The Statistical Description of Visual Signals

David Mumford

Abstract

Bayesian statistical methods are being successfully applied in speech recognition and language parsing, computer vision (i.e. image analysis and object recognition) and in medical expert systems. All of these are instances of Grenander's general conception of "Pattern Theory", the statistical analysis of patterned but noisy and distorted signals produced by the world. To apply these ideas to a class of signals, we need to construct probability models for the observed random variables and for the unobserved world variables which caused the signal, and we also need algorithms for inferring high probability values of the unobserved variables. This talk will introduce a series of such models for visual signals which incorporate successively deeper layers of unobserved variables. Model 0 involves only the observed signal and is the basic scale-invariant Gaussian process model. Model 1 introduces local feature variables such as line processes and it belongs to the class of Markov random field models. Model 2 introduces variables describing surfaces, subsets of the domain of the image and leads to the use of stochastic grammar formalisms. This class of models is the natural stage at which the three-dimensional structure of the world producing the signal is made explicit. Finally model 3 introduces templates for learned classes of objects, which must be matched to the observed signal by pointers, random variables whose values are addresses. These are examples of what I call "mixed Markov models" which I propose as the basic tool in object recognition.

0 Introduction

The study of visual signals, commonly referred to as *images*, and of the process of extracting meaning from them has traditionally been studied by two quite different groups. The first group consists of the psychophysicists and neurobiologists, going back to the great German psychophysicist von Helmholtz whose monumental work [HvH] started the whole field. This group asks how animals and man in particular can 'see', how they can use the pattern of light striking the retina to acquire and construct a mental representation of the world in front of them. The second group consists of the engineers who sought computer algorithms for such tasks as the grasping of objects seen through a video camera by a robot, the automatic navigation of vehicles without human drivers and the automatic reading medical scans and X-rays. David Marr was one of the most influential voices in bringing these groups together. In his book [Ma], he described what he called the *theory of the computation*, a level at which there was one problem of vision for animals and computers. It might be solved by different algorithms and certainly by using different implementations in these different classes of agents, but one could analyze the components of the problem and its computational complexity in a unified way.

As a mathematician, the issue remains however – what sort of a theory is it? For example, the AI (artificial intelligence) school has proposed studying the problem of vision, as one of many cognitive problems, using logic-like languages. They would transcribe into formal logic or into prolog databases the facts of the physics of light, the shapes of the objects of the world and of how these interact to produce observed images. They propose further to develop heuristics for efficiently searching the combinatorially explosive tree of combinations of these facts to arrive at a high-level scene description compatible with the observed image.

Ulf Grenander [Gre1], [Gre2] however pioneered a second approach based on statistics. From his perspective, the problem is to learn from extensive experience the statistics of images and of the objects represented in them and to find fast algorithms for the statistical estimation of the random variables not directly observed (such as the distance to and the identity of the objects being viewed), given those which are observed (the raw retinal or video signal). In this estimation problem, the Bayesian approach – of combining learned priors on the unobserved variables with an imaging model – has been the dominant approach. This statistical approach to vision has been gaining adherents though it is by no means universally accepted. Let me note briefly how similar trends have grown in related fields:

- In speech recognition, the Bayesian statistical theory of *hidden Markov models* and the asociated *EM algorithm* for learning the model parameters have totally dominated the field,
- In control theory, the Bayesian statistical tool of the *Kalman filter* is the central technique for dealing with noise and uncertainty,
- In AI itself, statistical theories have grown in importance in medical expert systems (see work of Pearl, Lauritzen and Spiegelhalter, e.g. [Pe])

and in the so-called PAC learning models ('probably approximately correct') of Valiant, e.g. [K-V].

This article will present one way to codify the statistical approach to vision by describing *a series of classes of probability models in which successively deeper layers of unobserved variables are incorporated.* The inspiration for describing the various stages in the vision computation in this way came from trying to understand the analogies between vision and speech/language.

To make clearer what we mean by a series of probability models which successively approximate a class of real signals, I want to end this introduction by giving samples from seven successively more refined models of English, developed in Shannon's early work on information theory [Sh]. While still 'low-level', i.e. there is no syntax nor semantics in these models, they make the point that statistics alone does capture a great deal of the structure of language. First, here is a random sequence of English letters (plus space) – the linguistic analog of white noise signal:

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGXYD QPAAMKBZAACIBZLHJQD

Second, we sample from a model in which the individual letter frequencies are those of English:

OCRO HLI RGWR NMIELWIS EU LL NBBESEBYA TH EEI ALHENHTTPA OO BTTV

Third, we sample from a model in which the letter *pairs* have their correct frequency – i.e. you compile a table of probabilities of the 27^2 events $x_i x_j$ in representative samples of English speech and prose and make a string by choosing each new letter using the conditional probability of its occurence following the previously chosen letter:

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE FUSO TIZIN ANDY TOBE SEACE CTISBE

In the same way, here are strings chosen randomly with the correct letter *triple* frequencies:

IN NO IST LAY WHEY CRATICT FROURE BERS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE

and *4-tuple* frequencies:

THE GENERATED JOB PROVIDUAL BETTER TRAND THE DISPLAYED CODE ABOVERY UPONDULTS WELL THE CODERST IN THESTICAL IT TO HOCK BOTHE

Modeling English further, we incorporate the lexicon and make strings using only valid English words with their correct frequency:

```
REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT
NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME
```

Finally, here is a string with correct word pair frequencies:

```
THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE
CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE
```

These models give a sense of steady convergence to the true probabilities in English speech or prose. What are the analogs of these random strings in vision? As we explore this, we will refer back to these and other descriptions of speech/language for comparison with each class of visual models.

1 Model 0: The Scale-invariant Gaussian Process

We begin by fixing notation. By an *image*, in the simplest case, we shall mean a rectangular array $\{I(i,j) | 1 \leq i \leq N, 1 \leq j \leq M\}$ of positive real numbers. These refer to the light intensity recorded at a rectangular grid of receptors in either a TV camera or the retina. The sample points (i,j) are called the *pixels* of the image. Of course, this may be generalized in many ways: the pixels may not be spaced on a rectangular grid (in fact, the retina uses an approximately hexagonal array in the fovea) nor even on a regular grid at all; the incident light may be sampled by frequency leading to a vector of color values $\{\vec{I}(i,j)\}$ rather than a scalar brightness; or one may pass to a continuum limit $I(x,y)$ of pixels or even $I(x,y,\lambda)$, λ being wavelength. Sticking with the simplest case, however, the main object of study is a probability distribution:

$$p(\cdots, I(i,j), \cdots) \prod_{i.j} dI(i,j)$$

on NM-dimensional space. This should not be taken too literally: p is meant to capture the statistics of what an average agent in the world 'sees' as it moves around in the world. Of course this varies from agent to agent – the visual world of a mouse living in the forest is quite different from that of a Professor working in a city – but one seeks a class of such probability models with many parameters which allow an agent to learn its environment and to capture its regularities. One can also ask for models using infra-red or radar images which will differ even more radically.

