelets, ridgelet, edgelets, curvelets by Donoho and Candes et al. in a series of papers.

Obviously the ensemble of natural images is quite different from those functional classes. Therefore, the image components must be adapted to natural images. This leads to an inspiring ideas in recent literature – *sparse coding* with *over-complete basis or dictionary* (Olshausen and Field, 1996). With over-complete basis, an image may be reconstructed by a small (sparse) number of bases in the dictionary. This often leads to 10-100 folds of dimension reduction. For example, an image of 200×200 pixels can be reconstructed approximately by about $100 - 500$ base images. Olshausen

[2]Correlations only measures second order moments while histograms include all the high order information, such as skewness (third order) and kurtosis (fourth order).

and Field then learned the over-complete dictionary from natural images. Figure 13 shows some of the bases. Added to this development is the independent component analysis (ICA) (Common 1994, VanHateren and Ruderman 1998). It is shown in harmonic analysis that the Fourier, wavelet, and ridgelet bases are independent components for various ensembles of mathematical functions (see Donoho et al 1998 and ref. therein). But for the ensemble of natural images, it is not possible to have an independent basis, and one can only compute a basis which maximize some measure of independence. Going beyond the image bases, recently (Zhu et al, 2002) proposed the texton representation with each texton consists of a number of image bases at various geometric, photometric, and dynamic configurations. If we compare the image bases to phonemes in speech, then the textons are larger structures corresponding to words.

Stream 3: the grouping of natural image elements

The third research stream originated from Gestalt psychology (Koffka, 1935). Human visual perception has strong tendency (bias) towards forming global percept ("whole" or pattern) by grouping local elements ("parts"). For example, Human vision completes illusory figures (Kaniza, 1979), and perceives hallucinatory structures from totally random dot patterns (Smith, 1986). In contrast to research streams 1 and 2, early work in stream 3 focused on *computational procedures and algorithms* that seem to demonstrate performance similar to human perception. This includes work on illusory figure completion and grouping from local edge elements (e.g. Guy and Medioni 1996).

While the Gestalt laws are quite successful in many artificial illusory figures, their applicability in real world images was haunted by ambiguities. A pair of edge elements may be grouped in one image but separated in the other image, depending on information that may have to be propagated from distant edge elements. So the Gestalt laws are not really *deterministic laws* but rather *heuristics* or *importance hypotheses* which are better used with probabilities.

Lowe (1985) was the first who computed the likelihoods (probabilities) for grouping a pair of line segments based on proximity, co-linearity, or parallelism respectively. Considering a number of line segments that are independently and uniformly distributed in terms of lengths, locations and orientations in a unit square, Lowe estimated the expected number for a pair of line segments at a certain configuration that are formed *accidentally* according to this uniform distribution. Lowe conjectured that the likelihood of grouping a pair of line segments in real images should be proportional to the inverse of this expected number – which he called *non-accidental property.* In a similar method, Jacobs (1993) calculated the likelihood for grouping a convex figure from a set of line segments. In a similar way, Moisan, Desolneux and Morel (2000) compute the likelihoods for "meaningful alignments". More advanced work include (Sarkar and Boyer, 1993) and (Dickinson et al. 1992) for generic object grouping and recognition (see Fig 20). Bienenstock, Geman, and Potter (1997) proposed a compositional vision approach for grouping of handwritten characters.

Generally speaking, the probabilities for grouping are, or can be reformulated to, posterior probability ratios of "grouping" vs "un-grouping", or "on" vs "off" an object. The probabilities are computed based on local features. Recently people started learning these probabilities and ratios from natural images with supervised input. For example, Geisler et al (2001) computed the *likelihood ratio* for the probability that a pair of edge elements appear in the same curve (grouped manually) against the probability that they appear in different curves (not grouped). Konishi et al (2003) computes probability ratio for a pixel on vs off the edge (object boundary) from some manually segmented images.

Stream 4: the modeling of natural image patterns

The fourth stream of research follows the Bayesian framework and develops explicit models for visual patterns. In the literature, Grenander (1976), Cooper (1979), and Fu (1982) were the pioneers using statistical models for various visual patterns. In the late 1980s and early 1990s, image models become popular and indispensable when people realized that vision problems, typically the shape-from-X problems, are fundamentally ill-posed. Extra information is needed to account for regularities in real world scenes, and the models represent our visual knowledge. Early models all assumed simple *smoothness* (sometimes piecewisely) of surfaces or image regions, and they were developed from different perspectives. For example, physically-based models (Terzopoulos, 1983, Blake and Zisserman, 1987), regularization theory (Poggio, Torre, and Koch, 1985), energy functionals (Mumford and Shah, 1989). Later these concepts all converged to statistical models which prevailed due to two pieces of influential work. The first work is the Markov random field (MRF) modeling (Besag, 1973, Cross and Jain, 1983) introduced from statistical physics. The second work is the Geman and Geman (1984) paper which showed that vision inference can be done rigorously by Gibbs sampler under the Bayesian framework. There were extensive literature on Markov random fields and Gibbs sampling in later 1980s. This trend went down in the early 1990s for two practical reasons. 1). Most of those Markov random field models are based on pair cliques and thus do not realistically characterize natural image patterns. 2). The Gibbs sampler is computationally very demanding on such problems.

Other probability models of visual patterns include deformable templates for objects, such as human face (Yuille, 1991) and hands (Grenander et al 1991). In contrast to the homogeneous MRF models for texture and smoothness, deformable templates are inhomogeneous MRF on small graphs whose nodes are labeled. We should return to more recent MRF models in later section.

2.2 Four categories of statistical models

The interactions of the research streams produce four categories of probability models. In the following, we briefly review the four types of models to set background for a mathematical framework that unifies them.

Category 1. Descriptive models.

Firstly, the integration of stream 1 and stream 4 yields a class of models which we call "descriptive models". Given an image ensemble and its statistics properties, such as the $1/f$-power law, scale invariant gradient histograms, studied in stream 1, one can alway construct a probability model which produces the same statistical properties as observed in the image ensemble. The probability is of the Gibbs (MRF) form following a maximum entropy principle (Jaynes, 1957). By maximum entropy, the model minimizes the bias while it satisfies the statistical descriptions. We call such models the descriptive models because they are constructed based on statistical descriptions of the image ensembles.

The descriptive model is attractive because a single probability model can integrate all statistical measures of different image features. For example, a Gibbs model of texture (Zhu, Wu, and Mumford, 1997) can account for the statistics extracted by a bank of filters, and a Gibbs model of shapes (2D simple curves) can integrate the statistics of various Gestalt properties: proximity, co-linearity, parallelism (Zhu, 1999). Such integration is not a simple product of the likelihoods or marginals on different features (like the projection pursuit method) but uses sophisticated energy functions to account for the dependency of these features. This provides a way to exactly measure the "non-accidental statistics" sought after by Lowe (1987). We shall deliberate on this point in latter section.

The descriptive models are all built on certain graph structures including lattices. There are two types descriptive models in the literature. 1). *Homogeneous models* where the statistics are assumed to be the same for all elements (vertices) in the graph. The random variables are the attributes of vertices, such as texture models. 2). *Inhomogeneous model* where the elements (vertices) of the graph are labeled, and different features and statistics are used at different sites. For example, deformable models of human faces.

Category 2. Variants of descriptive models and energy approximations

The descriptive models are often computationally expensive, due to the difficulty of computing the partition (normalizing) functions. This problem becomes prominent when the descriptive models have large image structures and account for high order image statistics. In the literature, there are a few variants to the descriptive models and approximative methods.

The first is causal Markov models. A causal MRF model approximates a descriptive model by imposing a partial (or even linear) order among the vertices of the graph such that the joint probability can be factorized as a product of conditional probabilities. The latter have lower dimensions and thus are much easier to learn and to compute. The Causal MRF models are still maximum entropy distributions subject to, sometimes, the same set of statistical constraints as the descriptive models. But the entropy is maximized in a limited probability space. Examples include texture synthesis in (Popat and Picard, 1993, Efros and Leung 1999) and the recent cut-and-paste work (Liang et al 2001, Efros and Freeman, 2001)

The second is called pseudo-descriptive model. Typical examples include texture synthesis methods by (Heeger and Bergen, 1995, DeBonet and Viola, 1997). They draw independent samples in the feature space, for example, filter responses at multiple scales and orientations at each pixel from the marginal or joint histograms. Though the sampled filter responses satisfy the statistical description in observed image, there is no image that can produce all these filter responses, as the latter are conflicting with each other. Then an image is synthesized by a pseudo-inverse method. Sampling in the feature space and the pseudo-inverse are often computationally convenient, but the whole method does not follow a rigorous model.

The other approximative approach for computing the descriptive model introduces a *belief* at each vertex. These beliefs are only normalized at single site or a pair of sites, and they do not necessarily form a legitimate (well normalized) joint probability for the whole graph. Thus it avoids computing the partition functions. This technique, originated in statistical physics, includes the mean field approximation, the Bethe and Kikuchi approximations (see Yedidia et al. 2000 and Yuille, 2001).

Category 3. Generative models

The principled way to tackling the computational complexity of descriptive models (no "hacks" or approximations) is to introduce hidden variables that can "explain away" the strong dependency in observed images. For example, the sparse coding scheme (Olshausen and Field, 1997) is a typical generative model which assumes an image being generated by a small number of bases. Other models include (Dayan et al 1995, Frey and Jojic, 1999). The computation becomes less intensive because of the reduced dimensions and the partially de-coupling of hidden variables. The generative model must engage some vocabulary of visual descriptions. For example, an over-complete dictionary for image coding. The elements in the vocabulary specify how images are generated from hidden variables.

The generative models are not separable from descriptive models, because the hidden variables must be characterized by a descriptive model, though in the literature the latter may often be a trivial iid Gaussian model or a causal Markov model. For example, the sparse coding scheme is a two layer generative model and assumes that the image bases are iid hidden variables. Hidden Markov models in speech and motion is a also two layer model whose hidden layer is a Markov chain (causal MRF model with linear order).

So descriptive and generative models must be integrated for developing richer and computationally tractable models. We should deliberate on this in latter sections. Thus we have a unified family of models for the descriptive (its variants) and generative models. These models are representational.

Category 4. Discriminative models

In contrast to the representational models (descriptive plus generative), some probabilities are better considered computational heuristics viewed from the general task of image parsing – the discriminative models used in stream 3 belong to this category.

In comparison, descriptive models and generative models are used as prior probabilities and likelihoods in the Bayesian framework, while discriminative models approximate the posterior probabilities of hidden variables (often individually) based on local features. As we shall show in later section that they are *importance proposal probabilities* which drive the stochastic Markov chain search for fast convergence. It was shown, through simple case, that the better the proposal probability approximate the posterior, the faster the algorithm converges (Mengersen and Tweedie, 1994).

The interaction between discriminative and generative models has not gone very far in the literature. Recent work include the data driven Markov chain Monte Carlo (DDMCMC) algorithms for image segmentation, parsing, and object recognition (Tu and Zhu et al 2000-02)[82, 83, 98].

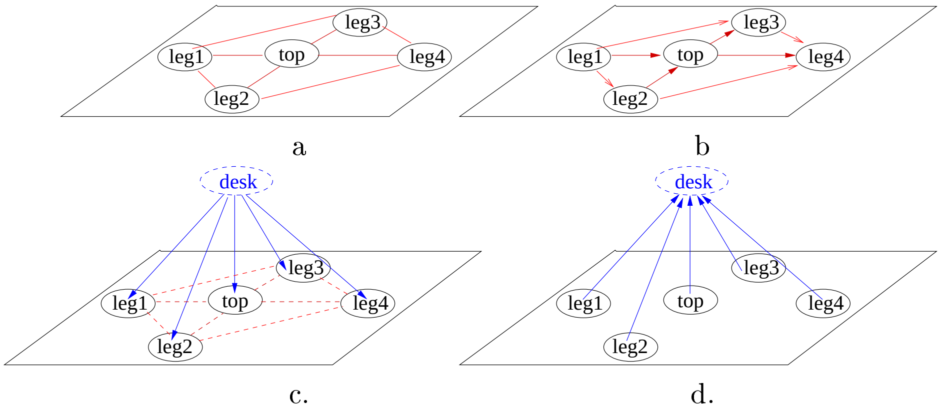


Figure 2: Four types of models for a simple desk object. a). descriptive (MRF), b). causal MRF, c). generative + descriptive d). discriminative.

Summary and justification of terminology

To clarify the terminology used above, Figure 2 shows a trivial example of the four models for a desk object. A desk consists of four legs and a top, denoted respectively by variables d, l_1, l_2, l_3, l_4, t for their attributes (vector valued). Figure 2.a shows the undirected graph for a descriptive model $p(l_1, l_2, l_3, l_4, t)$. It is in the Gibbs form with a number of energy terms to account for the spatial arrangement of the five pieces. The potential functions of the Gibbs assign low energies, and thus high probabilities, to more general configuration. This descriptive model accounts for the phenomological probability that the five pieces occur together without "understanding" a hidden concept of "desk" – denoted by hidden variable d. The causal MRF model assumes a directed graph in Figure 2.b. Thus it simplifies the descriptive model as $p(l_2)p(t|l_1, l_2)p(l_3|t, l_1)p(l_4|t, l_2, l_3)$. Figure 2.c is a two level generative model which involves a hidden variable d for the "whole" desk. The desk generates the five pieces by a model $p(l_1, l_2, l_3, l_4, t|d)$. d contains global attributes of the desk which controls the positions of the five parts. If we assume that the five pieces are conditionally independent, then it becomes a context free grammar (without the dashed lines). In general, we still need a descriptive model to characterize the spatial deformation by a descriptive model (see the dashed links). But this new descriptive model $p(l_1, l_2, l_3, l_4, t|d)$ is much less complicated than $p(l_1, l_2, l_3, l_4, t)$ in Figure 2.a. For example, if there are 5 types of desks, the descriptive model $p(l_1, l_2, l_3, l_4, t)$ must have complicated energy function so that it has 5 distinct modes (maxima). But if d contains a variable for the desk type, then $p(l_1, l_2, l_3, l_4, t|d)$ has a single mode for each type of desk and its potential is quite easy to compute. Finally Figure 2.d is a discriminative model, the links are pointed from parts to whole (reversing the generative arrows). It tries to compute a number of posterior probabilities $p(d|t)$, $p(d|l_i), i = 1, 2, 3, 4$. These probabilities are often treated as "votes" which are then summed up in a generalized Hough transform.

Syntactically the generative, causal Markov, and discriminative models can all be called Bayesian (causal, belief) networks as long as there are no loops in the graphs. But this terminology is very confusing in the literature. Our terminology for the four types of models are from a semantic perspective. We call it a generative model if the links are directed downwards in the conceptual hierarchy. We call it a discriminative model if the links are upwards. For example, the Bayes networks used by (Dickinson et al 1992, Sarkar and Boyer 1993-94, see Figure 20) are discriminative models. We call it a causal Markov model if the links are pointed to variables at the same conceptual level (also see Figure 17). For example we consider hidden Markov models in motion or speech as two layer generative models where the hidden variables is governed by a causal Markov (descriptive) model because they belong to the same semantic level. When a generative model is integrated with descriptive model, the integrated model can still be called generative model – a slight abuse of terminology.

It is worth noting that not all hidden variables are used in generative models. The mixed Markov model, as a variant of descriptive model, uses hidden variables to specify the neighborhood for variables at the same semantic level. These hidden variables are called "address variables" (Mumford, 1995). In contrast, the hidden variables in generative models represent entities of large structures.

