BAYESIAN IMAGE ANALYSIS

Donald Geman[1]
Department of Mathematics and Statistics
University of Massachusetts
Amherst, MA 01003, USA

Stuart Geman[2]
Division of Applied Mathematics
Brown University
Providence, RI 02912, USA

In: NATO ASI Series, Vol. F20, Disordered Systems and Biological Organization
Springer-Verlag, Berlin, 1986

I. INTRODUCTION.

In [8] we introduced a class of image models for various tasks in digital image processing. These models are multi-level or "hierarchical" Markov Random Fields (MRFs). Here we pursue this approach to image modelling and analysis along some different lines, involving segmentation, boundary finding, and computer tomography. Similar models and associated optimization algorithms appear regularly in other work involving immense spatial systems; some examples are the studies in these proceedings on statistical mechanical systems (e.g. ferromagnets, spin-glasses and random fields), the work of Hinton and Sejnowski [14], Hopfield [15], and von der Malsburg and Bienenstock [19], in neural modeling and perceptual inference, and other work in image analysis, e.g. Besag [2], Kiiveri and Campbell [17], Cross and Jain [5], Cohen and Cooper [4], Elliott and Derin [7], Deviver [6], Grenander [11], and Marroquin [20]. The use of MRFs and related stochastic processes as models for intensity data has been prevalent in the image processing literature for some time now; we refer the reader to [8] and standard references for a detailed account of the genealogy.

The aforementioned analogy between very large (usually spatial) stochastic systems such as those encountered in digital image processing, computer vision, and neural modelling, and the lattice-based systems of statistical mechanics has been an important theme of our past work. For instance, our computational algorithms are based on a new optimization technique called "simulated annealing", introduced by Cêrny [3] and

1. Research supported in part by the Office of Naval Research, Contract N000-14-84-K-0531, and The National Science Foundation, DMS-8401927.
2. Research supported in part by the Army Research Office, Contract DAAG-29-83-K-0116, and The National Science Foundation, DMS-8306507 and DMS-8352087.

Kirkpatrik et al [18]. Stochastic relaxation and simulated annealing are briefly discussed in § V and remain the basis of our reconstruction and segmentation algorithms. However, the focus here is on image modelling, statistical inference, and new applications.

Our image models are "hierarchical" and stochastic. First, we regard the "image" as a collection of attribute processes, only one of which is the usual array of intensity or brightness values. The other, mainly geometric, attribute processes are constructs, corresponding to edges, object locations, feature labels, and so forth; they are part of the image model but not of the physical data. We use the term "hierarchical" to reflect the fact that image attributes such as boundaries and texture labels involve increasingly global and contextual information and expectations.

We have chosen the family of MRF priors for "images" for several reasons. First, we believe this formulation provides a solid, theoretical basis for complex image modelling: the class of models is extremely rich and easily accommodates a multi-level framework. Indeed, spatially-invariant, geometric attributes such as edges, curves, and simple polygons (with arbitrary scale and location) can be incorporated in the model in a local fashion. This was illustrated in [8] with the addition of a "line process". Second, the duality between MRFs and Gibbs distributions (see § II) allows the modelling process to be explicit and constructive: we build energy functions to quantify our a priori expectations about imagery. Finally, for many types of degredations (see § III), the conditional independence (= Markov property) of the prior is inherited by the posterior distribution. This is crucial because it guarantees a satisfactory degree of computational feasibility; see §§ IV, V.

For many problems in "low-level" image processing and related fields the current models appear adequate; other needs are more pressing, such as reducing the computational load and developing a rational data-driven method for estimating parameters in the model (see § VI). It remains to be seen whether the hierarchical MRF framework can accomodate the necessary high degree of external knowledge to deal with problems in "high-level" vision (for instance object recognition and texture labeling). Basic concepts such as scale and shape must be merged into the graph and model structures, and in a way that is sufficiently local to avoid unrealistic amounts of computation. Some of our preliminary experiments, and those of others, are encouraging. This paper addresses a "middle-level" of problems in reconstruction and segmentation in which excellent results are possible with some degree of "knowledge engineering", coupled with a careful analysis of the degradation mechanism.

PRIOR DISTRIBUTIONS ON IMAGES.

Let $\underset{\sim}{X}^P = (X^P_{ij})$, $1 \leq i,j \leq N$ denote the pixel values associated with an N×N (digitized) picture. Usually, each X^P_{ij} represents the intensity of electromagnetic radiation in some frequency band that is emitted or reflected from a small region in the true "scene" or "object plane". (We regard these as the "ideal" intensities, uncorrupted by the recording system; in § III we will consider the nature of the actual, observed data.) Some examples we have in mind are grey-tone, infrared, and tomographic imagery, but the same analysis applies to other imagery, for example x-rays or a channel of multispectral landsat data.

As already discussed, we view the image as the realization of a compound stochastic process $\underset{\sim}{X} = (\underset{\sim}{X}^P, \underset{\sim}{X}^E, \underset{\sim}{X}^L, \ldots)$ in which $\underset{\sim}{X}^E$ might denote an array of "edge variables", $\underset{\sim}{X}^L$ certain "label variables", etc. In this paper we shall only consider an edge process $\underset{\sim}{X}^E$ in addition to the "pixel process" or "intensity process" $\underset{\sim}{X}^P$. Since our approach is Bayesian, we are going to impose a prior probability distribution on the set of possible values of $\underset{\sim}{X}$, which we denote by Ω. Thus, for example, if $\underset{\sim}{X}=\underset{\sim}{X}^P$ only, and we are only interested in binary imagery, one would assign probabilities to each of the 2^{N^2} elements of Ω.

In order to define our family of priors, we must specify exactly what we mean by $\underset{\sim}{X}^E$. Let s,t denote points in the square lattice. For each pair s,t of adjacent horizontal or adjacent vertical pixels we append an "edge site", denoted <s,t>, to the lattice; it corresponds to the "location" of a putative edge or boundary element between pixels s and t. In the simplest case, the edge variables are binary, with 0 and 1 representing the absence or presence of an edge at <s,t>. Then $\underset{\sim}{X}^E$ consists of the 2N(N-1) variables $X^E_{<s,t>}$.