The simplest probability distributions in high dimensional spaces are the Gaussian ones and this is what model 0 is. The general form of such a model would be:

$$p(I) = \frac{1}{Z} e^{-\sum a_{ij,kl} I(i,j) I(k,l)}$$

Here and in the rest of this paper, Z stands for whatever normalizing constant is needed to make the various functions p into probabilities.

Now if we consider the image I to be a function on the torus $\mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/M\mathbb{Z}$ rather than the rectangular grid, then we can ask that an image I and a translate $I'(i,j) = I(i+i_0, j+j_0)$ of I by some (i_0, j_0) should be equally likely, i.e. that p is translation-independent. If we let $\hat{I}(\xi, \eta)$ be the discrete Fourier transform of $I(i,j)$, the quadratic form a appearing in a translation-invariant Gaussian distribution is diagonalized in the new variables $\hat{I}(\xi,\eta)$, so that

$$p(I) = \frac{1}{Z} e^{-\sum c_{\xi,\eta} |\hat{I}(\xi,\eta)|^2}$$

where $c_{\xi,\eta}^{-1}$ represents the expected power of I in the spatial frequency band (ξ, η).

However visual signals have a fundamental property not shared by other types of sensory signals. *They have no distinguished scale.* This means that, if we now model images as functions $I(x,y)$ of continuous variables so as to allow all scales, then an image $I(x,y)$ and a rescaled image $I(\sigma x, \sigma y)$ are equally likely. The reason for this is that when an observer moves closer or further from some scene, then the image is rescaled: when the observer is closer, the image is enlarged by some factor σ, and when further away, the image is reduced by some factor σ. Since the distance of the observer is not fixed by anything, neither is the scale. This is not true of touch, because the the finger must actually contact the sensed object, so the object's size on the tactile array of sensors is always the same. And in audition, constants like the frequency of the vocal cords set a fixed scale on the time axis relative to which all other sound durations can be calibrated. If we go further and assume rotation invariance of the probability distribution p, this leads to model 0, a Gaussian model for images I, unique up to one parameter β:

$$p_0(I) = \frac{1}{Z} e^{-\beta \iint (\xi^2+\eta^2) |\hat{I}(\xi,\eta)|^2 d\xi d\eta}.$$

This expression poses several questions: a) is it well-defined and b) why is this distribution scale-invariant? Accepting that it is well defined in some sense, the simplest way to argue formally that it is scale-invariant is to note that it makes $|\hat{I}(\xi,\eta)|$ into independent normally distributed variables with variances

$1/2\beta(\xi^2+\eta^2)$, hence if $A(r_1,r_2)$ is the annulus in the (ξ,η)-plane with inner radius r_1 and outer radius r_2, then

$$\begin{aligned}\iint_{A(r_1,r_2)} E(|\hat{I}(\xi,\eta)|^2)d\xi d\eta &= \iint_{A(r_1,r_2)} \frac{d\xi d\eta}{2\beta(\xi^2+\eta^2)}\\ &= \frac{\pi}{\beta}\log(\frac{r_2}{r_1})\end{aligned}$$

Thus the expected amount of power in a spatial frequency band depends only on the ratio of the high and low frequencies, not on the frequencies themselves. On the other hand, letting the high frequency cutoff go to infinity, we find an infinite amount of power, hence the typical samples from this probability distribution cannot be continuous, or even locally L^2.

The stochastic process formally defined by the above probability distribution is well known to physicists but it is a process which is not supported on any space of measurable functions: rather its sample paths must be taken as distributions [R-C-L]. For instance, its samples on a torus are readily constructed as random fourier *series*

$$I(x,y)=\sum_{\xi\in\mathbb{Z}}\sum_{\eta\in\mathbb{Z}}\frac{1}{\sqrt{2\beta(\xi^2+\eta^2)}}a_{\xi,\eta}e^{2\pi i(x\xi+y\eta)}$$

where $a_{\xi,\eta}$ are an independent normal sequence with variance 1, mean 0 and $\bar{a}_{\xi,\eta}=a_{-\xi,-\eta}$. Such series 'barely' miss being measurable functions. On the other hand, a little reflection shows that we wouldn't expect sample scale-invariant images to be functions! Imagine you had the X-ray vision of superman and could see all the warts on everyone's face and all the mites crawling on a leaf, etc. By the laws of reflectance, these would cause black and white fluctuations in the image of the same size at arbitrarily small scale. Even on a macroscopic scale, it is well known to photographers that the visual world is *cluttered* with objects of every size so that good photographs must be carefully composed to emphasize the composition on one scale. This clutter is a basic problem for all computer vision algorithms too and will come up again in this article.

The Fourier series in the last formula enables one to construct readily by computer random samples from model 0. One of these is shown in figure 1. Note that it is quite reminiscent of many fractal everyday objects such as clouds, that one can 'see' in it various shapes by the use of imagination because it has lots of structure. By contrast, white noise is quite boring and featureless.

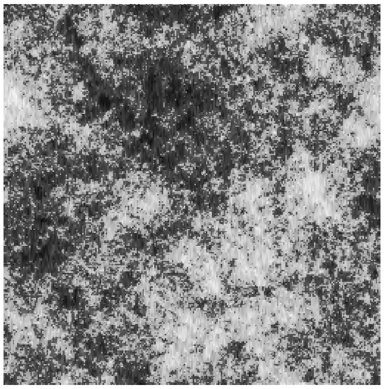

Figure 1: A sample from Model 0: an image with power $\sim 1/f^2$.

Nonetheless, it is certainly not a typical image of the world. In the linguistic analogy, model 0 is like the model of English strings in which the letter frequencies are correct.

Before leaving this model, however, it is helpful to use the inverse Fourier transform and express model 0 in terms of the original image I. By the rules for the Fourier transform of the derivative, we see the beautiful fact that:

$$p_0(I) = \frac{1}{Z} e^{-\beta \iint \|\nabla I\|^2 dx dy}.$$

Note from this that the dimension of β is 1/intensity^2, (which also followed from the expression of expected power in an annulus) and does not involve distance (confirming the scale-invariance of the model). A discrete form of this probability is obtained by replacing the gradient by sums over *adjacent* pairs of pixels:

$$p_0(I) = \frac{1}{Z} e^{-\beta \sum_{p,q \text{ adj.}} (I(p)-I(q))^2}$$

2 Model 1: Local Features using Markov Random Fields

The natural way to improve model 0 of images is to model the co-occurence of gray levels at adjacent pixels: i.e. if p, q are adjacent pixels, then one should attempt to model the so-called co-occurence statistics for the pair of intensity values $(I(p), I(q)$. The most obvious thing that happens in images is that there are *edges*. These are sharp discontinuities of image intensity primarily caused by pixel p being part of the surface of one object and pixel q being part of another, e.g. one on the foreground, the other on the background. There are also edges caused by surface markings, by abrupt changes in the surface normal (folds) and by other illumination effects such as shadows and highlights. It is clear that figure 1 lacks sharp edges.