3 Problem formulation

Now we start with a general formulation of visual modeling, from which we derive the descriptive and generative models for visual knowledge representation.

Let $\mathcal{E}$ denote the ensemble of natural images in our environment. As the number of natural images is so large, it makes sense to talk about a frequency $f(\mathbf{I})$ for images $\mathbf{I} \in \mathcal{E}$. $f(\mathbf{I})$ is intrinsic to our environment and our sensory system. For example, $f(\mathbf{I})$ would be different for fish living in deep ocean or rabbits living in prairie, or if our vision is 100 times more acute. The general goal of visual modeling is to estimate the frequency $f(\mathbf{I})$ by a probabilistic model $p(\mathbf{I})$ based on a set of observations $\{\mathbf{I}_1^{\text{obs}}, ..., \mathbf{I}_M^{\text{obs}}\} \sim f(\mathbf{I})$. $p(\mathbf{I})$ represents, exclusively, our understandings of image regularities and thus all of our representational knowledge for vision. [3]

It may sound quite ridiculous to estimate a density like $f(\mathbf{I})$ which is often in 256×256 space. But as we shall show in the rest of the paper, this is possible because of the strong regularities in natural images, and easy access to a very large number of images. For example, if a child sees 20 images per second, and open eyes 16 hours a day, then by the age of ten, he/she has seen 3

[3] A frequency $f(\mathbf{I})$ is an objective probability for the ensemble $\mathcal{E}$, while a model $p(\mathbf{I})$ is subjective and biased by the finite data observation and choice of model families.

billion images. The probability model $p(\mathbf{I})$ should approach $f(\mathbf{I})$ by minimizing a Kullback-Leibler divergence $KL(f||p)$ from f to p,

$$KL(f\,||\,p) = \int f(\mathbf{I}) \log \frac{f(\mathbf{I})}{p(\mathbf{I})} d\mathbf{I} = E_f[\log f(\mathbf{I})] - E_f[\log p(\mathbf{I})]. \tag{1}$$

Approximating the expectation $E_f[\log p(\mathbf{I})]$ by a sample average leads to the standard maximum likelihood estimator (MLE),

$$p^* = \arg\min_{p\in\Omega_p} KL(f\,||\,p) \approx \arg\max_{p\in\Omega_p} \sum_{m=1}^{M} \log p(\mathbf{I}_m^{\text{obs}}), \tag{2}$$

where Ω_p is a family of distributions where p^* is searched for. One general procedure is to search for p in a sequence of nested probability families,

$$\Omega_0 \subset \Omega_1 \subset \cdots \subset \Omega_K \to \Omega_f \ni f,$$

where K indexes the dimensionality of the space, e.g. the number of free parameters. As K increases, the probability family should be general enough to approach f to an arbitrary predefined precision.

There are two choices for the families Ω_p in the literature.

The first choice is the descriptive model. They are called exponential or log-linear models in statistics, and Gibbs models in physics. We denote them by

$$\Omega_1^d \subset \Omega_2^d \subset \cdots \Omega_K^d \to \Omega_f. \tag{3}$$

The dimension of the space Ω_i^d is augmented by increasing the number of *feature statistics* of $\mathbf{I}$.

The second choice is the generative model, or mixture models in statistics, denoted by

$$\Omega_1^g \subset \Omega_2^g \subset \cdots \subset \Omega_K^g \to \Omega_f \ni f. \tag{4}$$

The dimension of Ω_p is augmented by introducing hidden variables for the underlying image structures in $\mathbf{I}$.

Both families are general enough for approximating any distribution f. In the following sections, we deliberate on the descriptive and generative models and learning methods, and then discuss their unification and the philosophy of model selection.

4 Descriptive modeling

In this section, we review the basic principle of descriptive modeling, and show a spectrum of seven examples for modeling visual patterns from low to high level.

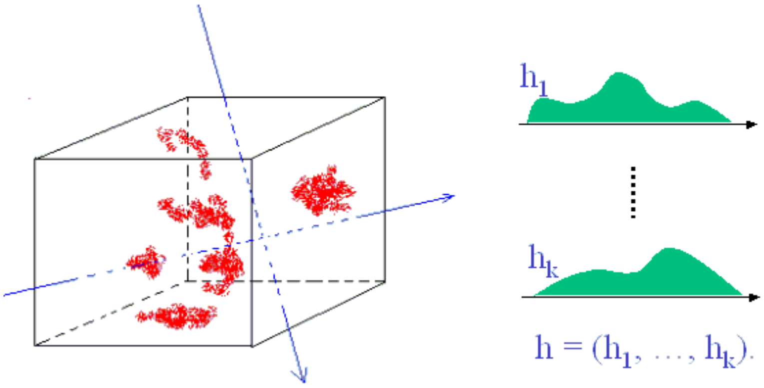


Figure 3: Descriptive modeling: estimating a high dimensional frequency f by a maximum entropy model p that matches the low dimensional (marginal) projections of f. The projection axes could be non-linear.

4.1 The basic principle of descriptive modeling

The basic idea of descriptive modeling is shown in Figure 3. Let $\mathbf{s} = (s_1, ..., s_n)$ be a representation of a visual pattern. For example $\mathbf{s} = \mathbf{I}$ could be an image with n pixels, and in general $\mathbf{s}$ could be a list of attributes for vertices in a random graph representation. An observable data ensemble is illustrated by a cloud of points in an n-space, and each point is an instance of the visual pattern. A descriptive method extracts a set of K **features** as *deterministic transforms* of $\mathbf{s}$, denoted by $\phi_k(\mathbf{s}), k = 1, ..., K$. For example, $\phi_k(\mathbf{I}) =< F, \mathbf{I} >$ is a projection of image $\mathbf{I}$ on a linear filter (say Gabor) F. These features (such as F) are illustrated by axes in Figure 3. In general, the axes don't not have to be straight lines and could be more than one dimensional. Along these axes we can compute the projected histograms of the ensemble (the right side of Figure 3). We denote these histograms as $\mathbf{h}_k^{\text{obs}}$ for features $\phi_k(\mathbf{s}), k = 1, 2, ..., K$. They are estimates to the marginal statistics of $f(\mathbf{s})$.

A model p must match the marginal statistics $\mathbf{h}_k^{\text{obs}}, k = 1, ..., K$ if it is to estimate $f(\mathbf{s})$. Thus, we have descriptive constraints:

$$E_p[h(\phi_k(\mathbf{s}))] = \mathbf{h}_k^{\text{obs}} \approx E_f[h(\phi_k(\mathbf{s}))], \quad k = 1, ..., K. \tag{5}$$

The least biased model that satisfies the above constraints is obtained by maximum entropy (Jaynes, 1957), and this leads to the FRAME model (Zhu, Wu, and Mumford, 1996-97),

$$p_{\text{des}}(\mathbf{s}; \boldsymbol{\beta}) = \frac{1}{Z(\boldsymbol{\beta})} \exp\{-\sum_{k=1}^{K} < \boldsymbol{\lambda}_k, h(\phi_k(\mathbf{s})) >\}. \tag{6}$$

The parameters $\boldsymbol{\beta} = (\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_K)$ are Lagrange multipliers and they are computed by solving the constraint equations (5). $\boldsymbol{\lambda}_k$ is a vector whose length is equal to the number of bins in the histogram $h(\phi_k(\mathbf{s}))$. As the features $\phi_k(\mathbf{s}), k = 1, 2, ..., K$ are often correlated, the parameters $\boldsymbol{\beta}$ are learned to weight these features. Thus $p_{\text{des}}(\mathbf{s}; \boldsymbol{\beta})$ integrates all the observed statistics. [4]

[4]In natural language processing, such Gibbs model was also used in modeling the distribution of English letters (Della Pietra, Della Pietra, and Lafferty 1997).

The selection of features in p_{des} is guided by a minimum entropy principle. For any new feature ϕ^+, we can define its non-accidental statistics following (Zhu, Wu, and Mumford, 1997).

Definition 1 : (non-accidental statistics). *Let $\mathbf{h}_f^+$ be the observed statistics for a novel feature ϕ^+ computed from the ensemble, i.e. $\mathbf{h}_f^+ \approx E_f[h(\phi_+(\mathbf{s}))]$ and $\mathbf{h}_p^+ = E_{p_{\text{des}}}[h(\phi_+(\mathbf{s}))]$ its expected statistics according to a current model p_{des}. Then the non-accidental statistics of ϕ^+, with its correlations to the previous K features removed, is a quadratic distance $d(\mathbf{h}_f^+, \mathbf{h}_p^+)$.*

$d(\mathbf{h}_f^+, \mathbf{h}_p^+)$ measures the statistics discrepancy of ϕ^+ which are not captured by the previous K features. Let p_{des}^+ be an augmented descriptive model with the K statistics in p_{des} plus the feature ϕ^+, then the following theorem is observed in (Zhu, Wu, and Mumford, 1997).

Theorem 1 (Feature Pursuit). *In the above notation, the non-accidental statistics of feature ϕ^+ is equal to the entropy deduction,*

$$d(\mathbf{h}_f^+, \mathbf{h}_p^+) = KL(f \,||\, p_{\text{des}}) - KL(f \,||\, p_{\text{des}}^+) = \text{entropy}(p_{\text{des}}) - \text{entropy}(p_{\text{des}}^+), \tag{7}$$

where $d(\mathbf{h}_f^+, \mathbf{h}_p^+)$ is a quadratic distance between the two histograms.

As entropy is the logarithmic volume of the ensemble governed by p_{des}, the higher the non-accidental statistics, the more informative feature ϕ^+ is for the visual pattern in terms of reducing uncertainty. Thus a feature ϕ^+ is selected sequentially following equation (7).

The Cramer and Wold theorem states that the descriptive model p_{des} can approximate any densities f using linear axes only (also see [95]).

Theorem 2 (Cramer and Wold) *Let f be a continuous density, then f is a linear combination of $\mathbf{h}$, the latter are the marginal distributions on the linear filter response $F^{(\xi)} * \mathbf{s}$, and f can be reconstructed by p_{des}.*

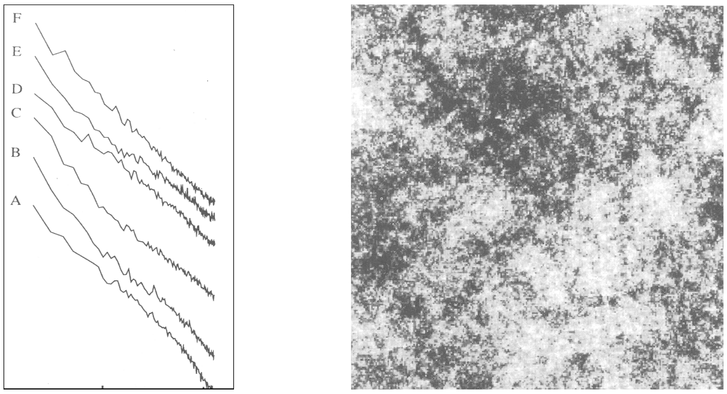
4.2 A spectrum of descriptive models for visual patterns

In the past few years, the descriptive models have successfully accounted for the observed natural image statistics (stream 1) and modeled a broad spectrum of visual patterns displayed in Fig. 1. In this subsection, we show seven examples.

Model D1. Descriptive model for $1/f$-power law of natural images

An important discovery in studying the statistics of natural images is the $1/f$ power-law (see review in stream 1). Let $\mathbf{I}$ be a natural image and $\hat{\mathbf{I}}(\xi, \eta)$ its Fourier transform. Let $A(f)$ be the Fourier amplitude $|\hat{\mathbf{I}}(\xi, \eta)|$ at frequency $f = \sqrt{\xi^2 + \eta^2}$ averaged over all orientations, then $A(f)$ falls off in a $1/f$-curve.

$$A(f) \propto 1/f, \quad \text{or} \quad \log A(f) = \text{const} - \log f.$$



a b

Figure 4: a). The log-Fourier-amplitude of natural images are plotted against $\log f$ (Field, 1987). b). A randomly sampled image with $1/f$ Fourier amplitude. (Mumford, 1995)

Figure 4.a is a result in logarithmic scale by Field (1987) for six natural images. The curves are fit well by straight lines in log-plot. This observation reveals that natural images contain equal Fourier power at each frequency band – scale invariance. That is,

$$\int\int_{f^2 \leq \xi^2+\eta^2 \leq (2f)^2.} |\hat{\mathbf{I}}^2(\xi,\eta)| d\xi d\eta = 2\pi \int_{f^2}^{4f^2} \frac{1}{f^2} df^2 = \text{const.}, \quad \forall f.$$

The descriptive model that accounts for such statistical regularity is surprisingly simple. It was showed by Mumford (1995) that a Gaussian Markov random field (GMRF) model below has exactly $1/f$-Fourier amplitude.

$$p_{1/\mathrm{f}}(\mathbf{I};\beta) = \frac{1}{Z} \exp\{-\sum_{x,y} \beta |\nabla \mathbf{I}(x,y)|^2\} \tag{8}$$

where $|\nabla \mathbf{I}(x,y)|^2 = (\nabla_x \mathbf{I}(x,y))^2 + (\nabla_y \mathbf{I}(x,y))^2$. ∇_x and ∇_y are the gradients. As the Gibbs energy is of a quadratic form and its matrix is real symmetric circulant, by a spectral analysis (see Priestley, 1981) its eigen-vectors are the Fourier bases and its eigen-values are the spectra.

This simply demonstrates that the much celebrated $1/f$-power law is nothing more than a second order moment constraint in the maximum entropy construction,

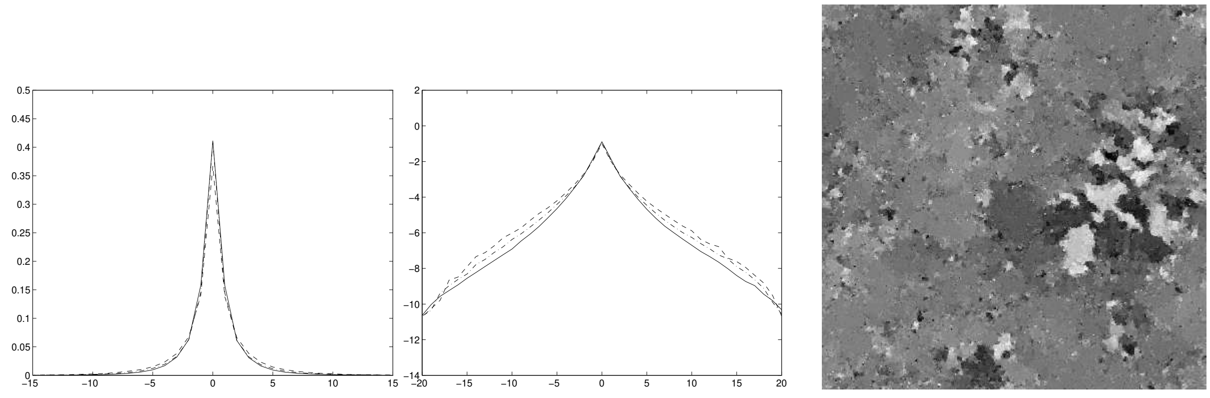
$$E_p[|\nabla \mathbf{I}(x,y)|^2] = \frac{1}{2\beta} \approx E_f[|\nabla \mathbf{I}(x,y)|^2], \quad \forall\, x, y. \tag{9}$$

This is equivalent to a $1/f$ constraint in the Fourier amplitude.