The totality of pixel and edge sites in denoted by S. Given a neighborhood system $\boldsymbol{G} = \{\boldsymbol{G}_\alpha, \alpha \in S\}$ (see [8]) a stochastic process X on S is a MRF if $P(\underset{\sim}{X}=x)>0$ for all $x \in \Omega$ and

$$P(X_\alpha = x_\alpha \mid X_\beta = x_\beta,\ \beta \neq \alpha) = P(X_\alpha = x_\alpha \mid X_\beta = x_\beta,\ \beta \in \boldsymbol{G}_\alpha)$$

for every $\alpha \in S$ and $x = (x_\gamma)$, $\gamma \in S$, in Ω. In words, the conditional probability of seeing the value x_α at site α given any other configuration for the remaining sites depends only on the states of the neighbors of α. In our case, the α's and β's denote pixel or edge sites. The size of the neighborhood determines the range of interactions, and we shall say that $\underset{\sim}{X}$ is "locally-composed" or just "local" if $|\boldsymbol{G}_\alpha|$ is small, say less than ten or twenty. Roughly speaking, these models are computationally feasible to the extent that $\underset{\sim}{X}$ is local. It is now well-known that a process $\underset{\sim}{X}$ on a graph $(S,\boldsymbol{G})$ is a MRF if and only if its joint probability distribution $\Pi(x) = P(\underset{\sim}{X}=x)$, $x \in \Omega$, is a Gibbs distribution on $(S,\boldsymbol{G})$. This means that Π has the form

$$\Pi(x) = e^{-U(x)}/Z. \qquad Z = \sum_x e^{-U(x)}$$

where the energy function U contains interactions confined to the cliques of the graph. Loosely speaking, this means that x_α and x_β may appear together in a term in U only if α and β are neighbors. Examples should make this clear and we again refer the interested reader to [8] for a complete discussion. Suffice it to say that the Gibbs formulation is convenient for modelling whereas the Markov property ensures that one can indeed examine samples from such a process.

We restrict ourselves to the following neighborhood system. Each pixel site has eight pixel neighbors, the nearest ones, and four edge neighbors; each edge site <s,t> has six edge neighbors (corresponding to the possible "continuations" of a boundary at <s,t>) and the two pixel neighbors s and t. Sites near the boundary of the lattice have fewer neighbors. Two of these neighborhoods, one for a pixel and one for a "vertical" edge site, are shown in Figure 1, in which the circles and pluses denote pixel and edge sites respectively. (We believe this edge graph originated in [13].)

Figure 1

To illustrate the functional form of the models suppose first that we were only interested in modeling "smoothness" or "regularity" in the intensity array, i.e. the tendency of nearby pixels to have similar intensities. Then a suitable model might be $\underset{\sim}{X}=\underset{\sim}{X}^P$ with

$$P(\underset{\sim}{X} = x) = Z^{-1} \exp\left\{\theta \sum_{(s,t)} \phi(x_s - x_t)\right\} \qquad (1)$$

where the sum extends over all neighbor pairs (s,t) of pixels. (Thus each interior pixel is included in eight terms in the summation.) Here $\phi=\phi(u)$ is even and decreasing for $u>0$, and θ is a parameter which corresponds to inverse temperature and controls the degree of regularity. The extreme cases are $\theta=0$, corresponding to pure noise, and $\theta=\infty$ in which case the distribution is concentrated on images of constant intensity.

We shall consider two examples of these "potentials" ϕ, depending on the possible intensity values, say $X_\alpha \in \Lambda$, $\alpha \in S$. If Λ is discrete, say $\Lambda=\{0,1,2,\ldots,L\}$, and L is small, then

one simple choice is

$$\phi_1(u) = \begin{cases} 1, & u = 0 \\ -1, & u \neq 0. \end{cases} \qquad (2)$$

In particular, the conditional probability that X_α=j given the eight neighboring values depends only on the number $N_s(j)$ of neighbors which agree with X_α. Specifically,

$$P(X_\alpha = j \,|\, X_\beta,\ \beta \neq \alpha) = \frac{\exp\,(2\theta N_\alpha(j))}{\sum_{k=0}^{L} \exp\,(2\theta N_\alpha(k))},\quad 0 \leq j \leq L.$$

For a binary image, this is a weighted majority rule: the log odds of a 1 to a 0 are $2\theta(N_\alpha(1) - N_\alpha(0))$.

If L is large or Λ is a <u>continuous</u> interval [0,L], then we have adopted potentials of the form

$$\phi_2(u) = \left(1 + \left|\frac{u}{C_1}\right|^{C_2}\right)^{-1},\quad u \in [-L,L] \qquad (3)$$

where C_1,C_2 are parameters; usually C_2=1.5 or 2.0 and C_1 depends on the dynamic range of the image. One reason for this choice is that if ϕ were to decrease too rapidly (e.g. $\phi(u)=-u^2$) we would a priori inhibit (almost <u>pro</u>hibit) adjacent, roughly homogeneous regions of highly separated intensities.

With the inclusion of the edge process X^E we incorporate our expectations about <u>both</u> the interactions between intensities and edges (i.e. where edges "belong") and about clusters of nearby edges. (It should be noted that, at this level of the hierarchy, we are not exactly modelling boundaries but rather <u>segments of boundaries</u>; except in the simplest imagery and with larger neighborhoods, it is essentially impossible to distinguish actual boundary segments from intensity gradients due to lighting, texture, etc.). We conclude this section with an example of an energy function U for $\underset{\sim}{X} = (X^P,X^E)$. The energy $U(x^P,x^E)$ consists of two terms, say $U = U^1(x^P,x^E) + U^2(x^E)$. We want to construct U^1 such that the most likely configurations will have $x^E_{<s,\,t>} = 1$ (resp. =0) when the intensity difference $|x^P_s - x^P_t|$ is large (resp. small). Put differently, we want to break the bond between pixels s and t when their values are "far" apart. Thus we choose

$$U^1(x^P,x^E) = -\sum_{(s,t)} (\theta_1 \phi_2(x^P_s - x^P_t) - \theta_2)\,(1 - x^E_{<s,t>}) \qquad (4)$$

where $\theta_1 > \theta_2 > 0$. The value of u for which $\theta_1\phi(u) = \theta_2$ represents an intensity difference for which we have no preference in regard to the state of an edge. Finally, the organization of nearby edges is controlled by

$$U^2(x^E) = -\theta_3 \sum_D V_D(x^E) \qquad (5)$$

($\theta_3>0$), where the sum extends over all subsets D of four neighboring edge sites (the maximal "cliques" in the edge graph), and V_D assigns weights in accordance with our expectations about edge behavior. More specifically, there are six possible clique states (up to rotations):

Figure 2

Here the slashes indicate that the edge variable at the indicated site is "on". Let $V_D=\xi_i$, $1\leq i\leq 6$, denote the weights assigned to the six above configurations in Figure 2. If we assume that most pixels are not next to boundaries, that edges should continue, and that boundary congestion is unlikely, than we might choose $\xi_1\leq\xi_2\leq\xi_3\leq\xi_4\leq\xi_5\leq\xi_6$. A specific, image-dependent choice is made in the block experiment; see § VII.

A final point: it is useful to rewrite the total energy, up to a constant, as

$$-U(x) = \theta_1 \sum_{(s,t)} \phi(x_s^P - x_t^P)(1 - x^E_{<s,t>}) + \theta_2 \sum_{<s,t>} x^E_{<s,t>} + \theta_3 \sum_D V_D(x^E) \qquad (6)$$

For inferential purposes, this shows that our model is an exponential family in $\theta = (\theta_1,\theta_2,\theta_3)$. In addition, the form in (6) is helpful for parameter interpretation; for instance, θ_2 is clearly a "reward" index for edges.