A simple way to increase the probability of sudden intensity changes is to replace the squared term $(I(p)-I(q))^2$ by a *robust* variant $\psi(I(p)-I(q))$, where $\psi(x)$ is a function which approximately x^2 for x small, but which approaches an upper bound for $|x|$ large. Examples would be $x^2/(1+x^2)$, $\tanh^2(x)$, $(1-e^{-x^2})$. Using such a ψ, we define model 1 via:

$$p_1(I) = \frac{1}{Z} e^{-\beta \sum_{p,q \text{ adj.}} \psi(I(p)-I(q))}.$$

There is a second very natural way in which model 1 arises, with a particular ψ. This is by introducing an auxiliary set of random variables, the *line process*. We imagine the rectangular grid of pixels as being the vertices of a graph with edges linking each pixel to its 4 horizontally and vertically adjacent neighbors. A line process is a function ℓ on the edges of this graph whose values are 0 or 1:

$$\ell : \{\text{prs of adj pixels } p, q\} \longrightarrow \{0,1\}.$$

The assertion $\ell(p,q) = 0$ means the bond between pixels p and q is intact, so that $I(p)$ and $I(q)$ try to be equal, while the assertion $\ell(p,q) = 1$ means the bond is broken and $I(p)$ and $I(q)$ are totally independent of each other. This may be expressed by the formula:

$$p_1'(I,\ell) = \frac{1}{Z} e^{-\sum_{p,q \text{ adj.}} \left[\beta (I(p)-I(q))^2 (1-\ell(p,q)) + \nu \ell(p,q)\right]}.$$

What is quite remarkable is that model 1 with a suitable choice of ψ is just the *marginal probability distribution on I from this* $p_1'(I,\ell)$ [G-Y]! More precisely:

$$\sum_{\text{all possible } \ell} p_1'(I,\ell) = \frac{1}{Z} e^{-\sum_{p,q \text{ adj.}} \psi(I(p)-I(q))}.$$

where

$$\psi(x) = \log(e^{-\beta x^2} + e^{-\nu})^{-1}.$$

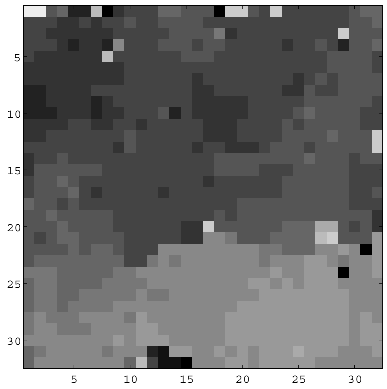

Figure 2: A sample from Model 1: a local part of an image with line processes.

It is not simple to sample from model 1, but we have used simulated annealing – hopefully with long enough time – and show in figure 2 the kind of sample we find. Unlike model 0, model 1 has a scale parameter, and the figure should be considered as a close-up of some scene in which two objects are visible.

Having broken the scale-invariance of model 0, this means that images sampled from model 1 should be regarded not as real world images but as simplifications of real images in which clutter on smaller scales has been rejected. The observed image should be regarded as a sample from model 1 plus fine detail. The simplest way to do this is to make the artifical assumption that this fine detail is white noise. This leads to the full version of model 1 in which there are three random processes: the actually observed image I, the simplified version without clutter which we now call the *cartoon* J and the line process ℓ. The full probability distribution is now:

$$p_1''(I,J,\ell) = \frac{1}{Z} e^{-\sum_p (I(p)-J(p))^2/2\sigma^2 - \sum_{p,q \text{ adj.}} \left[\beta(J(p)-J(q))^2(1-\ell(p,q))+\nu\ell(p,q)\right]}.$$

This model was introduced independently by S. and D. Geman [G-G] and by A. Blake and A. Zisserman [B-Z]. This model has been used together with Bayes's theorem to find segmentations of an image. One assumes I is given and uses the conditional distribution induced by p_1'' on the remaining variables J and ℓ in order to find probable segmentations of the image. One should not expect that such an elementary model, which still does not incorporate very much knowledge of the world, is going to find the correct segmentation of the

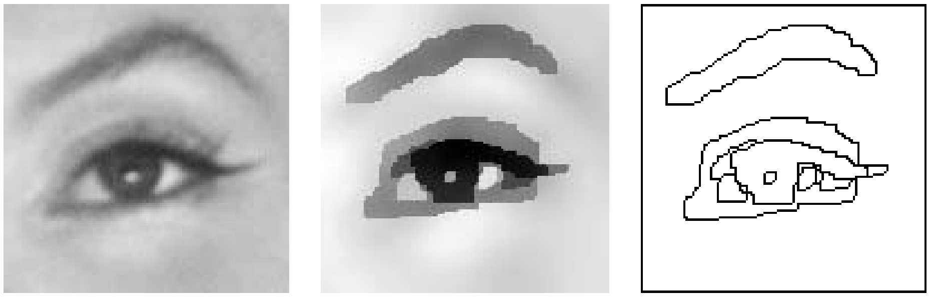

Figure 3: Image segmentation by Model 1. Left: an image I of an eye; center: the cartoon J; right: the line process ℓ.

image into multiple objects, but one does expect that the correct segmentation has relatively high probability. Figure 3 shows a typical application of this model in this way. The figure shows a real image of an eye, an estimate for the most probable cartoon J and the line process ℓ from this model. (Because this figure was generated by an algorithm approximating the most probable J, ℓ, the result is probably not optimal.)

A striking aspect of this approach is that, while the prior model on cartoons $\{J, \ell\}$ is very crude and while the imaging model $I = J + (\textit{white noise})$ is also crude, the results are rather reasonable. It seems as though the deficiencies in the prior model and the imaging model are of different sorts and the strengths of the each model help make up for the weaknesses of the other. Looking again at language analogies, the same phenomenon was found by Jelinek's group at IBM doing machine translation via similarly crude statistical models. They used word triple (called *trigram*) statistics to model English language strings; and using a corpus of 2.2 million French/English sentence pairs supplied by the Canadian parliament, they built a statistical English to French dictionary (example: `answer` becomes 44% of the time the noun `réponse`, 23% of the time the infinitive verb `repondre`, 7% of the time is omitted in the corresponding French sentence, etc.). For each French sentence F, they computed the English string E maximizing the product $p(E) \cdot p(F|E)$. In spite of the obvious deficiencies of both probability models and the total absence of any grammar rules, their translations were 45%-60% correct! Again, it seems that the strengths of each model compensate for the weaknesses of the other.