Since $p_{1/f}(\mathbf{I};\beta)$ is a Gaussian model, one can easily draw a random sample $\mathbf{I} \sim p_{1/\mathrm{f}}(\mathbf{I};\beta)$. Figure 4.b shows a typical sample image by Mumford (1995). It has very little structure in it! We will revisit the case in the generative model.

Model D2. Descriptive model for natural images with scale-invariant histograms

The second important discovery of natural image statistics is the scale-invariance of gradient histograms (Ruderman 1994, Zhu and Mumford, 1996-97). Take a natural image $\mathbf{I}$, and build



a b c

Figure 5: a). Gradient histograms over three scales. b). Logarithm of histograms. c). A randomly sampled images from a descriptive model $p_{\rm inv}(\mathbf{I};\boldsymbol{\beta})$. Courtesy of (Zhu and Mumford, 1996-97)

a pyramid with a number of n scales, $\mathbf{I} = \mathbf{I}^{(0)}, \mathbf{I}^{(1)}, ..., \mathbf{I}^{(n)}$. $\mathbf{I}^{(s+1)}$ is obtained by an average of 2×2 pixels in $\mathbf{I}^{(s)}$. The histograms $\mathbf{h}^{(s)}$ of gradients $\nabla_x \mathbf{I}^{(s)}(x,y)$ (or $\nabla_y \mathbf{I}^{(s)}(x,y)$) are plotted in Figure 5.a for three scales $s = 0, 1, 2$. Figure 5.b shows the logarithm of the histograms averaged over a number of images.

These histograms demonstrate high kurtosis and is amazingly consistent over a range of scales. Let $\mathbf{h}^{\rm obs}$ be the normalized histogram averaged over 3 scales, and impose constraints that a model p should produce the same histograms (marginal distributions),

$$E_p[h(\nabla_x \mathbf{I}^{(s)})] = E_p[h(\nabla_y \mathbf{I}^{(s)})] = \mathbf{h}^{\rm obs}, \quad s = 0, 1, 2, 3, \tag{10}$$

Zhu and Mumford (1996-97) derived a descriptive model,

$$p_{\rm inv}(\mathbf{I};\boldsymbol{\beta}) = \frac{1}{Z}\exp\{-\sum_{s=0}^{3} \sum_{(x,y)\in\Lambda^{(s)}} \lambda_x^{(s)}(\nabla_x \mathbf{I}^{(s)}(x,y)) + \lambda_y^{(s)}(\nabla_y \mathbf{I}^{(s)}(x,y))\}. \tag{11}$$

$\Lambda^{(s)}$ is the image lattice at scale s. $\boldsymbol{\beta} = (\lambda_x^{(0)}(), \lambda_y^{(0)}(), ..., \lambda_x^{(3)}(), \lambda_y^{(3)}())$ are the parameters and each $\lambda_x^{(s)}()$ is a 1D potential function quantized by a vector.

Figure 5.c shows a typical image sampled from this model by Gibbs sampler that was used in (Geman and Geman, 1984). This image has the scale-invariant histograms shown in Figure 5.a-b. Clearly the sampled image demonstrates some piecewise smoothness and consists of micro-structures of various sizes.

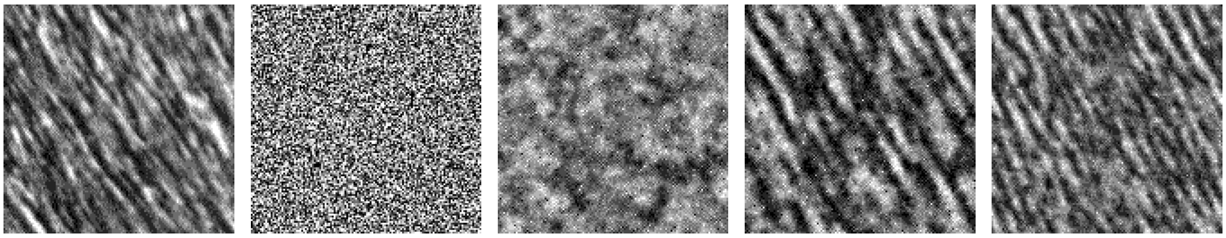
To make connection with other models, we remark on two aspects of $p_{\rm inv}(\mathbf{I};\boldsymbol{\beta})$.

Firstly, by choosing only one scale $s = 0$, the constraints in equation (10) is a super-set of the constraints in equation (9), as the histogram includes the variance. Therefore $p_{\rm inv}$ also observes the $1/f$-power law but with much more structures.

Secondly, with only one scale, $p_{\rm inv}$ reduces to the general smoothness models widely used in shape-from-X and denoising (see review of stream 4). The learned potential functions $\lambda_x()$ and

$\lambda_y()$ match pretty close to the manually selected energy functions. This bridges the learning of Gibbs model with PDEs in image processing (see details in [94]).

Model D3. Descriptive model for textures



a. observed b. $K = 0$ c, $K = 1$ d. $K = 2$ e. $K = 7$

Figure 6: Learning a sequence of descriptive models for a fur texture: a). The observed texture image, b),c),d),e), are the synthesized images as random samples from $p_{\rm tex}(\mathbf{I}; \boldsymbol{\beta})$ using $K = 0, 1, 2, 7$ filter histograms respectively. The images are obtained by Gibbs sampler. Courtesy of (Zhu, Wu, and Mumford, 1997)

The third descriptive model accounts for interesting psychophysical observations in texture study that histograms of a set of Gabor filters may be *sufficient statistics* in texture perception, i.e, two textures cannot be told apart in early vision if they share the same histograms of Gabor filters (Chubb and Landy, 1991).

Let $F_1, ..., F_K$ be a set of linear filters (such as Laplacian of Gaussian, Gabors), and $h(F_k * \mathbf{I})$ the histograms of filtered image $F_k * \mathbf{I}$ for $k = 1, 2, ..., K$. Each F_k corresponds to an axis and $h(F_k * \mathbf{I})$ a 1D marginal distribution in Fig.3. From an observed image, a set of histograms $\mathbf{h}_k^{\rm obs}, k = 1, 2, ..., K$ are extracted. By imposing the descriptive constraints

$$E_p[h(F_k * \mathbf{I})] = \mathbf{h}_k^{\rm obs}, \quad \forall k = 1, 2, ..., K. \tag{12}$$

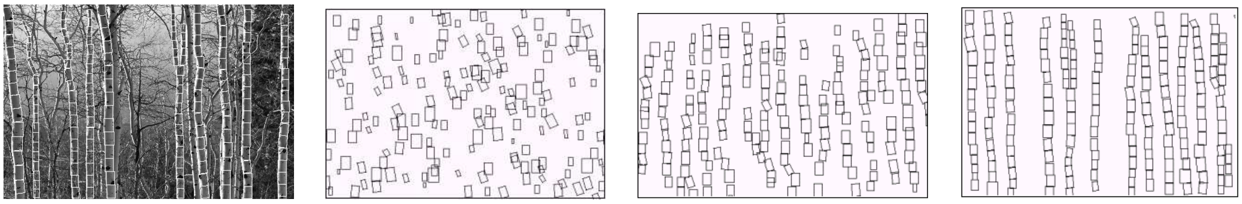
A FRAME model (Zhu, Wu, and Mumford, 1997-98) is obtained through maximum entropy.

$$p_{\rm tex}(\mathbf{I}; \boldsymbol{\beta}) = \frac{1}{Z} \exp\{- \sum_{(x,y)\in\Lambda} \sum_{k=1}^{K} \lambda_k (F_k * \mathbf{I}(x, y))\}. \tag{13}$$

where $\boldsymbol{\beta} = (\lambda_1(), \lambda_2(), ..., \lambda_K())$ are potential functions with each function $\lambda_i()$ being approximated by a vector. $p_{\rm tex}(\mathbf{I}; \boldsymbol{\beta})$ extends traditional Markov random field models (Besag, 1973, Cross and Jain, 1983) by replacing pairwise cliques with Gabor filters and by upgrading the quadratic energy to non-parametric potential functions which account for high order statistics.

Figure 6 illustrates the modeling of a texture pattern. As texture is homogeneous, it uses spatial average in a single input image in figure 6.a to estimate the ensemble average $\mathbf{h}_k^{\rm obs}, k = 1, 2, ..., K$. With $K = 0$ constraints, $p_{\rm tex}(\mathbf{I}; \Theta)$ is a uniform distribution and a *typical* random sample is a noise image shown in Figure 6.b. With $K = 1, 2, 7$ histogram constraints, the randomly sampled images from the learned Gibbs models $p_{\rm tex}(\mathbf{I}; \Theta)$ are shown in figures (6.c,d.e) respectively. The samples are drawn by Gibbs sampler[33] from $p_{\text{tex}}(\mathbf{I};\boldsymbol{\beta})$ and the selection of filters are governed by a minimax entropy principle[93]. A wide variety of textures are modeled in this way. In a similar way, one can put other statistics, such as filter correlations, in the model (See Portilla and Simoncelli, 2000).

Model D4. Descriptive model for texton (attributed point) process



a). observed b). $t = 1$ c). $t = 30$ d). $t = 332$

Figure 7: Different stages of simulating a wood pattern with local spatial interactions of textons. Each texton is represented by a small rectangle. After (Guo, Zhu, and Wu, 2001)

The descriptive models $p_{1/\text{f}}$, p_{inv}, and p_{tex} are all based on lattice and pixel intensities. Now we review a fourth model for texton (attributed point process) which extends lattices to graphs and extends pixel intensity to attributes. Texton processes are very important in perceptual organization. For example, Figure 1 shows a point process for the music band, and Figure 7.a shows a wood pattern where a texton represents a segment of the tree trunk.

Suppose a texton t has attributes x, y, σ, θ, c for its location, scale, orientation, and photometric contrast respectively. A texton pattern with an unknown number of n textons is represented by,

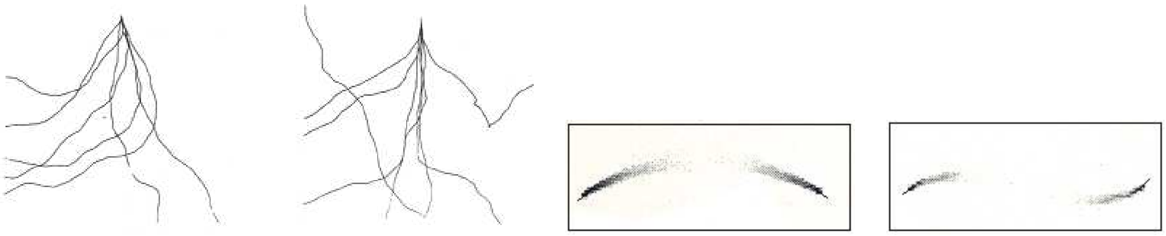
$$\mathbf{T} = (n, \{\ t_j = (x_j, y_j, s_j, \theta_j, c_j), j = 1, ..., n\ \}).$$

Each texton t has a neighborhood ∂t defined by spatial proximity, good continuity, parallelism or other Gestalt properties. It can be decided deterministically or stochastically. Once a neighborhood graph is decided, one can extract a set of features $\phi_k(t|\partial t), k = 1, 2, ..., K$ at each t measuring some Gestalt properties between t and its neighbors in ∂t. If the point patterns are homogeneous, then through constraints on the histograms, a descriptive model is obtained to capture the spatial organization of textons (Guo et al 2001),

$$p_{\text{txn}}(\mathbf{T}; \beta_o, \boldsymbol{\beta}) = \frac{1}{Z} \exp\{-\boldsymbol{\beta}_o n - \sum_{j=1}^{n} \sum_{k=1}^{K} \lambda_k(\phi_k(t_j|\partial t_j))\}, \tag{14}$$

p_{txn} is distinct from previous descriptive models in two respects. (1) The number of elements varies, thus a death-birth process must be used in simulating the model. (2) Unlike the static lattice, the spatial neighborhood of each element can change dynamically during the simulation.

Figure 7.a shows an example of a wood pattern with $\mathbf{T}$ given, from which a texton model p_{txn} is learned. Figure 7.b-d shows three stages of the MCMC sampling process of p_{txn} at $t = 1, 30, 332$ sweeps respectively. This example demonstrates that global pattern arises through simple local interactions in p_{txn}. More point patterns are referred to (Guo et al. 2001).



a b c d

Figure 8: a-b). Two sets of random sampled curves from the Elastica model, after (Mumford, 1994), c-d). The stochastic completion fields, After (Williams and Jacobs, 1997).

Model D5. Descriptive models for 2D open curves: Snake and Elastica

Moving up the hierarchy from point and textons to curves, we see most existing curve models are descriptive.

Let $C(s)$ $s \in [a,b]$ be an open curve, there are two curve models in the literature. One is the prior term used in the popular SNAKE or active contour model (Kass etc. 1987).

$$p_{\text{snk}}(C;\alpha,\beta) = \frac{1}{Z}\exp\{-\int_a^b \alpha|\nabla C(s)|^2 + \beta|\nabla^2 C(s)|^2 ds\},$$

where $\nabla C(s)$ and $\nabla^2 C(s)$ are the first and second derivatives.

The other is an Elastica model (Mumford, 1994) simulating a Ulenbeck process of a moving particle with friction, let $\kappa(s)$ be the curvature, then

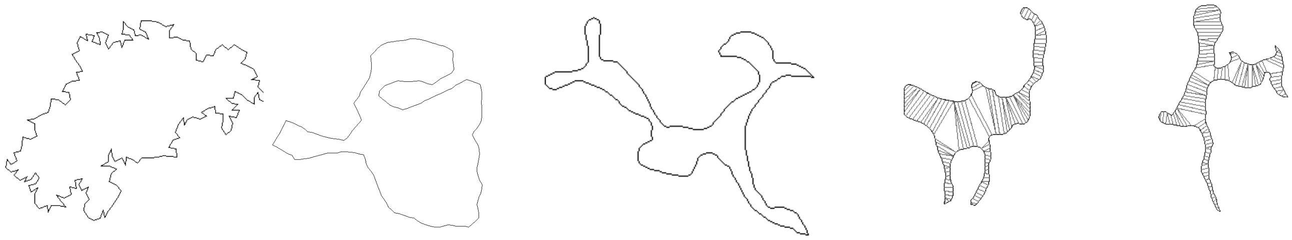
$$p_{\text{els}}(C;\beta) = \frac{1}{Z}\exp\{-\int_a^b [\alpha + \beta\kappa^2(s)]ds\}.$$

α controls the curve length as a decay probability for terminating the curve, like β_o in p_{txn}.

Figures 8.a-b show two sets of randomly sampled curves each starting from an initial point and orientation, the curves show general smoothness like the images in Figure 5.c. Williams and Jacobs (1997) adopted the Elastica model for curve completion. They define the so-called "stochastic completion field" between two oriented line segments (a source and a sink). Suppose a particle is simulated by a random walk, it starts from the source and ends at the sink. The completion fields shown in Figures 8.c-d display the probability that the particle passing a point (x,y) in the lattice (dark means high probability). This was used as a model for illusory contours.