III. DEGRADATIONS.

We actually observe some transformation $\underset{\sim}{Y}=\Gamma(\underset{\sim}{X}^P)$ of the intensity process, three examples being:

(i) Filtering and Deconvolution. In many cases the pixel intensities do not represent the radiant energy at the source; rather, this energy is transformed due to the detecting and recording system. This is true both for photochemical (e.g. film) and photoelectronic (e.g. video) systems, and both usually involve blur and noise. A generic model for (space-invariant) degradation is then

$$y_s = g\Big(\sum_{s-r \in D} H_{s-r}\, x_r^P\Big) + \eta_s. \qquad (7)$$

Here, D is a symmetric neighborhood of the origin, H is a blurring matrix, g accounts for nonlinearities in the recording device, and $\eta = (\eta_s)$ is a random noise due to the sensor, digitization, etc. In some cases the noise may be multiplicative (e.g. in synthetic aperture radar) or the blur may be anisotropic (e.g. in certain infrared scanners). One of the best features of the MRF formulation is that all such degradations are easily handled. We will assume that η is white Gaussian noise and statistically independent of $\underset{\sim}{X}$. The general restoration problem is then to recover $\underset{\sim}{X}^P$ from the data $\underset{\sim}{Y}=(y_s)$ assuming we are given g , H and the noise statistics.

(ii) Boundary-finding. Another type of information loss occurs in the segmentation of "natural scenes" and other imagery which, for all practical purposes, can be taken as uncorrupted. Since we regard the image as $\underset{\sim}{X} = (\underset{\sim}{X}^P, \underset{\sim}{X}^E)$, what is observed is $\underset{\sim}{X}^P$, whereas $\underset{\sim}{X}^E$ must be inferred. Of course this "transformation", a projection, is to some extent merely an artifact of the model. Nonetheless, from the viewpoint of statistical inference, the information loss is severe.

Whether the primary goal is segmentation or restoration, it can be useful to combine these tasks into a single algorithm. For example, we found in [8] that the inclusion of the edge process $\underset{\sim}{X}^E$ facilitated the restoration of images degraded in accordance with (7), especially when some a priori knowledge about the boundary behavior was available. Conversely, even if the object is segmentation, some de-blurring or noise removal may improve performance. Another advantage of the hierarchical MRF formulation is that these tasks can be combined into a single process. It is well-known that smoothing often degrades boundary behavior, thereby making the segmentation problem more difficult. See Marroquin [20] for a discussion of these issues in the context of surface reconstruction.

(iii) Single Photon Emission Tomography. In this case the pixel lattice corresponds to a discretization of a cross-section of tissue and $X_s = X_s^P$ represents the concentration of some isotope at site s. Particles are emitted in random directions from these sites and follow the usual Poisson laws for radiation counts. In particular, the number of particles emitted from s is a Poisson random variable with rate proportional to X_s. (The time interval is fixed, and hence can be ignored.) These particles are received and counted at banks of detectors which are placed around the lattice and in the same plane. However, there is attenuation due to the passage of the photons through the tissue or whatever media is storing the isotope. Thus, the number of received particles at a given detector k (which we denote y_k) is Poisson with rate

$$E\,[y_k] = \int_{L_\theta} X_t\, A_k(t)dt \qquad (8)$$

where L_θ is the line with direction θ, and A embodies the attenuation factor for the segment from t to the detector k as well as the details of the detector geometry. The object is to recover the isotope density $\{X_s\}$ from the detector counts. See [9] for more information.

IV. POSTERIOR DISTRIBUTIONS.

Given the data $\underset{\sim}{Y}=y$, the posterior distribution is

$$\pi_y(x) = P(\underset{\sim}{X}=x \,|\, \underset{\sim}{Y}=y), \quad x\in\Omega .$$

This is a powerful tool for image analysis: in principle we can construct the optimal (Bayesian) estimator for $\underset{\sim}{X}$, examine images sampled from π_y, design near-optimal statistical tests for the presence of special objects, and so forth.

If the data transformation Γ is sufficiently "local", then the conditional probability law of $\underset{\sim}{X}$ is also a MRF with a local graph structure. Let $U_y(x)$ denote the energy function in the representation of $\pi_y(x)$ as a Gibbs distribution;

$$\pi_y(x) = e^{-U_y(x)}/Z_y, \qquad Z_y = \sum_x e^{-U_y(x)} , \tag{9}$$

Then if Γ is local so is U_y. The practical import of this observation is that stochastic relaxation methods (such as the "heat-bath" and Metropolis algorithms) are feasible for analyzing π_y.

The types of degradation we have discussed are mostly "local". For example, the degradation in (7) leads to a locally-composed U_y whenever D is small and η is nearly white. In the case of boundary-finding, the posterior distribution is

$$\pi_y(x) = \pi_y(x^E) = P(\underset{\sim}{X}^E = x^E \mid \underset{\sim}{X}^P = x^P)$$

and the posterior energy U_y is then simply the expression in (6) with $y=x^P$ fixed; in particular, the posterior graph is just the subgraph for the edge sites. In contrast, tomogrophy leads to a non-local posterior energy, and potentially severe computational problems. So far, we have largely avoided these by employing more conventional reconstructions as starting points for our Bayesian algorithm (see § VII and Geman and McClure [9] for more details).

The mode(s) of π_y is called the maximum a posteriori or MAP estimator of x given

y, and much of our previous work has focused on the development of an algorithm to find near-MAP estimates. The computational problem is formidable. We seek to minimize the posterior energy function $U_y(x)$ over $x \in \Omega$. Typically, this function is highly non-linear, has an enormous number of (suitably defined) local minima, and the size of Ω is at least 2^{1000}, corresponding to a very small (32×32), binary intensity array and no edge units.

To illustrate the problem, consider the prior in (1) with $\Phi(u) = (1+u^2)^{-1}$ and additive white Gaussian noise with variance σ^2. Then a simple calculation gives

$$U_y(x) = - \sum_{<s,t>} \frac{\theta}{1+(x_s^P - x_t^P)^2} + \frac{1}{2\sigma^2} \sum_s (x_s^P - y_s)^2. \qquad (10)$$

The first term imposes smoothness and the second fidelity to the data, with relative emphasis in accordance with $\theta\sigma^2$.

V. STOCHASTIC RELAXATION AND SIMULATED ANNEALING.

Stochastic relaxation is an iterative, site-replacement procedure for generating a sample configuration from a Gibbs distribution, π. In our applications, $\pi = \pi_y$, the posterior distribution given $\underset{\sim}{Y}=y$. We refer the reader to [8] for a complete treatment, including the origins in physics, the precise mathematical formulation, and a comparison with so-called "probabilistic relaxation" [16] or "relaxation labeling". Suffice to say that the algorithm generates a (nonstationary) Markov chain $\underset{\sim}{X}(k)$, k = 0,1,2,..., with state space Ω and asymptotic distribution π:

$$\lim_{k\to\infty} P(\underset{\sim}{X}(k) = x \,|\, \underset{\sim}{X}(0) = x^1) = \pi(x) \qquad x, x^1 \in \Omega . \qquad (11)$$

To find the mode(s) of π, we pursue the analogy with statistical mechanics and regard these configurations as the ground states of an (imaginary) physical system with energy U(x). We then simulate the physical process of annealing, in which the slow decrease of temperature T forces the system into its low energy states. Roughly speaking, this occurs because the Boltzmann distribution at temperature T is $\pi_T = \exp(-U(x)/T)/Z_T$ and hence, whereas the ground states are unchanged, their relative weight increases as T decreases. Simulated annealing then refers to the slow decrease of a control parameter T during the generation of the Markov chain. Given a decreasing sequence T(k), k = 1,2,..., the annealing algorithm generates a new Markov chain $(\underset{\sim}{X}(k))$

whose asymptotic distribution as $k \to \infty$ is uniform over the set $\{z \in \Omega : U(z) = \min_x U(x)\}$. The only condition is that T(k) decrease sufficiently slowly, namely that

$$T(k) \geq C/\log(1 + k) \tag{12}$$

for a certain constant C=C(U); see [8], [10], and [12].