Going back to our table in the Introduction, model 1 is the most natural vision analog of the English string model which incorporates letter pair frequencies. It can only deal with homogeneous regions in an image with smoothly varying intensity and sudden jumps in intensity between such regions. But the real world is made up of many *textured* surfaces in which the local image is not homogeneous but have particular types of local statistics. Modeling these is analogous to looking for higher order letter frequency statistics. We need to model higher order co-occurence statistics for intensities at local clusters of pixels. There is no space to elaborate the various theories in this direction, especially because texture has not proved easy to describe mathematically. Instead I want to describe the basic mathematical formalism for the class of models of this type. The basic idea is to introduce *local feature descriptors* which respond when a certain pattern or texture is present and to try to group sets of pixels, or regions, where the same local feature descriptors are active.

The appropriate formalism for this is a *Markov random field.* In the generality which we need, we assume that the random variables in our model – the observed image and all the auxiliary local feature variables – form the vertices of a graph. We write these variables as $\{X_v\}_{v\in V}$, where V is the set of vertices. The edges of the graph are supposed to represent variables which have a direct influence on each other. A Markov Random Field is a probability space with these random variables with the following Markov property: when some subset of them $\{X_w\}_{w\in V_0}$ are fixed, and when v_1, v_2 are two vertices which cannot be joined by any path not containing a vertex in V_0, then X_{v_1} and X_{v_2} are conditionally independent. By the Hammersley-Clifford theorem, this is equivalent to the probabilities being given by a Gibbs formula:

$$p(\{X_v\}) = \frac{1}{Z} e^{-\sum_{\text{cliques } C} E_C(\{X_v\}_{v\in C}}.$$

Here a *clique* is a subset of the vertices of a graph all of whose vertices are joined by edges, and there is one term E_C in the exponential for each such clique.

The graph of the Markov random field used in model 1 is shown in figure 4. Many Markov random field models have been used to model textured images: they include further auxiliary vertices and edges linking nearby pixels over larger local neighborhoods. Figure 5 shows the output of an algorithm from the work of Zhu, Yuille and Lee [Z-L-Y]. Simple local statistics are used together with a 'region-growing' algorithm to find a high probability segmentation of the scene.

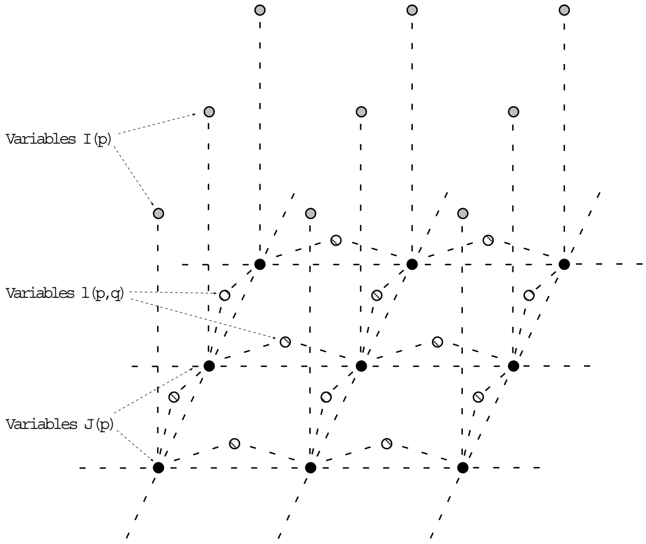

Figure 4: The graph for the Markov random field for segmenting an image I via a line process ℓ and a cartoon J.

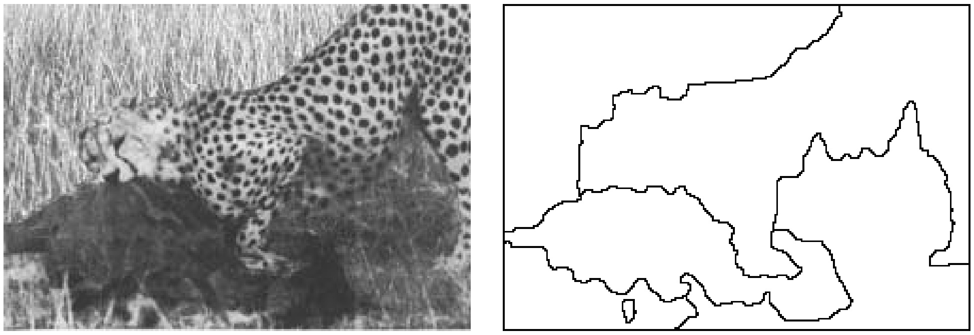

Figure 5: Segmenting a scene on the plains of Africa by local texture statistics modeled by a Markov random field.

3 Model 2: Surface Descriptors using Stochastic Grammars

While local patterns and structures arising from a Markov random field can create images with the local 'look and feel' of a real world image, there is much

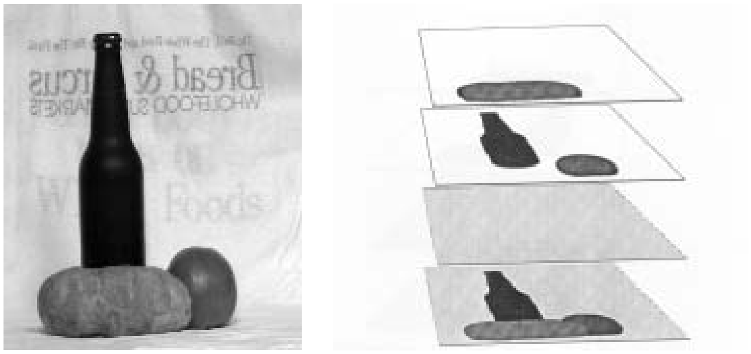

Figure 6: Left: the observed image; right: its representation as a set of three layers of increasing depth.

more to be captured. The next stage is to make explicit the larger structures which arise from the three-dimensional geometry of the world, especially objects in the world and the parts of their surface visible in the image. This is analogous to identifying in speech or language the larger groups of letters, first words, then grammatical phrases and clauses.