Model D6. Descriptive models for 2D closed curves

The next descriptive model generalizes the smoothness curve model to 2D shape models with both contour and region based features. Let $\Gamma(s)$, $s \in [0,1]$ be a simple closed curve of normalized length. One can always represent a curve by polygon with a large enough number of vertices. Some edges can be added on the polygon for spatial proximity, parallelism, and symmetry. Thus a random graph structure is established, and some Gestalt properties $\phi_k(), k = 1,2,...,K$ can be extracted at each vertex and its neighbors, such as co-linearity, co-circularity, proximity, parallelism



a). $K=0$ b). $K=2$ c). $K=5$ d). $K=5$ e). $K=5$

Figure 9: Learning a sequence of models $p_{\rm shp}(\Gamma;\boldsymbol{\beta})$ for silhouettes of animals and plants, such as cats, dogs, fish, and leaves. a-e are typical samples from $p_{\rm shp}$ with $K=0,2,5,5,5$ respectively. The line segments show the medial axis features. Courtesy of (Zhu, 1999)

etc. Through constraints on the histograms of such features, a descriptive model is obtained in (Zhu, 1999),

$$p_{\rm shp}(\Gamma;\boldsymbol{\beta}) = \frac{1}{Z}\exp\{\sum_{k=1}^{K}\int_0^1 \lambda_k(\phi_k(s))ds\}, \tag{15}$$

This model is invariant to translation, rotation and scaling. By choosing features $\phi_k(s)$ to be $\nabla, \nabla^2, \kappa(s)$, this model is a non-parametric extension of the SNAKE and Elastica models on open curves.

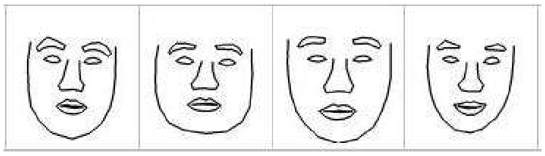
Figure 9 shows a sequence of shapes randomly sampled from $p_{\rm shp}(\Gamma;\boldsymbol{\beta})$. The training ensemble includes contours of animals and tree leaves. The sampled shapes at $K=0$ (i.e.. no features) are very irregular (sampled by Markov chain random walk under the hard constraint that the curve is closed and has no self-intersection. The MC starts with a circle), and become smooth at $K=2$ which integrates two features: co-linearity and co-circularity measured by the curvature and derivative of curvature $\kappa(s)$ and $\nabla\kappa(s)$ respectively. Elongated and symmetric "limbs" appear at $K=5$ when we integrates crossing region proximity, parallelism etc.

Model D7. Descriptive models for 2D human face

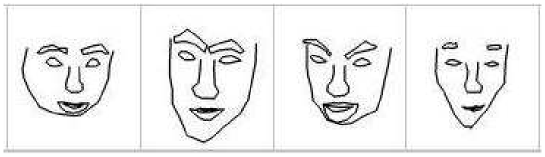
Moving up to high level patterns, descriptive models were used for modeling human faces (Yuille, 1991) and hand (Grenander et al. 1991), but early deformable models were manually designed, though in principle, they could be re-formulated in the maximum entropy form. Recently a descriptive face model is learned from data by (Liu etc. 2001) following the minimax entropy scheme.

A face is represented by a list of n (e.g. $n=83$) key points which are manually decided. Connecting these points forms the sketch shown in Figure 10. Thus each face is a point in a 166-space. After normalization in location, rotation and scaling, it has 162 dimensions. Figure 10.a shows four of example faces from the data ensemble.

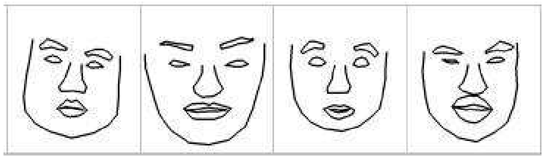
Unlike the previous homogeneous descriptive models where all elements in a graph (or lattice) are subject to the same statistical constraints, these key points on the face are labeled, and thus different statistical constraints are imposed at each location.



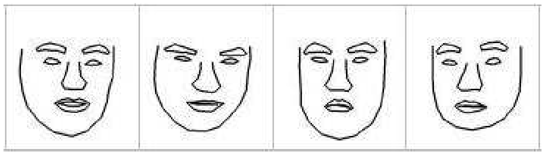
a. Observed examples of face



b. Random faces



c. Random faces with 4 features matched



d. Random faces with 17 features matched

Figure 10: Learning a sequence of face models $p_{\text{fac}}(V;\boldsymbol{\beta})$. a). Four of the observed faces from a training dataset, b,c,d). Four of the stochastically sampled faces with $K = 0, 4, 17$ statistics respectively. Courtesy of (Liu, Zhu, and Shum, ICCV, 2001)

Suppose we extract K features $\phi_k(V), k = 1, 2, ..., K$ on the graph V, then a descriptive model is,

$$p_{\text{fac}}(V;\boldsymbol{\beta}) = \frac{1}{Z}\exp\{-\sum_{k=1}^{K}\lambda_k(\phi_k(V))\}. \tag{16}$$

Liu etc. did a PCA to reduce the dimension first, and therefore the features $\phi_k(V)$ are extracted on the PCA co-efficients. Figure 10.b shows four sampled faces from a uniform model in the PCA-coefficient space bounded by the co-variances. The sampled faces in Figures 10.c-d become more pleasant as the number of features increases. When $K = 17$, the synthesized faces are no longer distinguishable from faces in the observed ensemble.

Summary: a continuous spectrum of models on the space of random graphs

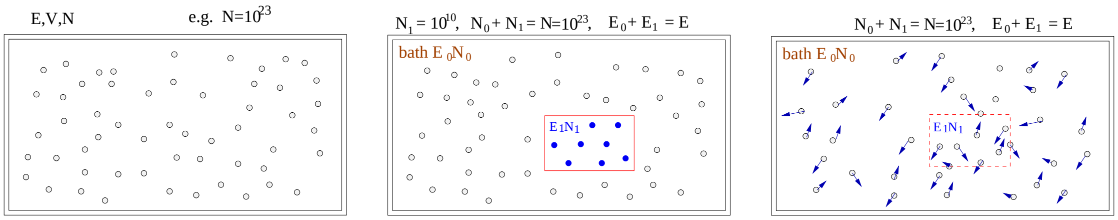
To summarize this section, visual patterns, ranging from generic natural images, textures, textons, curves, 2D shapes, and objects, can all be represented on attributed random graphs. All the descriptive models reviewed in this section are focused on different subspaces of a huge space of random graphs. Thus these models are examples in a "continuous" spectrum in the graph space (see eq.(3))! Though the general ideas of defining probability on random graphs were discussed in Grenander's pattern theory (Grenander, 1976-81), it will be a long way for developing such models as well as discovering a sufficient set of features and statistics on various graphs.

5 Conceptualization of visual patterns and statistics physics

Now we study an important theoretical issue associated with visual modeling: how do we define a visual pattern mathematically? For example, what is the definition of a human face, or a texture? In mathematics, a concept is equalized to a set. However a visual pattern is characterized by a probabilistic model as the previous section showed. The connection between a deterministic set and a statistical model was established in modern statistical physics.

5.1 Background: statistical physics and ensembles

Modern statistical physics is a subject studying macroscopic properties of a system involving massive amount of elements (see Chandler, 1987). Figure 11 illustrates three types of physical systems which are interesting to us.



a. micro-canonical ensemble b. canonical ensemble c. grand-canonical ensemble

Figure 11: Three typical ensembles in statistical mechanics.

1. Micro-canonical ensembles. Figure 11.a is an insulated system of N elements. The elements could be atoms, molecules, electrons in systems such as gas, ferro-magnetic material, fluid etc. N is really big, say $N = 10^{23}$ and is considered infinity. The system is decided by a configuration or state $s = (\mathbf{x}^N, \mathbf{m}^N)$, where $\mathbf{x}^N$ describes the coordinates of the N elements and $\mathbf{m}^N$ their momenta (Chandler, 1987). It is impractical to study the $6N$ vector s, and in fact, these microscopic states are less relevant, and people are more interested in the macroscopic properties of the system as a whole, say the number of elements N, the total energy $E(s)$, and total volume V. Other derivative properties are temperature and pressure etc.

If we denote by $\mathbf{h}(s) = (N, E, V)$ the macroscopic properties, at thermodynamic equilibrium all microscopic states that satisfy this property is called a *micro-canonical ensemble*,

$$\Omega_{mce}(\mathbf{h}_o) = \{s = (\mathbf{x}^N, \mathbf{m}^N) : \mathbf{h}(s) = \mathbf{h}_o = (N, V, E)\}.$$

s is an **instance** and $\mathbf{h}(s)$ is **summary** of the system state for practical purpose. Obviously Ω_{mce} is a deterministic set or an equivalence class for all states that satisfy a descriptive constraints $\mathbf{h}(s) = \mathbf{h}_o$.

An essential assumption in statistical physics is, as a first principle,

"all microscopic states are equally likely at thermodynamic equilibrium."

This is simply a maximum entropy assumption. Let $\Omega \ni s$ be the space of all possible states, then $\Omega_{mce} \subset \Omega$ is associated with a uniform probability,

$$p_{\text{unif}}(s; \mathbf{h}_o) = \begin{cases} 1/|\Omega_{mce}(\mathbf{h}_o)| & \text{for } \ s \in \Omega_{mce}(\mathbf{h}_o), \\ 0 & \text{for } \ s \in \Omega/\Omega_{mce}(\mathbf{h}_o). \end{cases}$$

2. Canonical ensembles. The canonical ensemble refers to a small system (with fixed volume V_1 and elements N_1) embedded in a micro-canonical ensemble, see Figure 11.b. The canonical ensemble can exchange energy with the rest system (called heat bath or reservoir). The system is relatively small, e.g $N_1 = 10^{10}$, so that the bath can be considered a micro-canonical ensemble itself.

At thermodynamic equilibrium, the microscopic state s_1 for the small system follows a Gibbs distribution,

$$p_{\text{Gib}}(s_1; \boldsymbol{\beta}) = \frac{1}{Z} \exp\{-\boldsymbol{\beta} E(s_1)\}$$

The conclusion was stated as a general theorem by Gibbs (1902),

> *"If a system of a great number of degrees of freedom is micro-canonically distributed in phase, any very small part of it may be regarded as canonically distributed."*

Basically this theorem states that the Gibbs model p_{Gib} is a conditional probability of the uniform model p_{unif}. This conclusion is extremely important because it bridges a deterministic set Ω_{mce} with a descriptive model p_{Gib}. We consider this is a true origin of probability for modeling visual patterns. Some detailed deduction of this conclusion in vision models can be found in (Wu and Zhu, 1999).

3. Grand-Canonical ensembles. When the small system with fixed volume V_1 can also exchange elements with the bath as in liquid and gas materials, then it is called a grand-canonical ensemble, see Figure 11.c. The grand-canonical ensemble follows a distribution,

$$p_{\text{gce}}(s_1; \beta_o, \beta) = \frac{1}{Z} \exp\{-\beta_o N_1 - \beta E(s_1)\},$$

where an extra parameter β_o controls the number of elements N_1 in the ensemble.

5.2 Conceptualization of visual patterns

In statistical mechanics, one is concerned with macroscopic properties for practical purpose and ignores the differences between the enormous number of microscopic states. Similarly, our concept of a visual pattern must be defined for a purpose. The purpose is reflected in the selection of some "sufficient" statistics $\mathbf{h}(s)$. That is, depending on a visual task, we are only interested in some global (macro) properties $\mathbf{h}(s)$, and ignore the differences between image instances within the set. This was clearly the case in Julesz psychophysics experiments in the 1960s-70s on texture discrimination (See Julesz, 1995). Thus we define a visual concept in the same way as the micro-canonical ensemble.

Definition 2 : (homogeneous visual patterns). *For any homogeneous visual pattern v defined on a lattice or graph Λ, let s be the visual representation (e.g. $s = \mathbf{I}$) and $\mathbf{h}(s)$ a list of sufficient feature statistics, then a pattern v is equal to a maximum set (or equivalence class), as Λ goes to infinity in the von Hove sense,*

$$\text{A pattern } v = \Omega(\mathbf{h}_o) = \{s_\Lambda : \ \mathbf{h}(s) = \mathbf{h}_o, \ \Lambda \to \infty\}. \tag{17}$$

As Λ goes to infinity and the pattern is homogeneous, the statistical fluctuations and the boundary condition effects both diminish. It makes sense to put the constraints $\mathbf{h}(s) = \mathbf{h}_o$.

In the literature, a texture pattern was first defined as a Julesz ensemble by (Zhu et al. 1999-2000). This can be easily extended to any patterns, including, generic images, texture, smooth surfaces, texton process, etc.

The connections between the three physical ensembles also reveals an important duality between a *descriptive constraints* $\mathbf{h}(s) = \mathbf{h}_o$ in the deterministic set $\Omega_{\rm mcn}(\mathbf{h}_o)$ and the parameters $\boldsymbol{\beta}$ in Gibbs model $p_{\rm Gib}$. In vision, the duality is between the image statistics $\mathbf{h}_o = (\mathbf{h}_1^{\rm obs}, ..., \mathbf{h}_K^{\rm obs})$ in equation (6) and the parameters of the descriptive models $\boldsymbol{\beta} = (\lambda_1, ..., \lambda_K)$.

The connection between set $\Omega(\mathbf{h}_o)$ and the descriptive model $p(s; \boldsymbol{\beta})$ is re-stated by the theorem below (Wu and Zhu, 1999).

Theorem 3 (Ensemble equivalence). *For visual signals $s_\Lambda \in \Omega(\mathbf{h}_o)$ on large (or infinity) lattice (or graph) Λ, then on any small lattice $\Lambda_o \subset \Lambda$, the signal s_{Λ_o} given its neighborhood $s_{\partial\Lambda_o}$ follows a descriptive model $p(s_{\Lambda_o} | s_{\partial\Lambda_o}; \boldsymbol{\beta})$.*

The duality between $\boldsymbol{\beta}$ and $\mathbf{h}_o$ is reflected by the maximum entropy constraints $E_{p(s;\boldsymbol{\beta})}[\mathbf{h}(s)] = \mathbf{h}_o$. More precisely, it is stated in the following theorem (Wu and Zhu, 1999).

Theorem 4 (Model and concept duality). *Let $p(s_\Lambda; \boldsymbol{\beta})$ be a descriptive model of a pattern v, and $\Omega_\Lambda(\mathbf{h})$ the set for pattern v, and let $\psi(\mathbf{h})$ and $\rho(\boldsymbol{\beta})$ be the entropy function and pressure defined as*

$$\psi(\mathbf{h}) = \lim_{\Lambda \to \infty} \frac{1}{|\Lambda|} \log |\Omega_\Lambda(\mathbf{h})|, \quad \text{and} \quad \rho(\boldsymbol{\beta}) = \lim_{\Lambda \to \infty} \frac{1}{|\Lambda|} \log Z(\boldsymbol{\beta}).$$

If $\mathbf{h}_o$ and $\boldsymbol{\beta}_o$ correspond to each other, then

$$\phi'(\mathbf{h}_o) = \boldsymbol{\beta}_o, \quad \text{and} \quad \rho'(\boldsymbol{\beta}_o) = \mathbf{h}_o,$$

in the absence of phase transition.

For visual patterns on finite graphs, such as a human face or a 2D shape of animal, the definition of pattern is given below.