The algorithm is computationally feasible to the extent that π is local because at iteration k a sample must be obtained from the conditional distribution

$$\pi_{T(k)}(X_{s(k)} = \cdot \,|\, X_r = X_r(k-1),\ r \in G_{s(k)}),$$

where {s(k)} is some pre-determined sequence for visiting the sites. (Of course, the computation time also depends heavily on the size of the graph, the intensity range, and other factors.) Finally, the algorithm is highly parallel in the sense that it can be executed by simple and alike processing units acting largely independently. Basically, one can cut the time in half with two processors, in thirds with three, etc. These processors would be assigned to collections of sites, and the exact degree of parallelism would depend on the chromatic index of the graph. For instance, with a nearest neighbor graph, one could update the "red" sites simultaneously, then all the "blacks", etc.

VI. PARAMETER ESTIMATION.

In our previous work on image restoration, certain parameters which appear in the prior MRF image models, for example θ in (1), were not estimated from the data, but rather divined, guided by experience and the persistent observation that the quality of the restorations was surprisingly insensitive to the choice of θ, at least over a fair-sized interval. However, in current experiments in tomography, segmentation, and computer vision the models are more complex and are likely to be still more so in envisioned work involving object recognition, texture analysis, etc. These new models (e.g. equation (6)) involve additional parameters whose interpretations are less apparent than, for instance, that of θ in (1). Moreover there is growing evidence that the algorithms are less robust. Consequently, one needs an accurate, data-driven method of parameter estimation.

Statistical inference is complicated by the high-dimensionality of the data and the severe loss of information in the transformation $\underset{\sim}{X} \to \underset{\sim}{Y}$. Thus it is perhaps not too surprising that the statistics and image processing literature contain very few papers that are relevant for situations akin to ours. The work of Besag ([2], and references therein) on "coding schemes" and "pseudo-likelihood" is an exception, although this work is primarily confined to the case of "complete data", i.e. $\underset{\sim}{Y} = \underset{\sim}{X}$. We of course are basically

interested in the case of "incomplete data", as illustrated in § III.

In the statistical terminology, our models are "exponential families", which refers to the fact that the parameters appear multiplicatively in the un-normalized log likelihoods. The major difficulty is that the normalizing constant Z, the so-called partition function in statistical mechanics, is a function of these parameters and entirely intractable.

To illustrate the pitfalls in conventional approaches, we are going to briefly consider the difficulties encountered in maximum likelihood estimation. To simplify matters, however, we make the following assumption:

(i) Γ is a projection, i.e. $\underset{\sim}{X} = (\underset{\sim}{W},\underset{\sim}{Y})$ where Y is observed and W is not observed. (Some authors refer to the elements of W as "hidden units".)

(ii) $\underset{\sim}{X}$ is MRF with a local graph structure.

(iii) The parameters $\theta = (\theta_1,\theta_2,...,\theta_J)$ appear multiplicatively in the representation of the distribution of $\underset{\sim}{X}$ as a Gibbs measure:

$$P_\theta(X=x) = Z^{-1}(\theta) \quad \exp -\sum_{j=1}^{J} \theta_j U_j(x), \qquad x = (w,y).$$

These restrictions are actually less severe than might be expected; indeed, all the examples we have seen so far satisfy (i) - (iii).

Example 1. Consider the case of simple filtering with additive white Gaussian noise with mean 0 and variance σ^2, no edge process, and a prior on $\underset{\sim}{X}^P$ of the form (1). Then with $\underset{\sim}{Y}=\underset{\sim}{X}^P+\eta$ and $\underset{\sim}{W}=\underset{\sim}{X}^P$, the pair $(\underset{\sim}{W},\underset{\sim}{Y})$ has distribution

$$P_\theta(\underset{\sim}{X}^P=x^P,\underset{\sim}{Y}=y) = Z^{-1}(\theta) \quad \exp -(\theta_1 U_1(x^P) + \theta_2 U_2(x^P,y)) \tag{13}$$

where $\theta_2 = (-2\sigma^2)^{-1}$, $U_1 = \sum_{(s,t)} \Phi(x_s^P - x_t^P)$, $U_2 = \sum_s (x_s^P - y)^2$, and $Z(\theta) = Z(\theta_1,\theta_2) = Z(\theta_1)\,(-\pi/\theta_2)^{N^2/2}$. The joint energy is clearly local. The same reasoning applies to more complex degradations of the family (7).

Example 2. Consider the pixel-edge model $\underset{\sim}{X} = (\underset{\sim}{X}^E,\underset{\sim}{X}^P)$. Taking $\underset{\sim}{W}=\underset{\sim}{X}^E$, $\underset{\sim}{Y}=\underset{\sim}{X}^P$, we see that the joint law is simply the prior distribution on $\underset{\sim}{X}$. An illustration of the parameter estimation problem is then to estimate θ_1,θ_2 and θ_3 in (6) based on observations of $\underset{\sim}{X}^P$. This is difficult for several reasons, the main one being that the marginal distributions of $\underset{\sim}{X}^E$ and $\underset{\sim}{X}^P$ have fully-connected graphs! In particular, the "likelihood function" $P_\theta(x^P)$ is intractable.

a) Maximum likelihood estimation. The distribution of the observed variables is

$$P_\theta(y) = \frac{Z(\theta|y)}{Z(\theta)}, \qquad Z(\theta|y) = \sum_w e^{\sum_j \theta_j U_j(w,y)}.$$

In the classical case of independent and identically distributed observations $y^{(1)},...,y^{(n)}$,

the (normalized) log-likelihood is

$$\frac{1}{n}\sum_{k=1}^{n} \log Z(\theta|y^{(k)}) - \log Z(\theta)$$

and the likelihood equations,

$$\nabla \log \prod_{k=1}^{n} P_\theta(y^{(k)}) = 0 \quad \text{reduce to}$$

$$E_\theta U_j = \frac{1}{n}\sum_{k=1}^{n} E_\theta(U_j|y^{(k)}), \quad j = 1,2,...,J. \tag{14}$$

This system is intractable as it stands: the expected values are impossible to calculate (for the same reasons that mean energies are in spin-glasses and the like) and even when they can be estimated (by sampling) there still remains the problem of solving (14). In particular, the log likelihood is highly non-convex.