I would like to make this clear from the simplest example, which is what Nitzberg and I have called the *2.1D sketch* [N-M]. Note that model 1, with its line process implicitly defines a decomposition of the set of pixels into regions. Namely, consider the set of pixels to be joined only by those edges where $\ell(p, q) = 0$, those edges which are intact after the line process breaks the rest. Let $\{R_i\}, 1 \leq i \leq n$ be the connected components of this graph. These are the connected regions resulting from cutting apart the image domain along the edges ℓ. In some cases, these may be the objects present in the scene but it may also happen that an object appears in several places, being partly occluded by a nearer object. Moreover, each edge has a 'belongingness' as Nakayama calls it: it is the edge of one of its two sides, that of the nearer object and lies in an accidental position on the farther one. There is a strong local cue for this three-dimensional structure. When one edge vanishes behind another edge, the set of edges forms a so-called *T-junction*, and the two objects seen on the stem of the 'T' must be further than the object above the top of the 'T'. An example is shown in figure 6: the observed image consists in a potato in front of an orange and a beer bottle at an intermediate distance all against a background consisting of a cardboard box with faint letters on it. Mentally, you represent this scene something like the diagram showing the three layers separately. Note the T-junctions where the beer bottle and the orange disappear behind the potato. Mathematically, we define a 2.1D sketch to be an ordered sequence $\{R_i\}_{1\leq i\leq n}$, of subsets of the image domain which can overlap in any ways. We assume the R_i's are objects projected onto the image domain and that R_i is nearer than R_j whenever $i < j$. In particlar, R_n is the background, which we assume to be the whole domain. Thus $R'_i = R_i - \cup_{j<i}(R_i \cap R_j)$ will be the visible part of object i. In the figure, a 2.1D sketch with 3 regions R_1, R_2 and R_3 has been computed as the most probable values of the 2.1D variables $\{R_i\}$ in a precise probablity model which relies heavily on the T-junctions.

Before giving details of the model, I want to note that such layered representations also arise from the analysis of binocular stereoscopic image pairs and from temporal image sequences. In figure 7, we show an example from the work of Wang and Adelson [W-A]: on the right, you see three frames from a movie with thirty images; on the left you see the decomposition of the scene into three layers, the foreground tree, the intermediate flower bed and the background house and trees. In fact, human infants are born able to segment visual signals into layers on the basis of relative motion, (using, presumably, to the brainstem structure called the *superior colliculus*). They develop the ability to segment into layers using stereoscopic vision later and the ability to perceive layers in single images last.

I claim that underlying the 2.1D sketch is an extremely simple stochastic grammar. Recall that a stochastic grammar is described by giving a set of symbols, called non-terminals and another set, the terminals, and a set of production rules of the form $A \rightarrow B_1B_2\cdots B_k$. For example, a simple class of sentences may be generated by the simple rules:

$$\begin{array}{llllll}
\mathbf{S} & \longrightarrow & \mathbf{NP} & \mathbf{VP}, & & \text{Prob.} = 1 \\
\mathbf{NP} & \longrightarrow & \mathbf{Adj} & \mathbf{NP}, & & \text{Prob.} = p \\
\mathbf{NP} & \longrightarrow & \text{N}, & & & \text{Prob.} = 1-p \\
\mathbf{VP} & \longrightarrow & \texttt{verb}\ v & & & \text{Prob.} = p(v) \\
\mathbf{Adj} & \longrightarrow & \texttt{adj.}\ a & & & \text{Prob.} = p(a) \\
\text{N} & \longrightarrow & \texttt{noun}\ n & & & \text{Prob.} = p(n).
\end{array}$$

where **S, NP, VP, Adj,** N are the non-terminals, $p/(1-p)$ is the expected number of adjectives in a random noun phrase, `v,a,n` stand for a large number of possible terminals in a lexicon, and the probabilities $p(v)$, $p(a)$ and $p(n)$ are the frequencies of occurrence of the various verbs, adjectives and nouns.

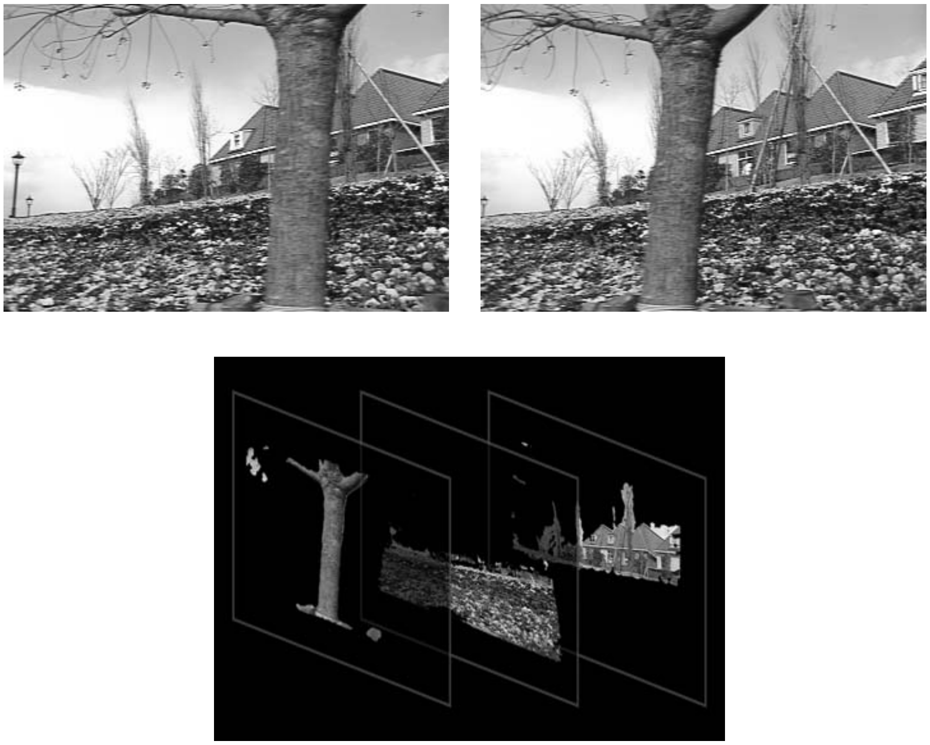

Figure 7: Top: two frames from a movie sequence of thirty images; bottom: their representation as a set of three layers of increasing depth (courtesy of Wang & Adelson).