Definition 3 (Finite patterns). *For visual pattern v on a finite lattice or graph Λ, let s be the representation, and $\mathbf{h}(s)$ its sufficient statistics, the visual concept is an ensemble governed by a maximum entropy probability $p(s; \boldsymbol{\beta})$,*

$$\text{pattern } \ \mathrm{v} = \Omega(\mathbf{h}_o) = \{(s, \ p(s : \boldsymbol{\beta})) : \ E_p[\mathbf{h}(s)] = \mathbf{h}_o\}. \tag{18}$$

Each pattern instance s is associated with a probability $p(s; \boldsymbol{\beta})$.

The ensemble is a set with each instance assigned a probability. Obviously, equation (17) is a special case of equation (18). That is, when $\Lambda \to \infty$, one homogeneous signal is enough to compute the expectation, i.e. $E_p[\mathbf{h}(s)] = \mathbf{h}(s)$. The limit of $p(s; \boldsymbol{\beta})$ is the uniform probability $p_{\text{unif}}(s; \mathbf{h}_o)$ as $\Lambda \to \infty$.

The probabilistic notion in defining finite visual signal is the root for errors in recognition, segmentation, and grouping. On any finite graph, the ensembles for two different patterns will overlap and the ability of distinguishing two patterns is limited by the Chernoff information which measures the distances of the two distributions. Some in-depth discussions on the relationship between performance bounds and models are referred to the order parameter theory (Yuille, et al.2001).

To conclude this section, we have the following equivalence for conceptualization of visual pattern.

$$\text{A visual pattern v} \longleftrightarrow \mathbf{h} \longleftrightarrow \boldsymbol{\beta} \in \Omega_K^d.$$

6 Generative modeling

In this section, we revisit the general MLE learning formulated in equations 2-4 , and review some progress in generative models of visual patterns and the integration with descriptive models.

6.1 The basic principle of generative modeling

Descriptive models are built on features and statistics extracted from the signal and use complex potential functions to characterize visual patterns. In contrast, generative models introduce hidden (latent) variables to account for the generating process of large image structures.

For simplicity of notation, we assume L-levels of hidden variables which generate image $\mathbf{I}$ in a linear order. At each level, W_i generates W_{i-1} with a dictionary (vocabulary) $\mathcal{D}_i, i = 1, ..., L$. The dictionary is a set of description, such as image bases, textons, parts, templates, lighting functions etc.

$$W_L \xrightarrow{\mathcal{D}_L} W_{L-1} \xrightarrow{\mathcal{D}_{L-1}} \cdots \xrightarrow{\mathcal{D}_2} W_1 \xrightarrow{\mathcal{D}_1} \mathbf{I}. \tag{19}$$

Let $p(W_{i-1}|W_i; \mathcal{D}_i, \boldsymbol{\beta}_{i-1})$ denote the conditional distribution for pattern W_{i-1} given W_i, with $\boldsymbol{\beta}_{i-1}$ being the parameter of the model. Then by summing over the hidden variables, we have an image model,

$$p(\mathbf{I}; \Theta) = \sum_{W_L} \cdots \sum_{W_1} p(\mathbf{I}|W_1; \mathcal{D}_1, \boldsymbol{\beta}_0) p(W_1|W_2; \mathcal{D}_1, \boldsymbol{\beta}_1) \cdots p(W_{L-1}|W_L; \mathcal{D}_L, \boldsymbol{\beta}_{L-1}). \tag{20}$$

$\Theta = (\mathcal{D}_1, ..., \mathcal{D}_L; \boldsymbol{\beta}_0, ..., \boldsymbol{\beta}_{L-1})$ are the parameters, and each conditional probability is often a descriptive model specified by $\boldsymbol{\beta}_i$.

By analogy to speech, the observable image pixels $\mathbf{I}$ is like the speech wave form. Then the first level dictionary $\mathcal{D}_1$ is like the set of *phonemes*, and $\boldsymbol{\beta}_1$ parameterizes the transition probability between phonemes. In image model, $\mathcal{D}_1$ is a set of image bases like Gabor wavelets. The second level dictionary $\mathcal{D}_2$ is like the set of *words*, each being a short sequences of phonemes in $\mathcal{D}_1$, and $\boldsymbol{\beta}_2$ parameterizes the transition probability between words. In image models, $\mathcal{D}_2$ is the set of textons. Going up the hierarchy, we need dictionaries like the *grammatic reproduction rules* for phrases and sentences in language, and probabilities for how frequently each reproduction rule is used, and so on.

A hidden variable W_i is fundamentally different from an image feature ϕ_i in descriptive models, though they may be closely related. W_i is a *random variable* that should be inferred from images, while ϕ_i is a *deterministic transform* of images.

Following the ML-estimate in equation(2), one can learn the parameters Θ in $p(\mathbf{I};\Theta)$ by EM-type algorithm, like stochastic gradients (Gu and Kong, 1998). Take derivative of the log-likelihood with respect to Θ, and set $\frac{d\partial \log p(\mathbf{I};\Theta)}{d\Theta}=0$, one gets

$$0=\sum_{W_L}\cdots\sum_{W_1}\quad[\frac{\partial \log p(\mathbf{I}|W_1;\mathcal{D}_1,\boldsymbol{\beta}_0)}{\partial(\mathcal{D}_1,\boldsymbol{\beta}_0)}+\cdots+\frac{\partial \log p(W_{L-1}|W_L;\mathcal{D}_L,\boldsymbol{\beta}_{L-1})}{\partial(\mathcal{D}_L,\boldsymbol{\beta}_{L-1})}]$$
$$\times p(W_1|\mathbf{I};\mathcal{D}_1,\boldsymbol{\beta}_0)\cdots p(W_L|W_{L-1};\mathcal{D}_L,\boldsymbol{\beta}_{L-1}) \tag{21}$$

In theory these equations can be solved with global optimum by iterating two steps (Gu and Kong, 1998):

1. The E-type step. Making inferences about the hidden variables by sampling from a sequence of posteriors,

$$W1\sim p(W_1|\mathbf{I};\mathcal{D}_1,\boldsymbol{\beta}_0),\quad\cdots,\quad W_L\sim p(W_L|W_{L-1};\mathcal{D}_L,\boldsymbol{\beta}_{L-1}). \tag{22}$$

Then we can approximate the summation (integration) by importance sampling.

2. The M-type step. Given the samples, one optimizes the parameters Θ. The learning results in Θ includes the visual dictionaries $\mathcal{D}_1,...,\mathcal{D}_L$ and the descriptive models $\boldsymbol{\beta}_0,...,\boldsymbol{\beta}_{L-1}$ that govern their spatial layouts of the hidden structures. It is beyond this review to discuss the algorithm.

6.2 Some examples of generative models

Now we review a spectrum of generative image models, starting again with a model for the $1/f$-power law.

Model G1. A generative model for the $1/f$-power law of natural images

The $1/f$-law of the Fourier amplitude in natural images was analytically modeled by a Gaussian MRF $p_{1/\mathrm{f}}$ (see eqn 8). We transform equation (8) into the Fourier domain, thus

$$p_{1/\mathrm{f}}(\mathbf{I};\beta)=\frac{1}{Z}\exp\{-\sum_{\xi,\eta}\beta(\xi^2+\eta^2)|\hat{\mathbf{I}}(\xi,\eta)|^2\}. \tag{23}$$

The Fourier bases are the independent components for the Gaussian ensemble governed by $p_{1/\mathrm{f}}$. From the above Gaussian model, one obtains a two-layer generative model (Mumford, 1995),

$$\mathbf{I}(x,y)=\sum_{\xi}\sum_{\eta}\frac{1}{2\beta(\xi^2+\eta^2)}a(\xi,\eta)e^{2\pi i\frac{x\xi+y\eta}{N}},\quad a(\xi,\eta)\sim N(0,1). \tag{24}$$

The dictionary $\mathcal{D}_1$ is the Fourier basis, and the hidden variables are the Fourier coefficients $a(\xi,\eta)$ $\forall \xi,\eta$ which are iid normal distributed. Only two parameters are used in $\boldsymbol{\beta}=(0,1)$ for specifying the normal density. Therefore,

$$W_1 = \{a(\xi,\eta) : \forall \xi,\eta\} \text{and} \mathcal{D}_1 = \{\ \mathbf{b}(\mathbf{I};\xi,\eta) = e^{2\pi i \frac{x\xi+y\eta}{N}} : \forall \xi,\eta\ \}.$$

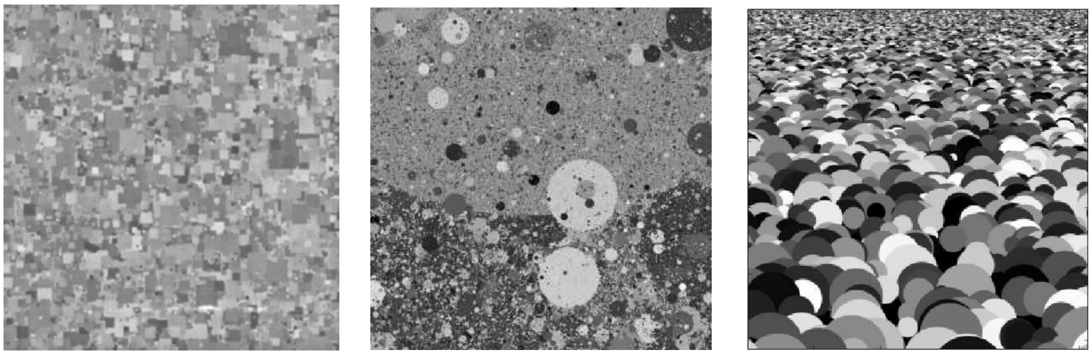
One can sample a random image $\mathbf{I} \sim p_{1/\mathrm{f}}(\mathbf{I};\beta)$ according to equation (24): 1). drawing the iid Fourier coefficients, 2). generating the synthesis image $\mathbf{I}$ by linear superposition of the Fourier bases. A result is displayed in Figure 4.b.

To our knowledge, this is the only image model whose descriptive and generative versions are analytically transferable. Such happy endings perhaps only occur in the Gaussian family!

In the literature, (Ruderman, 1997) explains the $1/f$-law by an occlusion model. It assumes that image $\mathbf{I}$ is generated by a number of independent "objects" (rectangles) of size subject to a cubic law $1/r^3$. A synthesis image is shown in Figure 12.a.

Model G2. A generative model for scale-invariant gradient histograms

The scale-invariance of gradient histograms in natural images inspired a number of research for generative models in parallel with the descriptive model p_{inv}. The objective is to search for some "laws" that governs the distribution of objects in natural scenes.



a b c

Figure 12: Synthesized images from three generative models. a). Ruderman, 1997. b) Lee et al 2000, c). Chi and S. Geman, 1998. See text for explanations.

The first is the random collage model (Lee et al. 2000), which is also called the "dead leaves" model (see Stoyan et al 1985). It assumes that an image is generated by a number of n opaque disks. Each disk is represented by hidden variables x, y, r, α for center, radius, and intensity respectively.

$$W_1 = (n, \{x_i, y_i, r_i, \alpha_i) : i = 1, 2, ..., n\}), \quad \mathcal{D}_1 = \{disk(\mathbf{I}; x, y, r) : \forall (x,y) \in \Lambda, r \in [r_{\min}, r_{\max}].\}$$

The dictionary $\mathcal{D}_1$ includes disk templates at all possible sizes and locations. Therefore let $a \oslash b$ denote that a occludes b, the image is generated by generated by

$$\mathbf{I} = disk(x_n, y_n, r_n, \alpha_n) \oslash disk(x_{n-1}, y_{n-1}, r_{n-1}, \alpha_{n-1}) \oslash \cdots \oslash disk(x_1, y_1, r_1, \alpha_1). \qquad (25)$$

Lee et al.(2000) showed that if $p(n)$ is Poisson distributed, and the disk location (x, y) and intensity α are iid uniform distributed, and the radius r_i subject to a $1/r^3$-law,

$$p(r) = c/r^3, \quad \text{for } r \in [r_{\min}, r_{\max}] \tag{26}$$

Then the generative model $p(\mathbf{I}; \Theta)$ has scale invariance gradient histograms. Figure 12.b shows a typical image sampled from this model.

The second model is studied by Chi and S.Geman (1998). This offers a beautiful 3D generative explanation. It assumes that the disks (objects) are sitting vertically on a 2D plane (the ground) facing the viewer. The sizes of the disks are iid uniformly distributed, and they proved that the 2D projected (by perspective projection) sizes of the objects then follow the $1/r^3$ law in equation (26). The locations and intensities are iid uniformly distributed like the random collage model. A typical image sampled from this model is shown in Figure 12.c. More rigorous studies and laws along this vein are in (Mumford and Gidas, 2001). These results put a reasonable explanation for the origin of scale invariance in natural images. Nevertheless these models are all biased by the object elements they choose, as they are not maximum entropy models, in comparison with $p_{\text{inv}}(\mathbf{I})$.

Model G3. Generative model for sparse coding: learning the dictionary

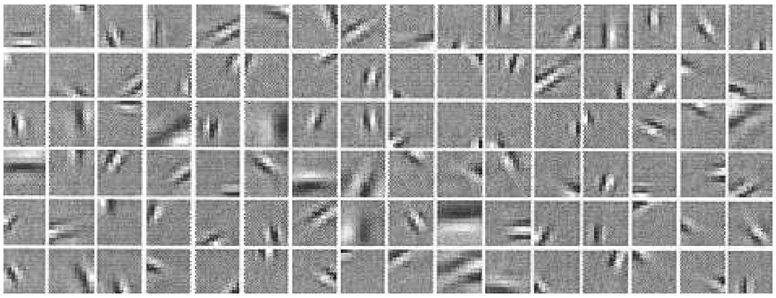


Figure 13: Some of the linear bases (dictionary) learned from natural images by (Olshausen and Field,1997)

In research stream 2 (image coding, wavelets, image pyramids, ICA etc) discussed in section (2.1), a linear additive model is widely assumed and an image is a superposition of some local image bases from a dictionary plus a Gaussian noise image $\mathbf{n}$.

$$\mathbf{I} = \sum_i^n \alpha_i \cdot \psi_{\ell_i, x_i, y_i, \tau_i, \sigma_i} + \mathbf{n}, \qquad \psi_i \in \mathcal{D}, \ \forall i. \tag{27}$$

ψ_ℓ is a base function, for example, Gabor, Laplacian of Gaussian etc. It is specified by hidden variables $x_i, y_i, \tau_i, \sigma_i$ for position, orientation, and scale. Thus a base is indexed by hidden variables $b_i = (\ell_i, \alpha_i, x_i, y_i, \tau_i, \sigma_i)$. The hidden variables and dictionary are,

$$W_1 = (n, \{b_i : i = 1, 2, ..., n\};\ \mathbf{n}), \quad \mathcal{D}_1 = \{\psi_\ell(x, y, \tau, \sigma) : \forall x, y, \tau, \sigma, \ell.\}$$

$x_i, y_i, \tau_i, \sigma_i$ are assumed iid uniformly distributed, and the coefficients $\alpha_i \sim p(\alpha)$, $\forall i$ follow an iid Laplacian or mixture of Gaussian for sparse coding,

$$p(\alpha) \sim \exp\{-|\alpha|/c\} \quad \text{or} \quad p(\alpha) = \sum_{j=1}^{2} \omega_j N(\alpha, \sigma_j). \tag{28}$$

According to the theory of generative model (section (6.1)), one can learn the dictionary from raw images in the M-step. Olshausen and Field (1995-97) used the sparse coding prior $p(\alpha)$ learned a set of $144 = 12 \times 12$ pixels bases, some of which are shown in Figure 13. Such bases capture some image structures and are believed to bear resemblance to the responses of simple cells in V1 of primates.