The "EM" algorithm is an iterative scheme designed for solving systems such as (14), although mainly in more conventional settings involving familiar densities (normal, Poisson, etc.) and much lower dimensional data. The algorithm does not seem suitable in its customary form and we have developed a number of modifications. The basic idea is to generate a sequence $\hat{\theta}^{(k)}$ of estimates intended to converge to a local maximum of the likelihood. We "update" $\hat{\theta}^{(k)}$ by first sampling from the posterior distribution $P_\theta(\underset{\sim}{W}=w|\underset{\sim}{Y}=y)$ at $\theta=\hat{\theta}^{(k)}$ and then choosing $\hat{\theta}^{(k+1)}$ to maximize or simply increase the joint likelihood $P_\theta(\hat{w}^{(k)},y)$. Results and experiments will be reported elsewhere.

b) A Priori Constraints. We have started using (a more complex version of) the model in equation (6) for the segmentation of "natural scenes" such as faces and houses, and for the segmentation of infrared imagery. However in addition to the generic difficulties discussed earlier, statistical inference for $\underset{\sim}{\theta} = (\theta_1,\theta_2,\theta_3)$ from the intensity image $\underset{\sim}{X}^P$ is further complicated by the fact that relatively disparate values of the parameter $\underset{\sim}{\theta}$ may induce essentially the same marginal distribution on $\underset{\sim}{X}^P$. Moreover, not all of these values may correspond to "good segmentations"; for example, the likely states of $P_\theta(\underset{\sim}{X}^E=\cdot\,|\underset{\sim}{X}^P)$ may not conform to our prior expectations of where the boundaries "belong" in $\underset{\sim}{X}^P$. Therefore, estimation based on the intensity image is not possible.

Fortunately, we can use these prior expectations to restrict the parameter space. In fact, we can sometimes identify a small region of the parameter space, $\Lambda\subset R^3$, with the following property: given a class of very simple "training images" $\underset{\sim}{X}_Y$ for which a desired segmentation $\underset{\sim}{X}^E_Y$ is simply and unambiguously defined, the posterior distribution $P_\theta(\cdot\,|\underset{\sim}{X}^P=\underset{\sim}{X}_Y)$ will be maximized at $x^E=\underset{\sim}{X}^E_Y$ only if $\theta\in\Lambda$. We refer to this as "reparametrization" because the set Λ depends on other parameters which directly reflect our characterization of "good segmentations". For example, one such parameter might be that value of the minimum difference across the boundary between a (candidate) "object"

and "background" such that we have no preference whether or not to segment the object. Typically, Λ turns out to be a line. Estimation is then reduced to one scale parameter corresponding to temperature, and this is rather easily handled by the variations on EM discussed earlier.

VIII EXPERIMENTAL RESULTS.

There are three sets of experiments, intended to illustrate a variety of image and degradation models previously described.

a) Blocks. These results appear in [8] and are reproduced here to illustrate the power of the hierarchical approach for image restoration. The original image, Figure 3(a), is "hand-drawn". We added Gaussian noise with mean 0 and variance σ^2=.49 to produce Figure 3(b). We then attempted restorations with and without $\underset{\sim}{X}^E$. Figure 3(c) is the restoration with simulated annealing with the prior in (1) with θ=1/3 and Φ as in (2). The inclusion of $\underset{\sim}{X}^E$ yields significant improvement - Figure 3(d). The model is essentially (6) with Φ above, θ_1=1, θ_2=0, θ_3=.9, and the following clique weights: ξ_1=0, ξ_2=1, ξ_3=ξ_4=2, ξ_5=ξ_6=3. The reason for favoring straight lines is obvious from Figure 3(a), and nicely illustrates the use of prior knowledge. The second set of block pictures illustrates the flexibility of the model in regard to different degradations g, H, etc. The original was corrupted (Figure 4(a)) according to $y_s = (H(x^P)_s)^{1/2} \cdot \eta_s$, where H puts weight 1/2 on the central pixel and 1/16 on the eight nearest neighbors. The (multiplicative) noise has mean 1 and σ=.1; the model is the same as before. The restoration, Figure 4(b), is nearly perfect.

b) Infrared. The upper left panel in Figure 5 is an infrared picture. There is one vehicle, with engine running. The intensity data represents corrupted thermal radiation. As with most photoelectronic systems, the imaging system consists of an optical subsystem, arrays of detectors, and a scanner.

There are a number of sources of blur and noise. For example, there is "background noise" due to the fluctuations of black body radiation, noise in the conversion of photons to electric current, and digitization noise. In addition the detectors cause spatial and temporal blurring. Finally, there is attenuation and diffraction at the optical stage.

No effort has been made to model each of these effects, except to note that the model in (7) offers a good first approximation with appropriate choices of g, H and η. Instead, the picture was segmented and restored under the simple degradation model $y_s = x_s^P + \eta_s$, $\{\eta_s\}$ i.i.d. $N(\theta,\sigma^2)$. The variance, σ^2, was estimated from the raw grey-level data "by eye", to be 16. The Gibbs prior was on the pixel-edge process $\underset{\sim}{X} = (\underset{\sim}{X}^P, \underset{\sim}{X}^E)$,

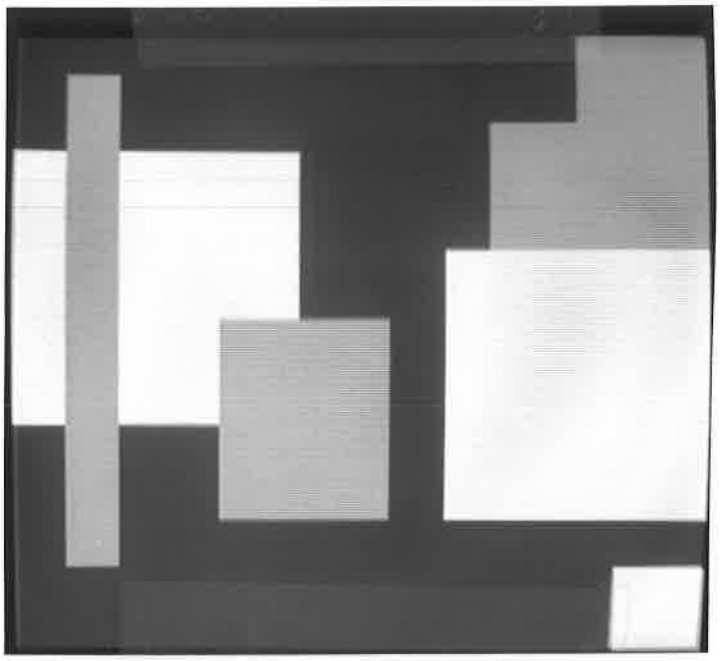
and the energy was a slight modification of (6), the modification allowing for the breaking of diagonal pixel bonds by the presence of suitable contiguous pairs of edges. The parameter values were chosen by "reparametrization" (see § VI), with the "scale" fixed after experimenting with a range of values.