In exactly this way, the 2.1D sketch is generated by the stochastic grammar:

$$\begin{array}{lclll}
\mathbf{Im} & \longrightarrow & \mathbf{Bkg} & \mathbf{Frg}, & \text{Prob.} = 1 \\
\mathbf{Frg} & \longrightarrow & \mathbf{Obj} & \mathbf{Frg}, & \text{Prob.} = p \\
\mathbf{Frg} & \longrightarrow & \mathbf{Obj}, & & \text{Prob.} = 1 - p \\
\mathbf{Bkg} & \longrightarrow & \mathtt{D} & & \text{Prob.} = 1 \\
\mathbf{Obj} & \longrightarrow & \mathtt{R} & & \text{Prob.} = p_2(R)
\end{array}$$

where **Im, Frg, Bkg, Obj** are the non-terminals, $1/(1-p)$ is the expected number of foreground objects in a random image, $\mathtt{D}$, $\mathtt{R}$ are the non-terminals, where $\mathtt{D}$ is the whole domain of the image and $\mathtt{R}$ can be any subset of the image

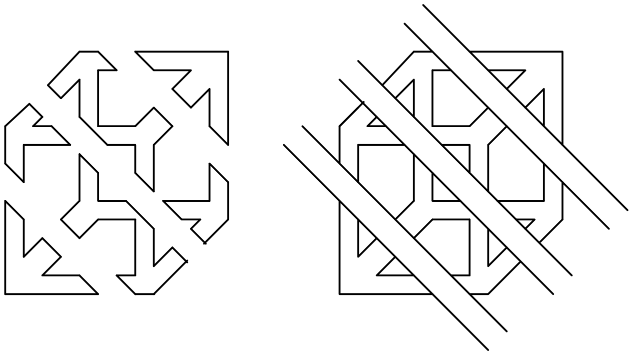

Figure 8: In center: 3D percept of three bars in front of cube; right: 3D percept persists without bars; left: figure separates into 2D shapes when only the occluding part of the bars are present (after Kanizsa).

domain and finally we define the probabilities $p_2(R)$ as:

$$p_2(R) = \frac{1}{Z} e^{-\int_{\partial R} \phi(\kappa) ds}$$

where ∂R is the boundary of R, κ is the curvature of this boundary, ds is arc length on this boundary and $\phi(x)$ is some function like $a + bx^2$. This prior on regions R encourages regions to have short smooth boundaries and, in particular, it will try to reconstruct the hidden edges of partially occluded objects by curves which minimize this functional (a class of curves invented by Euler and called by him *elastica*).

Tied together with a simple imaging model, such as:

$$p_2(I|\{R_i\}) = \frac{1}{Z} e^{-\sum_i \mathrm{Variance}_{R_i'}(I) \cdot \mathrm{Area}(R_i')}$$

we get model 2, the simplest grammatical model for generating global structure in an image. Just like an ordinary linguistic grammar, the purpose is to pick out large subsets of the signal that must be interpreted together and which may be interrupted by other structures. Thus foreground objects occluding part of the surface of a more distant object are like relative clauses embedded in a larger clause.

This idea of a grammar of images was invented by the Gestalt school of psychology in the early part of this century. They discovered many laws of grouping which they typically 'proved' by testing human responses to elegantly constructed images. An example is shown in figure 8, which is due to Kanisza

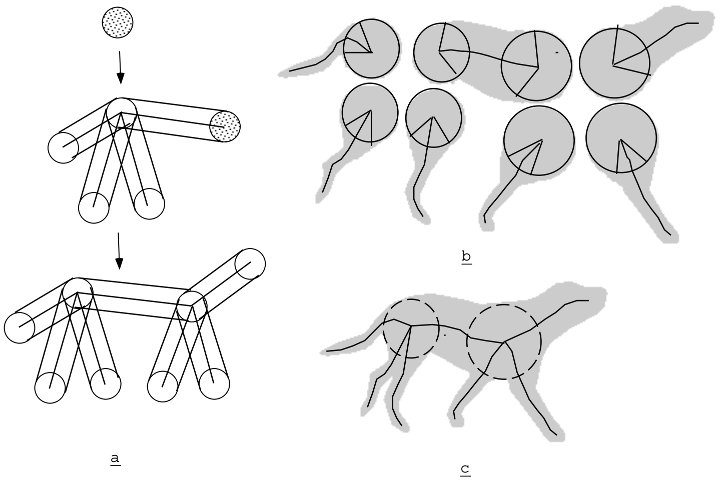

Figure 9: a) Generating by a grammar the part description of the dog using non-terminals symbols; b) the subsets of the image domain given by the corresponding terminal variables; c) the final dog silhouette and its medial axis.

[Ka]. Here you see in the middle a wire frame cube occluded by a set of parallel diagonal bars. Note that this percept still persists on the right where the edges of the wire frames end abruptly (a kind of T-junction with invisible top stroke); but this percept is absent on the left where the edges of the wire frame are joined, making each fragment into a self-contained two-dimensional shape.

More complex grammars are called for to deal with other aspects of images. In particular, there is a set of grammatical rules for the decomposition of complex articulated objects into ribbons and blobs with protrusions. These ideas go back to the work of Blum [Bl] and Fu [Fu]. Again, there is a set of abstract non-terminals and an infinite number of realizations of these as terminals which are subsets of the image domain $\mathtt{D}$. The ribbons produce worm-like shapes described by their axis and their width, the protrusions produce fin-like shapes described by an angular sector of a circle whose radius is a function of the angle. In figure 9, we give an example of this type of decomposition from the work of Zhu and Yuille [Z-Y].

A general characteristic of these grammatical models is that they incorporate a new class of variable. These are variables whose value is a subset $\mathtt{R}$ of the image domain. It is hard to force these into the Markov random field framework: not only are these global entities, but there has to be an unlimited supply of them – a shelf, as it were, of region variables waiting to be called upon. While perfectly satisfactory from a mathematical point of view, this raises a big problem when you try to imagine how the brain manipulates such entities. The brain, looked at neuro-anatomically, is a hard-wired graph of neurons very reminiscent of the kind of graph in a Markov random field. The model which I favor tries to reconcile these two using an *adaptive pyramid* architecture of the sort introduced by Hong and Rosenfeld [H-R]. In their construction, a series of successively coarser pixel grids are made into the levels of a pyramid with the original high resolution image at the bottom. They are linked, each level to the next lower level and the next higher level, by a many-to-many correspondence (in the original proposal, each pixel had 4 possible 'parents' and 16 possible 'children'), making the whole pyramid into a three-dimensional graph. Now you add a vertical line process which can cut or leave intact vertical links, or perhaps give them some weight in between: using these, pixels at higher levels can be adaptively linked to very general subsets R of the original, lowest level, thus creating subset variables.

4 Model 3: Object Templates using Mixed Markov Models

The final class of models I want to discuss are those incorporating the *semantics* of visual signals. Semantics deals with the construction of a database of individual things the agent has encountered and of categories of these things. In language you learn the names of objects and the meanings of words so as to use language correctly. In vision you learn the shape and appearance of objects and the clustering of objects into categories so as to recognize the object or instances of the category anew (and, in robotic applications, use this knowledge for navigation, grasping, etc.).