Model G4. A generative model for texton and texture

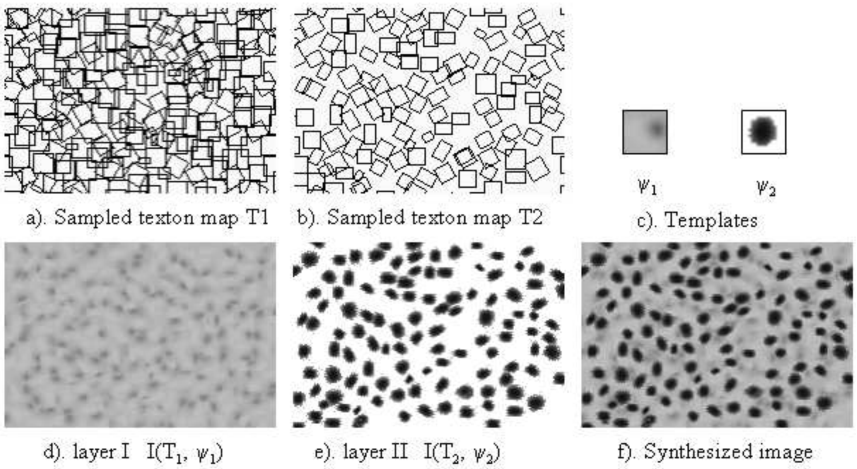


Figure 14: An example of integrating descriptive texton model and a generative model for a cheetah skin pattern. After (Guo et al 2001).

In the previous three generative models, the hidden variables are assumed to be iid distributed. Such distributions can be viewed as degenerated descriptive models. But obviously these variables and objects are not iid, and sophisticated descriptive models are needed for the spatial relationships between the image bases or objects.

The first work that integrates the descriptive and generative model was presented by (Guo et al. 2001) for texture modeling. It assumes that a texture image is generated by two levels (a foreground and a background) of hidden texton processes plus a Gaussian noise. Figure 14 shows an example of cheetah skin pattern. Figure 14.a-b shows two texton patterns $\mathbf{T}_1, \mathbf{T}_2$, which are sampled from descriptive textons models $p_{\text{txn}}(\mathbf{T}; \boldsymbol{\beta}_{o,1}, \boldsymbol{\beta}_1)$ and $p_{\text{txn}}(\mathbf{T}; \boldsymbol{\beta}_{o,2}, \boldsymbol{\beta}_2)$ respectively. The models are learned from an observed cheetah skin (raw pixel) image. Each texton is symbolically illustrated by an oriented window. Then two base functions ψ_1, ψ_2 are learned from images and shown in Figure 14.c. The two image layers are shown in Figure 14.d-e. The superposition (with occlusion) of the two layers renders the synthesized image in Figure 14.f. More examples and discussions are referred to (Guo, et al ICCV 2001).

Model G5. A generative rope model of curve processes

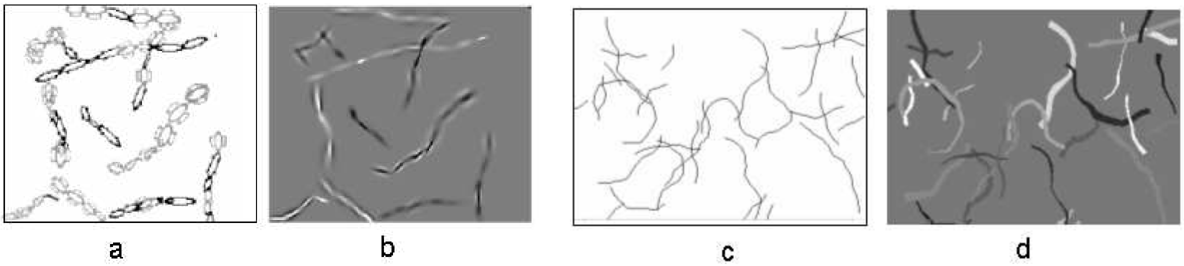


Figure 15: a-b). A rope model is a Markov chain of knots and each knot has 1-3 image bases shown by the ellipses. c-d). The smooth curve model on intensity. After (Tu and Zhu, ECCV 2002)

A three-layer generative model for curve, called a "rope model", was studied by (Tu and Zhu, ECCV 2002). The model extends the descriptive model for SNAKE and Elastica $p_{\rm snk}$ and $p_{\rm els}$ by integrating it with base and intensity representation.

Figure 15.a shows a sketch of the rope model which is a Markov chain of knots. Each knot ζ has 1-3 linear bases, for example, difference of Gaussian (DoG), and difference of offset Gaussians (DooG) at various orientations and scales

$$\begin{aligned} W_2 &= (n, \zeta_1, \zeta_2, ..., \zeta_n), \quad \text{with} \quad \zeta_i = (\alpha_{ij}, \ell_{ij}, x_{ij}, y_{ij}, \tau_{ij}, \sigma_{ij})_{j=1}^{k}, \ k \le 3. \\ W_1 &= (N, \{b_{ij} : i = 1, 2, ..., n; j = 1, ..., 3\}). \end{aligned}$$

Figure 15.b shows a number of random curves (image not pure geometry) sampled from the rope model. The image $\mathbf{I}$ is the linear sum of the bases in W_1.

This additive model is insufficient for occlusion etc. Figures 15.c-d show a occlusion type curve model. Each curve is a SNAKE/Elastica type Markov chain model with width and intensity at each point. Fig. 15.c is the sampled curve skeleton, and Fig 15.d is the image. Smoothness are assumed for both geometry, width, and intensity.

Summary. The generative models used in vision are still very preliminary, and they often assume a degenerated descriptive model for the hidden variables. To develop richer generative models, one needs to integrate generative and descriptive models.

7 Conceptualization of patterns and their parts: revisited

With generative models, we now revisit the conceptualization of visual patterns in a more general setting.

In Section (5.2), a visual pattern v with representation s is equalized to a statistical ensemble governed by a model $p(s; \boldsymbol{\beta})$ or equivalently a statistical description $\mathbf{h}_o$. In reality, the representation s is given in a supervised way and is not observable unless s is an image. Thus we need to define visual concepts based on images so that they can be learned and verified from observable data.

Following the notation is Section (6.1), we have the following definition extending from definition 3.

Definition 4 (visual pattern) *A visual pattern v is a statistical ensemble of image* $\mathbf{I}$ *governed by a generative model $p(\mathbf{I}; \Theta^v)$ with L layers,*

$$\text{pattern v} = \Omega(\Theta^v) = \{\ (\mathbf{I}, p(\mathbf{I}; \Theta^v)) : \Theta^v \in \Omega^g_K\ \},$$

where $p(\mathbf{I}; \Theta^v))$ is defined in equation (20).

In this definition, a pattern v is identified by a vector of parameters in the generative family Ω^g_K, which include the L dictionaries and L descriptive models,

$$\text{A visual pattern v} \longleftrightarrow \Theta^v = (\mathcal{D}^v_1, ..., \mathcal{D}^v_L, \beta^v_0, ..., \beta^v_{L-1}) \in \Omega^g_K$$

By analogy to speech, Θ^v defines the whole language system, say $v = English$ or $v = chinese$, and it includes all the hierarchic descriptions from waveforms to phonemes, and to sentences – both the vocabulary and models.

Therefore, the definition of many intuitive but vague concepts, such as textons, meaningful parts of shape etc, must be defined in the context of a generative model Θ. It is meaningless to talk about a texton or part without a generative image model

Definition 5 (visual vocabulary) *A visual vocabulary, such as textons, meaningful parts of shape etc. are defined as an element in the dictionaries $\mathcal{D}_i, i = 1, ..., L$ associated with the generative model of natural images $p(\mathbf{I}; \Theta)$.*

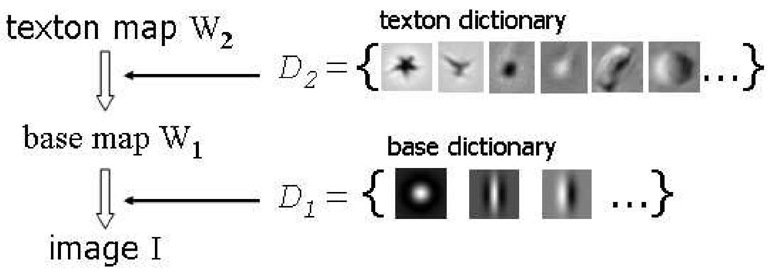


Figure 16: A three level generative image model with textons. After (Zhu, et al. ECCV 2002).

To show some recent progress, we show a three-level generative model for textons. It assumes that an image $\mathbf{I}$ is generated by a linear superposition of bases W_1 in equation (28). These bases are in turn generated by a smaller number of textons W_2. Each textons is a deformable template consists of a few bases in a graph structure. The dictionary $\mathcal{D}_1$ includes a number of base functions, such as Laplacian of Gaussian, Gabor etc. This is like the phonemes in speech. The dictionary $\mathcal{D}_2$ includes a larger number of texton templates. Each texton in $\mathcal{D}_2$ represents a small iconic object at distance, such as stars, birds, cheetah blobs, snowflakes, beans etc. It is expected that natural images have levels of vocabularies with sizes $|\mathcal{D}_1| = O(10)$ and $|\mathcal{D}_2| = O(10^3)$. These must be learned from natural images.

8 Variants of descriptive models

In this section, we review the third category of models which are two variants of descriptive models – causal MRF and pseudo-descriptive models. These variants are most popular due to their computational convenience. However people should be aware of their limitations and use them with caution.

8.1 Causal Markov Models

Let $s = (s_1, ..., s_n)$ be the representation of a pattern. As Figure 2.b illustrates, a causal Markov model imposes a partial order in the vertices and thus factorizes the joint probability into a product of conditional probabilities,

$$p_{\text{cau}}(s; \boldsymbol{\beta}) = \prod_{i=1}^{n} p(s_i \,|\, \text{parent}(s_i); \beta_i). \tag{29}$$

parent(s_i) is the set of parent vertices which point to s_i. Though the graph is directed in syntax, this is not a generative model because the variables are at the same semantic level. $p_{\text{cau}}(s)$ can be derived from the maximum entropy learning scheme in Section (4.1).

$$p^*_{\text{cau}} = \arg\max - \sum_{s} p_{\text{cau}}(s) \log p_{\text{cau}}(s).$$

Thus $p_{\text{cau}}(s; \boldsymbol{\beta})$ is a special class of descriptive model. When the dimension of $p(s_i \,|\, \text{parent}(s_i))$ is not high, (e.g. $|\text{parent}(s_i)|+1 \leq 4$), the conditional probability are often estimated by a non-parametric Parzen window.

There are many causal Markov models for texture in the 1980s and early 1990s (See Popat and Picard, 1994 and references therein). In the following, we review two pieces of interesting work appeared recently.

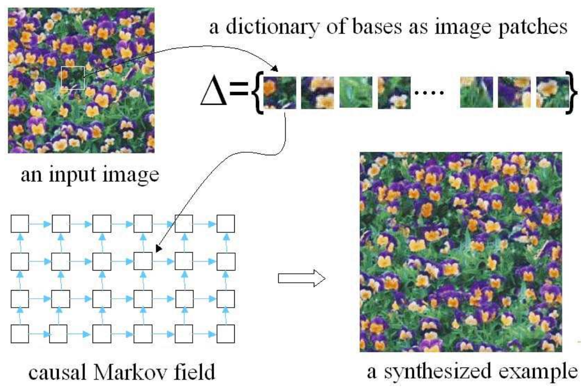
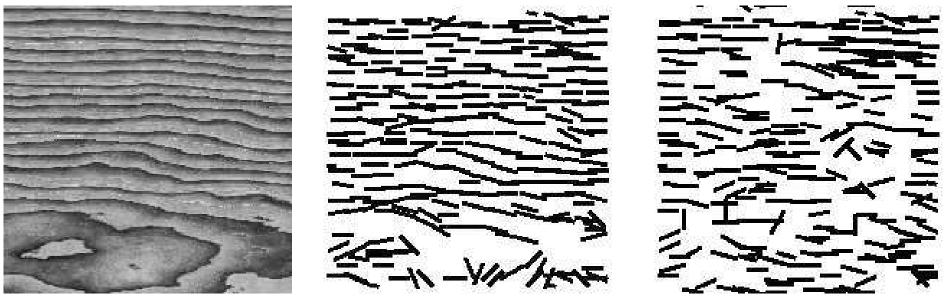


Figure 17: A causal MRF model for example-based texture synthesis (Efros et al 1999-2001, Liang et al 2001).

One is the work on example-based texture synthesis by Efros and Leung (1999), Liang et al. (2001), and Efros and Freeman (2001). Hundred of realistic textures can be synthesized by a patching technique. Figure 17 re-formulates the idea in a causal Markov model. An example texture image is first chopped into a number of image patches of a pre-defined size. These patches form a vocabulary $\mathcal{D}_1 = \Delta$ of image "bases" specific to this texture. Then a causal Markov field is set up with each element being chosen from Δ conditional on two other previous patches (left and below). The patches are pasted one by one in a linear order by sampling from a non-parametric conditional distribution. A synthesized image is shown to the lower-right side. The vocabulary Δ greatly reduces the search space and thus the causal model can be simulated extremely fast. The model is biased by the dictionary and the causality assumption.



a). input b). image sketch c). a synthesized sketch.

Figure 18: A causal Markov model for texture sketch. After (Wu et al. 2002).

Another causal Markov model was proposed by (Wu et al. ECCV 2002). Wu et.al. represent an image by a number of bases from a generic base dictionary (Log. DoG. DooG) as in sparse coding model. Each base is then symbolically represented by a line segment, as Figures 18.a-b show. This forms a base map similar to the texton (attributed point) pattern in Figure 7. Then a causal model is learned based on Figure 18.b for the base map. The graph structure is more flexible than the grid in Figure 17. A random sample is drawn from the model and shown in Figures 18.c.

8.2 Pseudo-descriptive models

While causal Markov models approximate the Gibbs distributions p_{des} and have sound probabilities p_{cau}, the second variant, called descriptive model in this paper, approximates the Julesz ensemble.

For example, the texture synthesis work by (Heeger and Bergen, 1995) and (De Bonet and Viola, 1997) belong to this family. Given an observed image $\mathbf{I}^{\text{obs}}$ on a large lattice Λ, suppose a number of K filters $F_1, F_2, ..., F_K$ are chosen, say Gabors at various scales and orientations. Convolving the filters with image $\mathbf{I}^{\text{obs}}$, one obtains a set of filter responses

$$S^{\text{obs}} = \{\, F_i^{\text{obs}}(x,y) = F_i * \mathbf{I}^{\text{obs}}(x,y) : i = 1,2,...,K, \quad (x,y) \in \Lambda \,\}.$$

Usually $K > 30$ and thus S^{obs} is a very redundant representation of $\mathbf{I}^{\text{obs}}$. In practice, to reduce the dimensionality and computation, these filter responses are organized in a pyramid representation with low frequency filters sub-sampled (see Fig.19).