The upper right panel of Figure 5 is the restored and segmented picture, via simulated annealing. The locations of edges (slightly displaced to coincide with pixel sites) are shown in black. The lower left panel is the original picture corrupted by adding zero mean white Gaussian noise, with variance 16. The same restoration/segmentation was applied to the corrupted picture, except that the assumed noise variance, σ^2, was adjusted for the additional degradation: $\sigma^2 = 16+16 = 32$. The result is the lower right panel.

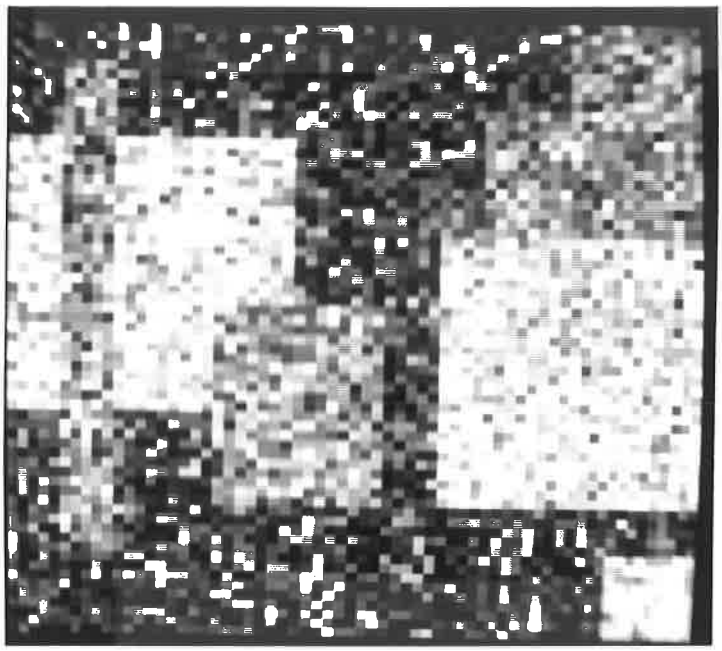
c) Tomography. The details of this experiment are in [9]. The object is to reconstruct the idealized isotope concentration shown in Figure 6(a). The observable photon counts $\{y_k\}$ were simulated in accordance with the degradation model given in (8). The detector geometry and assumed attenuation function are incorporated into A. Figure 6(b) shows the attenuation function, which is proportional to the probability of photon absorption per length of travel.

Reconstructions were generated under the prior in (1) with Φ as in (3), except that diagonal bonds were reduced by a factor of $1/\sqrt{2}$ (again, see [9] for the full story). Figures 6(c), 6(d), and 6(e) are approximate MAP esimators at θ=.25, θ=2.0, and θ=6.0, respectively, and C_1=.7, C_2=2. Obviously, the value of θ is important. We have begun to experiment with estimation of θ using the variations of EM discussed earlier (see § VI), and the preliminary results suggest that satisfactory estimates may be possible from single observations of the Poisson process (photon counts) $\underset{\sim}{Y}=\{y_k\}$.

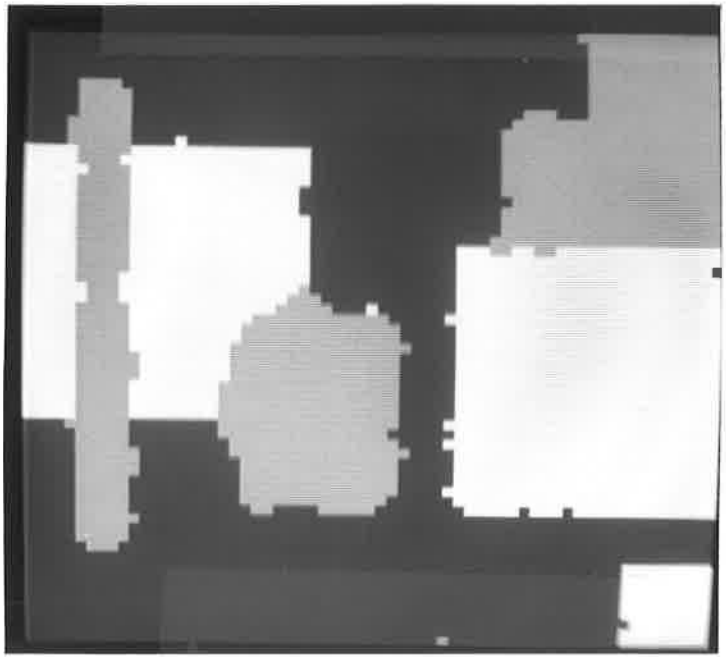
These reconstructions did not involve annealing. Instead, they were obtained by a simple gradient descent of the posterior energy, starting from the maximum likelihood reconstruction. The latter was achieved by an implementation of EM due to N. Accomando [1]. For comparison, the maximum likelihood reconstruction is shown in Figure 6(f).