I want to start with an extremely simple example: in figure 10, from the work of Yuille, Hallinan and Cohen [Y-H-C], you see an image of a face on which an outline eye – consisting of two parabolas for the edges of the eyelids and a circle for the pupil – has been drawn by computer (more or less correctly). The theory, which goes back to early work of Fischler and Elschlager [F-E], is that to identify objects belonging to a category of known but variable shape in a given observed image, you must find the pixels in the image where some set of feature points in a model of the object are located. This approach goes under the name of *flexible templates.* In speech recognition, *time warping* plays a

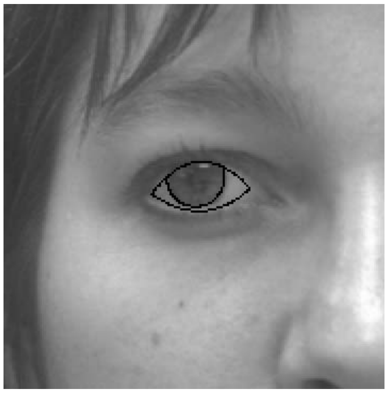

Figure 10: A flexible template of an eye, fit to a face image (from [Y-H-C]).

similar role in matching the expected temporal-frequency pattern of a specific phoneme with the observed sound. In general, we imagine that to recognize objects, you must learn a model for the object, called a template, which may be a typical image of the object or it may be cartoon-like with abstract points and edges or some combination of the two. You must also learn the typical amount of geometric variability of this template and of how the intensity values of the image should match the model. Then to recognize an instance of the object, this model must be fit to the present image.

A very extensively studied example of this approach is known as *model-based matching*. Here a precise geometric description of some object, like a machine part in a factory assembly line, is available. The assumption is that this object will be seen from an unknown point of view with unknown lighting conditions. The matching strategy employed is to identify the edges of the object in the image and also various special points, such as its corners. (A highly successful trick has been to look for bitangent lines, straight lines in the image domain tangent at two points to the edge of the object.) You must then solve for the viewpoint from which the outlines of the model would match up reasonably closely with the edges detected in the image, often under conditions of partial occlusion, so that the entire outline of the model cannot be seen in the image. People are remarkably good at this jigsaw puzzle like ability.

The difficulty of recognizing objects is highly dependent on the type of object. An unoccluded alphanumeric character from a known font or a flat machine part such as a gasket in good lighting conditions lie at the easiest end of the spectrum. Other 'objects', like a bunch of grapes, possess seemingly unlimited

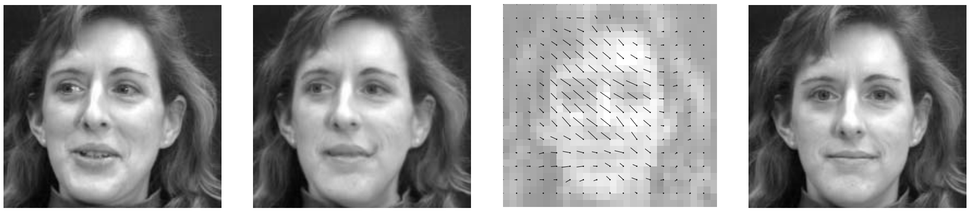

Figure 11: Left: the input face; next: the input matched by a warping of the template; next: the template with the warping indicated by arrows; right: the template itself. Note how the template is stretched in the mouth area because the mouth in the input is open.

variability. The case which has received the most recent attention because of its many applications, is face recognition. This is of intermediate difficulty: while faces are very stereotyped, their gray-level appearance is especially dependent on lighting conditions and they have only a small number of sharp internal edges. There has been extensive work in modeling the variability caused by a) viewpoint, b) lighting, c) expression, d) gross individual differences like glasses, facial hair and e) subtle individual characteristics like inter-ocular distance, shape of nose, etc. which identify each person.

To develop these ideas, I will describe one recent model for face recognition, which comes from the PhD thesis of Peter Hallinan [Ha1], [Ha2]. A quite similar model has been developed by Cootes, Lanitis and Taylor [L-T-C]. Since the face as a whole has few edges, Hallinan's model involves a dense set of feature points, i.e. the template face will be matched to the observed image via a diffeomorphism $\Phi : D_0 \hookrightarrow D$ of the domain of the template D_0 into the domain of the image D. On the other hand, to model arbitrary lighting conditions, let $J(p, \phi, \theta)$ be the gray-level image resulting from illuminating the template face with a spot light from the angle (ϕ, θ). Sampling the face with N pixels, J gives us a set of points $\vec{J}(\phi, \theta) \in \mathbb{R}^N$. We then take the principle components of this cluster in $\mathbb{R}^N$ and use the first 5 of them, $\{J_k(p)\}_{1 \leq k \leq 5}$, to approximate arbitrarily illuminated faces as a linear combination $\sum_{k=1}^{5} c_k J_k$. The final probability model is:

$$p_3(I, \vec{c}, \Phi, \sigma) = \frac{1}{Z} e^{-C_1 \iint_{D_0} (I(\Phi(x,y)) - \sum_k c_k J_k(x,y))^2 \, dx dy - C_2 \iint_{D_0} ||D\Phi^t \circ D\Phi - \sigma^2||^2 \, dx dy}$$

where $D\Phi$ is the Jacobian matrix of the diffeomorphism Φ, $\vec{c}$ is the vector

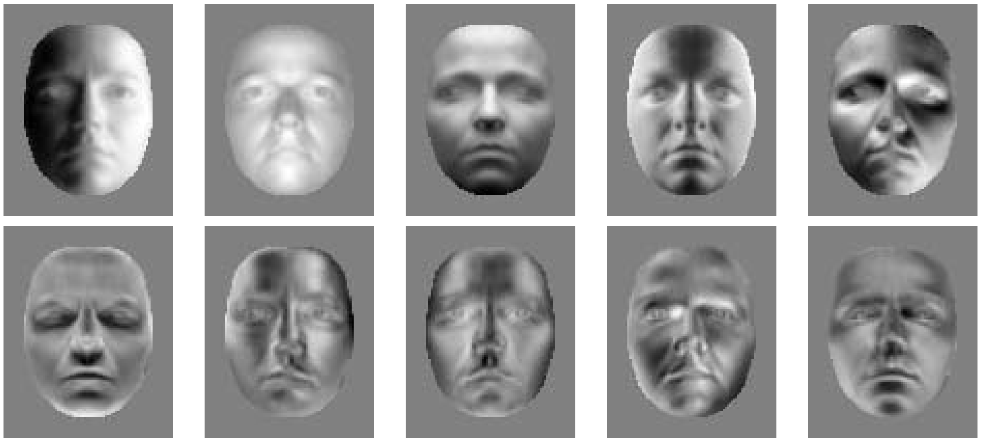

Figure 12: Ten *eigenfaces*: the principle components of the set of images obtained by all possible illuminations of the same face.

of lighting components and σ is a scale parameter. The prior here is quite crude as the distortions typical of rotating the head, changing expression and changing facial proportions should all be modeled explicitly. We give this example, however, to show that, in principle, probability models involving flexible templates and variable illumination can be built. In figure 11, we give an example of the warping Φ and in figure 12 we show the largest principle components J_k (also known as 'eigenfaces') for one individual.