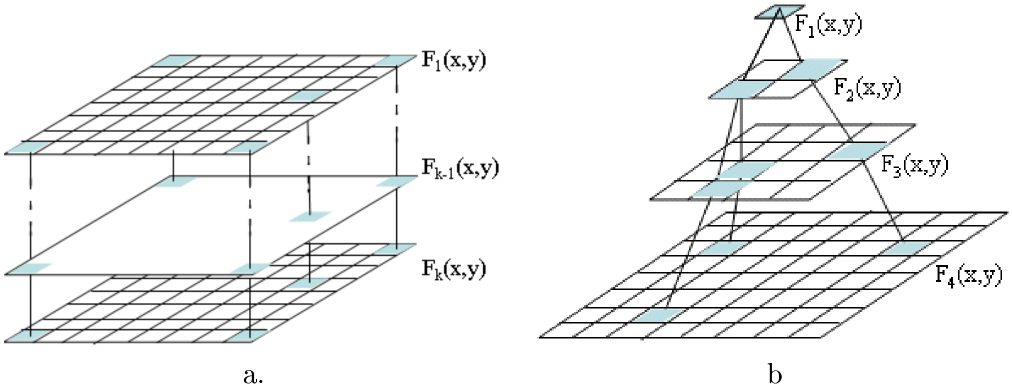


Figure 19: a. Extracting feature vectors $(F_1(x,y),...,F_K(x,y)$ for every pixels in a lattice and thus obtain $K|\Lambda|$ filter responses. b. Extracting the feature vectors in a pyramid. (Heeger and Bergen, 1995) and (De Bonet and Bergen, 1997).

Let $\mathbf{h}^{\mathrm{obs}} = \mathbf{h}(\mathbf{I}^{\mathrm{obs}}) = (\mathbf{h}_1^{\mathrm{obs}}, ..., \mathbf{h}_K^{\mathrm{obs}})$ be the K marginal histograms of the filter responses. A Julesz ensemble (or texture) is defined by $\Omega(\mathbf{h}^{\mathrm{obs}}) = \{\mathbf{I} : \mathbf{h}(\mathbf{I}) = \mathbf{h}^{\mathrm{obs}}\}$. (Heeger and Bergen, 1995) sampled the $K \cdot |\Lambda|$ filter responses independently according to $\mathbf{h}^{\mathrm{obs}}$, which is computationally very convenient. Obviously the sampled filter responses $F_i(x,y), i = 1,...,K, (x,y) \in \Lambda$ produce histograms $\mathbf{h}_o$ (or very closely), but these filter responses are inconsistent as they are sampled independently. There is no image $\mathbf{I}$ that can produce these filter responses. Usually one finds an image $\mathbf{I}$ that has least square error by pseudo-inverse. In fact this employs an image model,

$$p_{\mathrm{ps}_{\mathrm{d}}\mathrm{es}}(\mathbf{I}) \propto \exp\{-\sum_{i=1}^{K} \sum_{(x,y)\in\Lambda} (F_i^{\mathrm{syn}}(x,y) - F_i * \mathbf{I}(x,y))^2/\sigma^2\}, \quad F_i^{\mathrm{syn}}(x,y) \overset{\mathrm{iid}}{\sim} \mathbf{h}_i^{\mathrm{obs}}, \ \forall i, \forall (x,y). \quad (30)$$

Of course, the image computed by pseudo-inverse usually do not satisfy $\mathbf{h}(\mathbf{I}) = \mathbf{h}^{\mathrm{obs}}$. So we call it "pseudo-descriptive" model. The work by De Bonet and Viola (1997) was done in the same principle, but it used a K-dimensional joint histogram for $\mathbf{h}(\mathbf{I})$. As K is very high in their work (say $K = 128$), sampling the joint histogram is almost equal to shuffling the observed image.

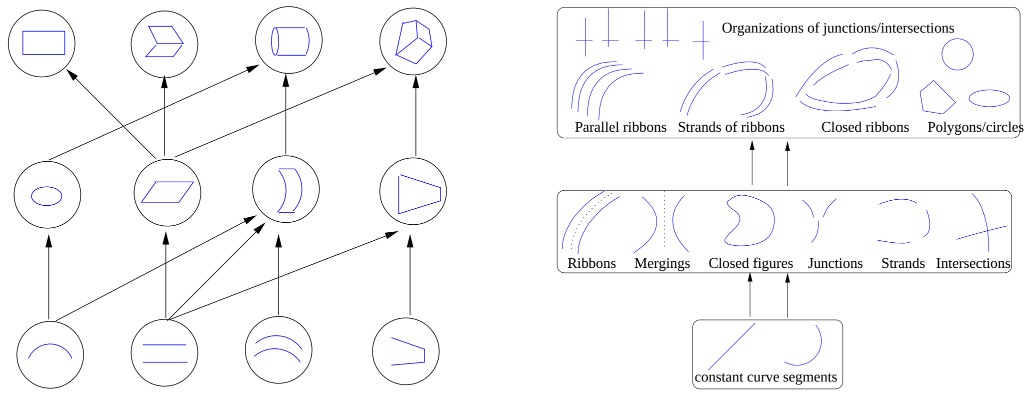
In a descriptive model or Julesz ensemble, the number of constraints in $\mathbf{h}(\mathbf{I}) = \mathbf{h}^{\mathrm{obs}}$ is much lower than the image pixels Λ. In contrast, a pseudo-descriptive model puts $K|\Lambda|$ constraints and produces to an empty set.

9 Discriminative models

Many perceptual grouping work (research stream 3) fall in category 4 – discriminative models. In this section, we briefly mention some typical work, and then focus on the theoretical connections between discriminative models to the descriptive and generative models. A good survey of grouping literature is given in (Boyer and Sarkar, 1999).

9.1 Some typical discriminative models

The objective of perceptual grouping is to compose image elements into larger and larger structures in a hierarchy. Figure 20 shows two influential work in the literature. (Dickinson et al, 1992) adopted a hierarchic Bayesian network for grouping short line and curve segments into generic object facets, and the latter are further grouped into 2D views of 3D object parts. (Sarkar and Boyer, 1994) used the Bayesian network for grouping edge elements into hierarchic structures in aerial images. More recent work is (Amir and Lindenbaum, 1999).



a b

Figure 20: Hierarchic perceptual grouping. a). After (Dickinson et al. 1992) b). After (Sarkar and Boyer, 1994)

If we represent the hierarchic representation by a linear order for ease of discussion, the grouping proceeds in the inverse order of the generative model (see equation (19). Figure 2).

$$\mathbf{I} \longrightarrow W_1 \longrightarrow W_2 \longrightarrow \cdots \longrightarrow W_L \tag{31}$$

As the grouping must be done probabilistically, both (Dickinson et al. 1992) and (Sarkar and Boyer, 1994) adopted a list of conditional probabilities in their Bayesian networks. Reformulated in the above notation, they are,

$$q(W_1|\mathbf{I}), \quad q(W_2|W_1), \quad ..., \quad q(W_L|W_{L-1}).$$

Again, we use linear order here for clarity. There may be express-ways for computing objects from edge elements directly, such as generalized Hough transform. In the literature, most of these conditional probabilities are manually estimated or calculated in a similar way to (Lowe, 1987).

9.2 The computational role of discriminative models

The discriminative models are effective and useful in vision and pattern recognition. However, there are a number of conceptual problems suggesting that they should perhaps not be considered *representational models*, instead they are *computational heuristics*. In the desk example of Figure 2, the presence of a leg may, as a piece of evident, "suggests" the presence of a desk but it does not "cause" a desk. A leg can also suggest chairs and a dozen other types of furniture which have legs. It is the desk concept that causes 4 legs and a top at various configurations in the generative model.

What is wrong with the inverted arrows in discriminative models? A key point associated with Bayes (causal, belief) networks is the idea of "explaining-away" or "lateral inhibition" in a neuroscience term. If there are multiple competing causes for a symptom, then the recognition of one cause will suppress the other causes. In a generative model, if a leg is recognized as belonging to a desk during computation, then the probability of a chair at the same location is reduced drastically. But in a discriminative model, it appears that the four legs are competing causes for the desk, then one leg should drive away the other three legs in explanation! This is not true. Without the guidance of generative model, the discriminative methods could create combinatorial explosions.

In fact, the discriminative models are approximations to the posteriors,

$$q(W_1|\mathbf{I}) \sim p(W_1|\mathbf{I};\mathcal{D}_1,\boldsymbol{\beta}_0), \quad \cdots, \quad q(W_L|W_{L-1}) \sim p(W_L|W_{L-1};\mathcal{D}_L,\boldsymbol{\beta}_{L-1}). \tag{32}$$

Like most pattern recognition methods, the approximative posteriors $q()$'s use only local deterministic features at each level for computational convenience. For example, suppose W_1 is an edge map, then it is usually assumed that $q(W_1|\mathbf{I}) = q(W_1|\Phi_1(\mathbf{I}))$ with $\Phi_1(\mathbf{I})$ being some local edge measures (Konishi, 2003). For the other levels, $q(W_{I+1}|W_i) = q(W_{i+1}|\Phi_i(W_i))$ with $\Phi_i(W_i)$ being some *compatibility functions and metrics* (Sarkar and Boyer 1994, Bienenstock et al 1997).

For ease of notation, we only consider one level of approximation: $q(W|\mathbf{I}) = q(W|\Phi(\mathbf{I})) \approx p(W|\mathbf{I};\mathcal{D},\boldsymbol{\beta})$. By using local and deterministic features, information is lost in each approximation. The amount of information loss is measured by the Kullback-Leibler divergence. Therefore, the best set of features is chosen to minimize the loss.

$$\Phi^* = \arg\min_{\Phi\in\text{Bank}} KL(p\,||\,q) = \arg\min_{\Phi\in\text{Bank}} \sum_W p(W|\mathbf{I};\mathcal{D},\boldsymbol{\beta}) \log \frac{p(W|\mathbf{I};\mathcal{D},\boldsymbol{\beta})}{q(W|\Phi(\mathbf{I}))}$$

Now we have the following theorem for what are most discriminative features.[5]

Theorem 5 *For linear features Φ, the divergence $KL(p \,||\, q)$ is equal to the mutual information between variables W and image $\mathbf{I}$ minus the mutual information between W and $\Phi(\mathbf{I})$.*

$$KL(p(W|\mathbf{I};\mathcal{D},\boldsymbol{\beta}) \,||\, q(W|\Phi(\mathbf{I}))) = MI(W,\, \mathbf{I}) - MI(W,\, \Phi(\mathbf{I})).$$

$MI(W,\mathbf{I}) = MI(W,\Phi(\mathbf{I}))$ *if and only if* $\Phi(\mathbf{I})$ *is the sufficient statistics for* W.

This theorem leads to a *maximum mutual information principle* for discriminative feature selection, and it is different from the most informative feature for descriptive models.

$$\Phi^* = \arg\max_{\Phi\in\text{Bank}} MI(W,\Phi(\mathbf{I})) = \arg\min_{\Phi\in\text{Bank}} KL(p(W|\mathbf{I};\mathcal{D},\boldsymbol{\beta}) \,||\, q(W|\Phi(\mathbf{I})))$$

[5]This proof was given in a unpublished note by Wu and Zhu. A similar conclusion was also given by a variational approach by Wolf and George (1999), who sent an unpublished manuscript to Zhu.

The main problem with the discriminative models is that they do not pool global and top-down information in inference. In our opinion, the discriminative models are *importance proposal probabilities* for sampling the true posterior and inferring the hidden variables. Thus they are crucial in computation for both Bayesian inference and for learning generative models (see the E-step in equation (22)). In both tasks, we need to draw samples from the posteriors through Markov chain Monte Carlo (MCMC) techniques. The latter need to design some proposal probabilities $q()$'s to suggest the Markov chain moves.

The convergence of MCMC critically depend on how well $q()$ approximate $p()$. This is stated in the theorem below by (Mengersen and Tweedie, 1994)

Theorem 6 *Sampling a target density $p(x)$ by independence Metropolis-Hastings algorithm with proposal probability $q(x)$. Let $P^n(x_o, y)$ be the probability of a random walk to reach point y at n steps from an initial point x_o. If there exists $\rho > 0$ such that,*

$$\frac{q(x)}{p(x)} \geq \rho, \quad \forall x.$$

then the convergence measured by a L_1 norm distance

$$||P^n(x_o, \cdot) - p|| \leq (1 - \rho)^n.$$

This theorem, though on a simple case, states the computational role of discriminative model. The idea of using discriminative models, such as edge detection, clustering, Hough transforms, are used in a data-driven Markov chain Monte Carlo (DDMCMC) framework for generic image segmentation, grouping, and recognition (Zhu, Zhang and Tu, CVPR 2000, Tu and Zhu, ICCV 2001, ECCV 2002).

10 Discussion

The modeling of visual patterns is to pursue a probability model $p()$ to estimate an ensemble frequency $f()$ in a sequence of nested probability families which integrate both descriptive and generative models. These models are adapted and augmented in four aspects.

1. Learning the parameters of the descriptive models.
2. Pursuing informative features and statistics in descriptive models.
3. Selecting address variables for the descriptive model.
4. Introducing hidden variables in the generative models.

The main challenge in modeling visual patterns is the choice of models which cannot be answered unless we understand the purposes of vision.

1. What is the ultimate goal of learning? Where does it end?

Our ultimate goal is to find the "best" generative model. Starting from the raw images, each time when we add a new layer of hidden variables, we make progress in discovering the hidden structures. At the end of this pursuit, suppose we dig out all the hidden variables, then we will have a physically-based model which is the ultimate generative model denoted by p^*_{gen}. This model cannot be further compressed and we reach the Komogorov complexity of the image ensemble.

For example, the chemical diffusion-reaction equations with a few parameters may be the most parsimonious model for rendering some textures. But obviously this is not a model used in human vision. Why didn't human vision pursue such ultimate model? This leads to the second question below.

2. How do you choose a generative model from many possible explanations?

There are two extremes of models. At one extreme, theorem 2 states that the pure descriptive model p^*_{des} on raw pixels, i.e. no hidden variables at all, can approximate the ensemble frequency $f(\mathbf{I})$ as long as we put a huge number of features statistics. At the other extreme end, we have the ultimate generative model p^*_{gen} mentioned above. In graphics, there are also a spectrum of models, ranging from image based rendering to physically-based ray tracing. Certainly our brains choose a model somewhere between p^*_{des} and p^*_{gen}.

We believe that the choice of generative models is decided by two aspects. The first is the purposes of vision for navigation, grasping not just for coding. Thus it makes little sense to justify models by a simple minimum description length principle or other statistics principles, such as AIC/BIC. The second is the computational effectiveness. It is hopeless to have a quantitative formulation for vision purposes at present. We only have some understanding on the second issue.

A descriptive model uses features $\Phi()$ which is deterministic and thus easy to compute (filtering) in a bottom-up fashion. But it is very difficult to do synthesis using features. For example, sampling the descriptive model (such as FRAME) is expensive. In contrast, the generative model uses hidden variables W which has to be inferred stochastically and thus expensive to compute (analysis). But it is easier to do top-down synthesis using the hidden variables. For the two extreme models, p^*_{des} is infeasible to sample (synthesis), and p^*_{gen} is infeasible to infer (analysis). For example it is infeasible to infer parameters of a reaction-diffusion equation from observed texture images. The choice of generative model in the brain should make both analysis and synthesis convenient. As vision can be used for many diverse purposes, there will be many models co-exist.

3. Where do features and hidden variables (i.e. visual vocabulary) come from?

The mathematical principles (minimax entropy or maximum mutual information) can choose "optimal" features and variables from pre-defined sets, but the creation of these candidate sets often come from three sources: 1). observations in human vision, such as psychology and neuroscience, thus related to purposes of vision. 2). physics models, or 3). artist models. For example, the Gabor filters, and Gestalt laws are found to be very helpful in visual modeling. At present, the visual vocabulary is still far from being enough.