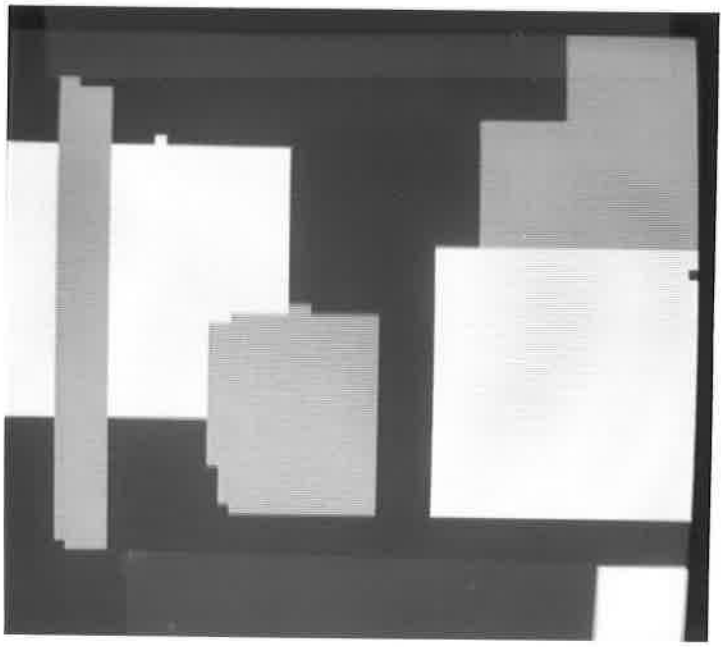
(a)
Original



(b)
Original corrupted by added noise

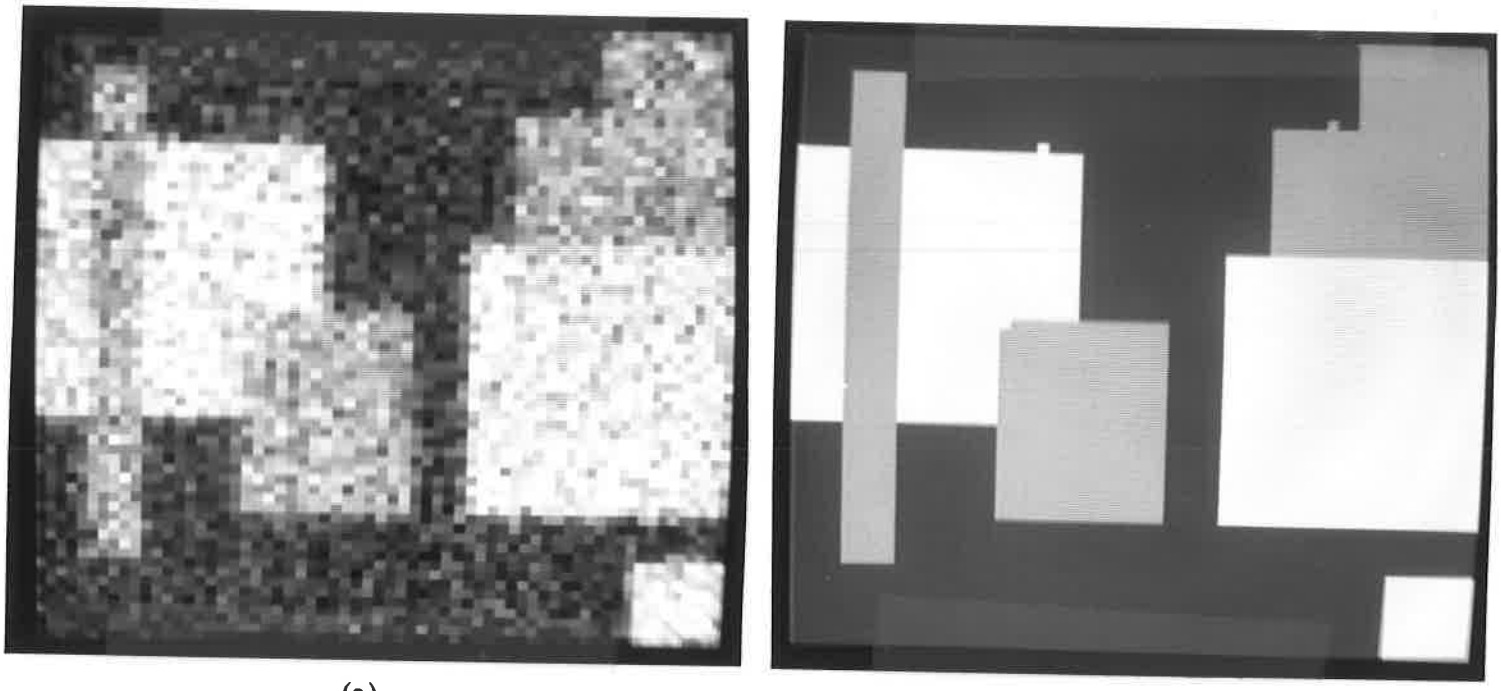


(c)
Restoration without edge process



(d)
Restoration with edge process

Figure 3



(a)

Original corrupted by blur, nonlinear transformation, and multiplicative noise

(b)

Restoration with edge process

Figure 4

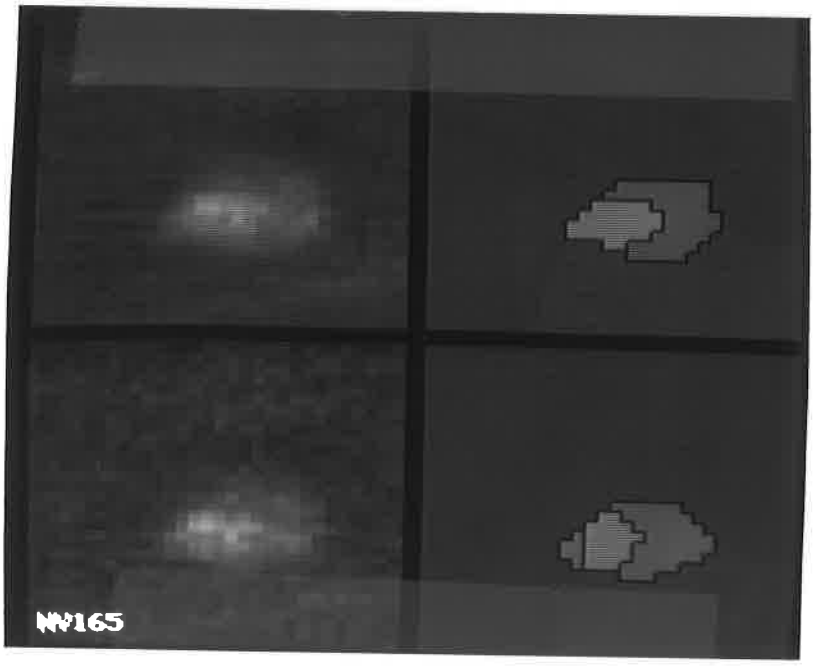
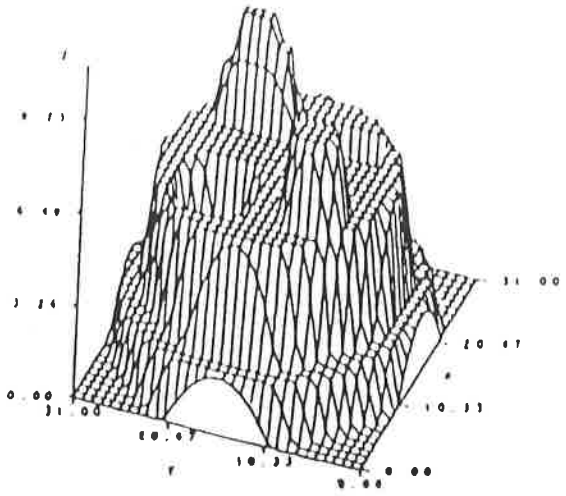
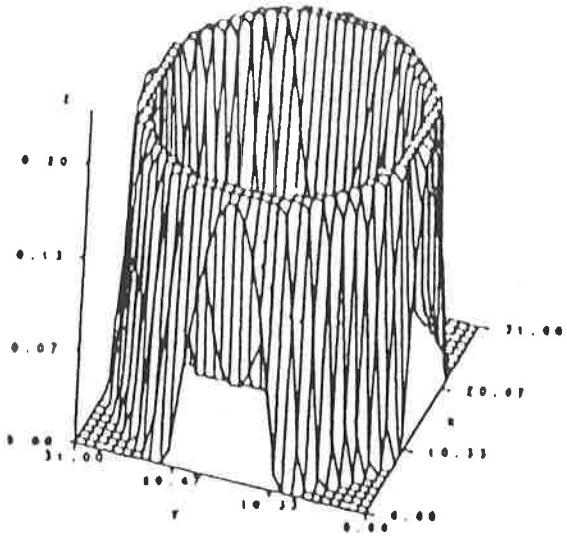