What is new in this type of model? Fitting any template, whether it has a small number of feature points like the eye in figure 10 or a dense set as in figure 11, involves computing the pixels where template points are found in the observed image: these pixels are not intensities or weights or Boolean values, but are the *addresses* of other variables, namely the image values $I(p)$. Thus we have address-valued random variables or *pointers* as they are called in programming language theory. A natural way to incorporate such variables into the framework of Markov random fields is to imagine that all the edges of the model are not hard-wired, but some may be chosen 'at run time'. More formally, imagine a graph $G = (V, E)$ whose vertices V are divided into two groups $V = V_t \cup V_p$. We suppose a random variable X_v is given for each vertex v, that the value of each variable $X_v, v \in V_t$, is a real number while the value of each variable $X_v, v \in V_p$, is a vertex $w \in A(v)$ for some restricted subset

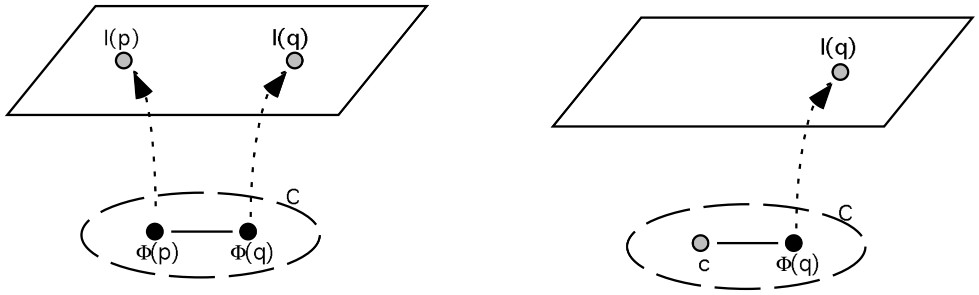

Figure 13: Mixed Markov models for face recognition: on left, the clique for $||D\Phi^t \circ D\Phi - \sigma^2||^2$; on right, the clique for $(I \circ \Phi - \sum c_k J_k)^2$.

$A(v) \subset V$ given as part of the definition. The effect of assigning values to the variables $\{X_v\}_{v \in V_p}$ is to augment the graph G by a new set of dynamic edges. This creates a new graph G^*. We call this set up a *mixed Markov model*. There seem to be several ways to define Gibbs distributions associated to such mixed Markov models. One of these is the following 'pull-back' definition:

$$p(\{X_t, X_p\}) = \frac{1}{Z} e^{-\sum_{\text{cliques } C \text{ in } G} E_C(\{X_v | v \in C\}, \{X_w | w = \bar{X}_v, v \in C \cap V_p\})}.$$

where $\bar{X}_v$ is the value of the variable X_v. This definition involves 'pulling back' the random variables referred to dynamically by members of a clique in G. This model includes the probability model p_3, as we show diagramatically in figure 13. It would be quite interesting to find a generalization of the Hammersley-Clifford theorem to mixed Markov models.

References

[B-Z] A.Blake & A.Zisserman: *Visual Reconstruction*, MIT Press, 1987.

[Bl] H.Blum, A transformation for extracting new descriptors of shape, *Symp. Models for Perception of Speech and Visual Form*, MIT Press, 1967.

[F-E] M.Fischler & R.Elschlager: The Representation and Matching of Pictorial Structures, *IEEE Trans. on Computers*, **22**, 1973.

[Fu] K.S.Fu: *Syntactic Pattern Recognition and Applications*, Prentice-Hall, 1982.

[G-Y] D.Geiger & A.Yuille, A common framework for image segmentation, *Int. J. Comp. Vis.*, **6**, 1991.

[G-G] S.Geman & D.Geman: Stochastic Relaxation, Gibbs Distribution and Bayesian Restoration of images, *IEEE Trans. Patt. Anal. Mach. Int.*, 1984.

[Gre1] U.Grenander: *Lectures in Pattern Theory*, Springer-Verlag, 1976.

[Gre2] U.Grenander: *General Pattern Theory*, Oxford Univ. Press (1994)

[Ha1] P.Hallinan: A low-dimensional representation of human faces for arbitrary lighting conditions, *IEEE Conf. Comp. Vis. Patt. Rec.*, 1993.

[Ha2] P.Hallinan: *PhD thesis*, Divison of Applied Science, Harvard University, (1995)

[HvH] H.von Helmholtz: *Physiological Optics*, original German edition, 1909; Dover Publ. translation, 1962.

[H-R] T.Hong & A.Rosenfeld: Compact region extraction using weighted pixel linking in a pyramid, *IEEE Trans. Patt. Anal. Mach. Int.*, **6**, pp.222-229, 1984.

[Ka] G.Kanizsa, *Organization in Vision*, Praeger, 1979.

[K-V] M.Kearns & U.Vazirani, *Computational Learning Theory*, MIT Press, 1994.

[L-T-C] A.Lanitis, C.Taylor & T.Cootes, A Unified Approach to Coding and Interpreting Face Images, *Proc. IEEE 5th Int. Conf. Comp. Vis.*, 1995.

[Ma] D.Marr: *Vision*, W.H.Freeman, 1982.

[N-M] M.Nitzberg & D.Mumford: The 2.1D Sketch, *Proc. 3rd IEEE Int. Conf.Comp.Vision*, 1990.

[Pe] J.Pearl, *Probabilistic Reasoning in Intelligent Systems*, M. Kaufmann, 1988.

[R-C-L] M.Reed, P.Colella & O.Lanford, lectures in *Constructive Quantum Field Theory*, Springer Lecture Notes in Physics, **25**, 1973.

[Sh] C.Shannon, A mathematical theory of communication, *Bell System Tech. J.*, **27**, 1948.

[W-A] J.Wang & E.Adelson, Representing moving images with layers, *IEEE Trans. Image Proc.*, **3**, 625-638, 1994.

[Y-H-C] A.Yuille, P.Hallinan & D.Cohen: Feature Extraction from Faces using Deformable Templates, *Int. J. Comp. Vision*, **6**, 1992.

[Z-L-Y] S.C.Zhu, T.S.Lee & A.Yuille, Region competition: Unifying snake, region growing and Bayes/MDL for multi-band image segmentation, submitted to *IEEE Trans. Patt. Anal. Mach. Int.*.

[Z-Y] S.-C.Zhu & A.Yuille: A framework for object representation and recognition, *Proc. 1st Int.Conf.Image Proc.*, 1994.

Address:

D. Mumford, Department of Mathematics, Harvard University, Cambridge, MA 02138, USA; email: `mumford@math.harvard.edu`.