This may sound ad hoc to someone who likes analytic solutions! Unfortunately, we may never be able to justify such vocabulary mathematically, just as physicists cannot explain why they have to use forces or basic particles and why there are space and time. Any elegant theory starts from some creative assumptions. In this sense, we have to accept that

"The far end of sciences is art".

Acknowledgment

The work is supported by an NSF grant IIS-00-92664 and an ONR grant N-000140-110-535. The author 'd like to thank David Mumford, Yingnian Wu, and Alan Yuille for extensive discussions which lead to the development of this paper, and thank Zhuowen Tu, Cheng-en Guo for their assistance.

References

[1] A. Amir and M. Lindenbaum, "Ground from figure discrimination", *Computer Vision and Image Understanding*, Vol. 76, No.1, pp.7-18, 1999.

[2] J.J. Atick and A.N. Redlich, "What does the retina know about natural scenes?", *Neural Computation*, 4:196-210, 1992.

[3] F. Attneave, "Some informational aspects of visual perception", *Pschological Review*, 61:183-193, 1954.

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

[5] H.B. Barlow, "Possible principles underlying the transformation of sensory messages". In *Sensory Communication*, ed. W.A. Rosenblith, pp217-234, MIT Press, Cambridge, MA, 1961.

[6] J. Besag, "Spatial interaction and the statistical analysis of lattice systems (with discussion)." *J. Royal Stat. Soc., B*, **36**, 192-236. 1973.

[7] E. Bienenstock, S. Geman, and D. Potter, "Compositionality, MDL Priors, and Object Recognition", NIPS, 1997.

[8] A. Blake and A. Zisserman. *Visual Reconstruction.* Cambridge, MA. MIT press, 1987.

[9] K.L. Boyer and S. Sarkar, "Perceptual organization in computer vision: status, challenges, and potentials", *Computer Vision and Image Understanding*, Vol. 76, no. 1, pp. 1-5, 1999.

[10] E.J. Candes, Rightlets: Theory and Applications, Ph.D thesis, Stanford University, Statistics, 1998.

[11] C.R. Carlson, "Thresholds for perceived image sharpness", *Photographic Sci. Eng.*, **22**, 69-71, 1978.

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

[13] Z.Y. Chi, "Probabilistic models for complex systems", Doctoral dissertation with Stu Geman, Division of Applied Math, Brown University, 1998.

[14] C. Chubb and M.S. Landy, "Othergonal distribution analysis: a new approach to the study of texture perception", in M.S. Landy (eds.) *Comp. Models of visual processing*, Cambridge, MA, MIT Press, 1991.

[15] R.W. Cohen, I. Gorog, and C.R. Carlson, "Image descriptors for displays", *Technical Report Cont. No. N00014-74-C-0184*, Office of Navy Research, 1975.

[16] R.R. Coifman and M.V. Wickerhauser, "Entropy based algorithms for best basis selection." *IEEE Trans. on Information Theory.*, Vol.38, pp713-718, 1992.

[17] P. Common, "Independent component analysis — a new concept?", *Signal Processing*, 36, 287-314, 1994.

[18] D. Cooper, "Maximum likelihood estimation of Markov process blob boundaries in noisy images", *IEEE Trans. on PAMI*, vol. 1, pp372-384, 1979.

[19] G. R. Cross and A. K. Jain, "Markov random field texture models", *PAMI*, **5**, 25-39, 1983.

[20] P. Dayan, G. E. Hinton, R. Neal, and R. S. Zemel, "The Helmholtz machine", *Neural Computation*, 7. 1022-1037, 1995.

[21] J. S. De Bonet and P. Viola, "A non-parametric multi-scale statistical model for natural images", *Advances in Neural Information Processing*, 10, 1997.

[22] S. Della Pietra, V. Della Pietra, and J. Lafferty, "Inducing features of random fields", *IEEE Trans. on PAMI*, vol. 19, no.4, April, 1997.

[23] N. G. Deriugin, "The power spectrum and the correlation function of the television signal", *Telecommunications*, **1** (7) 1-12, 1957.

[24] S. J. Dickinson, A. P. Pentland, and A. Rosenfeld, "From volumes to views: an approach to 3D object recognition", *CVGIP: Image Understanding*, Vol.55, No.2, pp. 130-154, March 1992.

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

[26] A. Efros and T. Leung, "Texture synthesis by non-parametric sampling", *Proc. of ICCV*, 1999.

[27] A. Efros and W. T. Freeman, "Image Quilting for Texture Synthesis and Transfer", *SIGGRAPH*, 2001.

[28] D.J. Field, "Relations between the statistics and natural images and the responses properties of cortical cells", *J. Optical Society of America*, A 42379-2394, 1987.

[29] D.J. Field, "What is the goal of sensory coding?", *Neural Computation.* 6, 559-601, 1994.

[30] B. Frey and N. Jojic, "Transformed component analysis: joint estimation of spatial transforms and image components", *Proc. of Int'l Conf on Computer Vision*, Corfu, Greece, 1999.

[31] K.S. Fu, *Syntactic Pattern Recognition*, Prentice-Hall, 1982.

[32] W.S. Geisler, J.S. Perry, B.J. Super, D.P. Gallogly, "Edge co-occurence in natural images predicts contour grouping performance", *Vision Research*, 41, pp711-724, 2001.

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

[34] J. W. Gibbs, *Elementary Principles of Statistical Mechanics.* Yale University Press, 1902.

[35] J.J. Gibson, *The Perception of the visual world*, Houghton Mifflin, Boston, 1966.

[36] U. Grenander, *Lectures in Pattern Theory I, II, and III*, Springer, 1976-1981.

[37] U. Grenander, Y. Chow, and K. M. Keenan, *Hands: A Pattern Theoretical Study of Biological Shapes*, Springer-Verlag, New York, 1991.

[38] U. Grenander and A. Srivastava, "Probability models for clutter in natural images", *IEEE trans. on PAMI.* vol. 23, no.4, 2001.

[39] M. G. Gu and F.H. Kong, "A stochastic approximation algorithm with MCMC method for incomplete data estimation problems", *Proc. of National Academy of Sciences*, USA, **95**, pp 7270-7274, 1998.

[40] C. E. Guo, S. C. Zhu, and Y. N. Wu, "Visual learning by integrating descriptive and generative methods", *ICCV*, 2001.

[41] G. Guy and G. Medioni, "Inferring global perceptual contours from local features", *Int'l J. of Comp. Vis.*, 20, 113-133, 1996.

[42] D. J. Heeger and J. R. Bergen, "Pyramid-based texture analysis/synthesis." *SigGraph*, 1995.

[43] D.W. Jacobs, "Recognizing 3D objects using 2D images", Doctoral dissertation, MIT AI lab. 1993.

[44] E. T. Jaynes, "Information theory and statistical mechanics", *Physical Review* 106, 620-630, 1957.

[45] B. Julesz, "Textons, the elements of texture perception and their interactions", *Nature*, 290, 91-97, 1981.

[46] B. Julesz, *Dislogues on Perception*, 1995.

[47] G. Kanizsa, *Organization in Vision*, Praeger, New York, 1979.

[48] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: active contour models", *Proc. of Int'l Conf. on Computer Vision*, London, 1987.

[49] D. Kersten, "Predictability and redundancy of natural images", *J. Opt. Soc. Am.*, A 4(12):2395-2400, 1987.

[50] S.M. Konishi, J.M. Coughlan, A.L. Yuille, and S.C. Zhu, "Fundamental bounds on edge detection: an information theoretic evaluation of different edge cues", *IEEE Trans. on PAMI*, vol. 25, no. 1, 2003.

[51] K. Koffka, *Principles of Gestalt Psychology*, 1935.

[52] A. Koloydenko, *Modeling natural microimage statistics*, Ph.D. Thesis, Dept. of Math and Stat., UMass, Amherst, 2000.

[53] A.B. Lee, J.G.Huang, and D.B. Mumford, "Random collage model for natural images", *Int'l J. of Computer Vision*, oct. 2000.

[54] L. Liang, Liu, Y. Xu, B.N. Guo, H.Y. Shum, "Real-time texture synthesis by patch-based sampling", MSR-TR-2001-40, March 2001. To appear in *Graphics and Its Applications.*

[55] C. Liu, S. C. Zhu, H. Y. Shum, "Learning inhomogeneous Gibbs model of face by minimax entropy", *ICCV*, 2001.

[56] L. D. Lowe, *Perceptual organization and visual recognition*, Kluwer Academic Publishers, 1985.

[57] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation", *IEEE Trans. on PAMI*, vol.11, no.7, 674-693, 1989.

[58] K.L. Mengersen and R.L. Tweedie, "Rates of convergence of the Hastings and Metropolis algorithms", *Annals of Statistics*, 24:101-121, 1994.

[59] Y. Meyer, "Principe d'incertitude, bases hilbertiennes et algebres d'operateurs", *Bourbaki Seminar*, no.662, 1985-86.

[60] Y. Meyer, "Ondelettes et Operateurs", Hermann, 1988.

[61] L. Moisan, A. Desolneux and J-M. Morel, "Meaningful Alignments" , *IJCV*, vol. 40:1, 7-23, 2000

[62] D. B. Mumford, "Elastica and Computer Vision", in C.L. Bajaj (ed.) *Algebraic Geometry and Its Applications*, Springer-Verlag, New York, 1994.

[63] D. B. Mumford and J. Shah, "Optimal Approximations of Piecewise Smooth Functions and Associated Variational Problems", *Comm. in Pure and Appl. Math.*, vol.42, 1989.

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

[65] D. B. Mumford, "The statistical decsription of visual signals", In ICIAM 95, edited by K.Kirchgassner, O.Mahrenholtz and R.Mennicken, Akademie Verlag, 1996..

[66] D. B. Mumford and B. Gidas, "Stochastic models for generic images", *Quarterly of Applied Math.*, vol. LIX, no. 1, 85-111, 2001.

[67] B. A. Olshausen and D. J. Field, "Sparse coding with an over-complete basis set: A strategy employed by V1?", *Vision Research*, 37:3311-3325, 1997.

[68] M.B. Priestley, *Spectral Analysis and Time Series*, Academic Press, London, 1981.

[69] T. Poggio, V. Torre and C. Koch, "Computational vision and regularization theory", *Nature*, vol. 317, pp 314-319, 1985.

[70] K. Popat and R. W. Picard, "Novel Cluster-Based Probability Model for Texture Synthesis, Classification, and Compression." *Proc. of SPIE Visual Comm. and Image Proc.*, Boston, MA, pp. 756-768, 1993.

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

[72] D.L. Ruderman, "The statistics of natural images", *Network: Computation in Neural Systems*, **5**, 517-548, 1994.

[73] D.L. Ruderman, "Origins of scaling in natural images", *Vision Research*, 37:3385-3398, December, 1997.

[74] S. Sarkar and K. L. Boyer, "Integration, inference, and management of spatial information using Bayesian networks: perceptual organization", *IEEE Trans. on PAMI*, vol. 15, No. 3, 1993.

[75] S. Sarkar and K. L. Boyer, *Computing Perceptual Organization in Computer Vision*, World Scientific Pub. Co. Singapore, 1994.

[76] C. Shannon, "A mathematical theory of communication", *Bell System Tech. Journal*, **27**, 1948.

[77] E.P. Simoncelli, W.T. Freeman, E.H. Adelson, D.J. Heeger, "Shiftable multiscale transforms", *IEEE Trans. on Info. Theory*, 38(2): 587-607, 1992.

[78] E.P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation", *Annual Rev. Neurosci.*, 24:1193-1216, 2001.

[79] B.J. Smith, "Perceptual organization in a rnadom stimulus", in *Human and Machine Vision*, Eds. A. Rosenfeld, Academic Press, San Diego, 1986.

[80] D. Stoyan, W. S. Kendall, J. Mecke, *Stochastic Geometry and its Applications*, 1985.

[81] D. Terzopoulos, "Multilevel computational process for visual surface reconstruction", *Computer Vision, Graphics, and Image Processing*, 24, 52-96, 1983.

[82] Z.W. Tu, S.C. Zhu, "Image segmentation by Data Driven Markov chain Monte Carlo", *IEEE Trans. on PAMI* vol 24, no.5, May 2002.

[83] Z.W. Tu and S.C. Zhu, "Parsing images into region and curve processes", *Proc. of ECCV*, Copenhagen, Denmark, 2002.

[84] J.H. Van Hateren and D.L. Ruderman, "Independent component analysis of natural image sequences yields spatiotemproal filters similar to simple cells in primary visual cortex", *Proc. Royal Soc. of London.* 265, 1998.

[85] L. R. Williams and D. W. Jacobs, "Stochastic completion fields: a neural model of illusory contour shape and salience", *Neural Computation*, 9, 837-858, 1997.

[86] D. R. Wolf and E.I. George, "Maximally informative statistics", Unpublished manuscript, 1999.

[87] Y. N. Wu and S. C. Zhu "Equivalence of Julesz and Gibbs Ensembles", *Proc. of ICCV*, Corfu, Greece, 1999.

[88] Y.N. Wu, S.C. Zhu, and C.E. Guo, "Statistical modeling of image sketch", Submitted to ECCV, 2002.

[89] J.S. Yedidia, W.T. Freeman and Y. Weiss, "Generalized belief propagation", TR-2000-26, Mitsubishi Elec. Res. Lab., 2000

[90] A. L. Yuille, "Deformable templates for face recognition", *J. of Cognitive Neuroscience*, vol.3 (1), 1991.

[91] A. L. Yuille, J. M. Coughlan, Y. N. Wu, and S. C. Zhu, "Order Parameter for Detecting Target Curves in Images: How Does High Level Knowledge Helps?", *Int'l Journal of Computer Vision*, vol. 41, No. 1/2, pp.9-33, 2001.

[92] A.L. Yuille, "CCCP Algorithms to Minimize the Bethe and Kikuchi Free Energies: Convergent Alternatives to Belief Propagation". *Neural Computation*, 2001.

[93] S. C. Zhu, Y. N. Wu and D. B. Mumford, "Minimax entropy principle and its application to texture modeling", *Neural Computation* Vol. 9, no 8, Nov. 1997.

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

[95] S.C. Zhu, Y. N. Wu and D. B. Mumford, "Filters, Random fields, And Maximum Entropy (FRAME): towards a unified theory for texture modeling", *Int'l Journal of Computer Vision*, 27(2) 1-20, 1998. (First appeared in CVPR96).

[96] S. C. Zhu, "Embedding Gestalt laws in Markov random fields", *IEEE Trans. on PAMI.* vol. 21, no.11, 1999.

[97] S.C. Zhu, X.W. Liu, and Y.N. Wu, "Exploring Julesz Texture Ensemble by Effective Markov Chain Monte Carlo", *IEEE Trans. PAMI*, vol. 22, no.6, 2000.

[98] S. C. Zhu, R. Zhang, and Z. W. Tu, "Integrating top-down/bottom-up for object recognition by DDMCMC", *Proc. of CVPR*, Hilton Head, SC, 2000.

[99] S.C. Zhu, C.E. Guo, Y.N. Wu, and Y.Z. Wang, "What are textons", *Proc. of ECCV*, Copenhagen, Denmark, 2002. (a long version is submitted to IJCV)