Figure 5

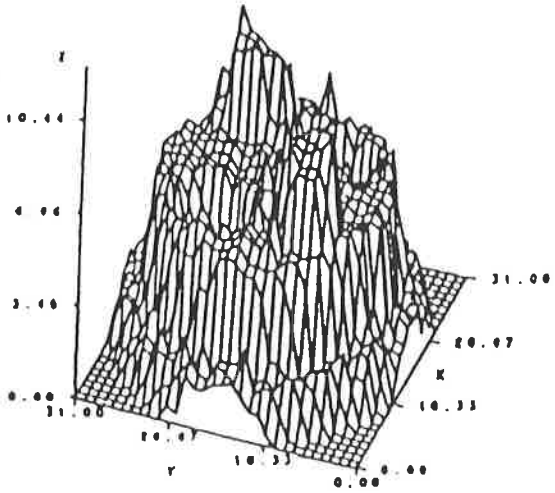
upper left: infrared image, including one vehicle with hot engine
upper right: original restored and segmented
lower left: original corrupted by added noise
lower right: corrupted restored and segmented



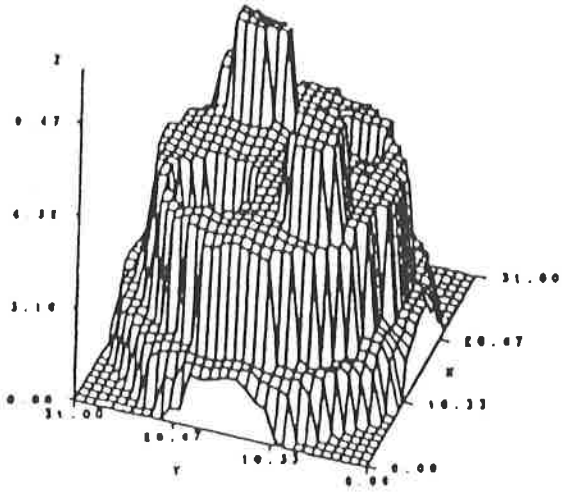
(a)
Idealized density



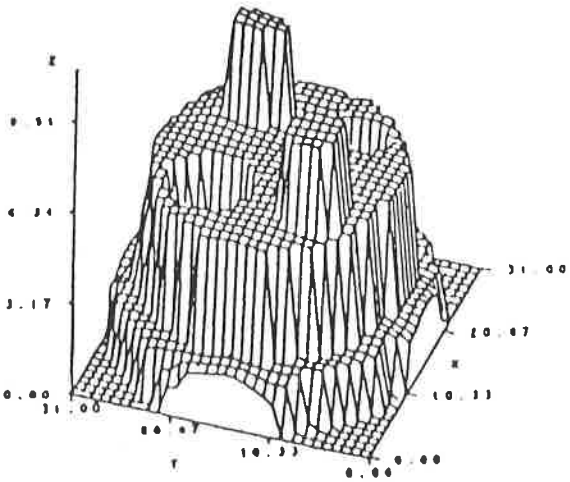
(b)
Attenuation



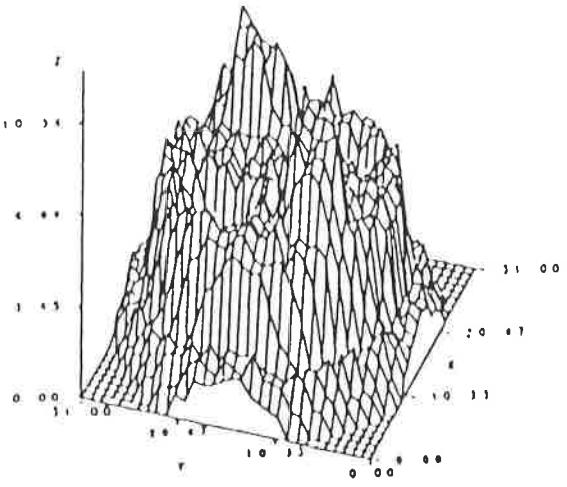
(c)
Reconstruction at θ = .25



(d)
Reconstruction at θ = 2.0



(e)
Reconstruction at θ = 6.0



(f)
EM Reconstruction

Figure 6

REFERENCES

1. N. Accomando, "Maximum likelihood reconstruction of a two dimensional Poisson intensity function from attenuated projections," Ph.D. thesis, Division of Applied Mathematics, Brown University, 1984.

2. J. Besag, "On the statistical analysis of dirty pictures," preprint, Department of Statistics, of Durhan, U.K., 1985.

3. V. Cêrný, "A thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm," preprint, Institue of Physics and Biophysics, Comenius Unive, Bratislava, 1982.

4. F.S. Cohen and D.B. Cooper, "Simple parallel hierarchical and relaxation algorithms for segmenting noncausal Markovian random fields," preprint, Brown University, 1984.

5. G.C. Cross and A.K. Jain, "Markov random field texture models," IEEE Transaction Pattern Analysis Machine Intelligence., vol. PAMI-5, pp.25-39, 1983.

6. P.A. Devijver, "Probabilistic labeling in a hidden second order markov mesh," technical report, Philips Research Laboratory, Brussels, Belgium, 1985.

7. H. Elliott and H. Derin, "Modelling and segmentation of noisy and textured images using Gibbs random fields," preprint, Department of Electrical and Computer Engineering, University of Massachusetts, 1984.

8. S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," IEEE Trans. Pattern Anal. Machine Intell., vol. PAMI-6, pp.721-741, 1984.

9. S. Geman and D.E. McClure, "Bayesian image analysis: An application to single photon emission tomography," 1985 Proceedings of the American Statistical Association. Statistical Computing Section. (to appear).

10. B. Gidas, "Nonstationary Markov chains and convergence of the annealing algorithm," Journal of Statistical Physics, Vol. 39, pp.73-131, 1985.

11. U. Grenander, "Tutorial in Pattern Theory," Division of Applied Mathematics, Brown University, 1984.

12. B. Hajek, "Cooling schedules for optimal annealing," preprint, Department of Electrical Engineering and the Coordinated Science Laboratory, University of Illinois at Champaign-Urbana, 1985.

13. A.R. Hanson and E.M. Riseman, "Segmentation of natural scenes," Computer Vision Systems. New York: Academic Press, 1978.

14. G.E. Hinton and T.J. Sejnowski, "Optimal perceptual inference," in Proceedings IEEE Conference Computer Vision Pattern Recognition, 1983.

15. J.J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," Proceedings of the National Academy of Sciences USA, Vol. 79, pp.2554-2558, 1982.

16. R.A. Hummel and S.W. Zucker, "On the foundations of relaxation labeling processes," IEEE Transaction Pattern Analysis Machine Intelligence, vol. PAMI-5, pp.267-287, 1983.

17. H.T. Kiiveri and N.A. Campbell, "Allocation of remotely sensed data using Markov models for spectral variables and pixel labels," preprint, CSIRO Division of Mathematics and Statistics, Sydney, Australia, 1985.

18. J. Kirkpatrick, C.D. Gelatt, Jr. and M.P. Vecchi, "Optimization by simulated annealing," IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y. 1982.

19. C. von der Malsburg, E. Bienenstock, "Statistical coding and short term synaptic plasticity: A scheme for knowledge representation in the brain," (this volume).

20. J.L. Marroquin, "Surface reconstruction preserving discontinuities," Laboratory for Information and Decision Systems, M.I.T., Cambridge, MA, 1984.