Proceedings of the International Congress of Mathematicians
Berkeley, California, USA, 1986

Markov Random Field Image Models and Their Applications to Computer Vision

STUART GEMAN AND CHRISTINE GRAFFIGNE

1. Introduction. Computer vision refers to a variety of applications involving a sensing device, a computer, and software for restoring and possibly interpreting the sensed data. Most commonly, visible light is sensed by a video camera and converted to an array of measured light intensities, each element corresponding to a small patch in the scene (a picture element, or "pixel"). The image is thereby "digitized," and this format is suitable for computer analysis. In some applications, the sensing mechanism responds to other forms of light, such as in infrared imaging where the camera is tuned to the invisible part of the spectrum neighboring the color red. Infrared light is emitted in proportion to temperature, and thus infrared imaging is suitable for detecting and analyzing the temperature profile of a scene. Applications include automated inspection in industrial settings, medical diagnosis, and targeting and tracking of military objects. In single photon emission tomography, as a diagnostic tool, individual photons, emitted from a "radiopharmaceutical" (isotope combined with a suitable pharmaceutical) are detected. The object is to reconstruct the distribution of isotope density inside the body from the externally-collected counts. Depending on the pharmaceutical, the isotope density may correspond to local blood flow ("perfusion") or local metabolic activity. Other applications of computer vision include satellite imaging for weather and crop yield prediction, radar imaging in military applications, ultrasonic imaging for industrial inspection and a host of medical applications, and there is a growing role for video imaging in robotics.

The variety of applications has yielded an equal variety of algorithms for restoration and interpretation. Unfortunately, few general principals have emerged and no common foundation has been layed. Algorithms are by and large ad hoc; they are typically dedicated to a single application, and often critically tuned to the particulars of the environment (lighting, weather conditions, magnification, and so on) in which they are implemented. It is likely that a

Research partially supported by Army Research Office Contract DAAG29-83-K-0116, National Science Foundation Grant DMS-8352087, and the General Motors Corporation.

coherent theoretical framework would support more robust and more powerful algorithms. We have been exploring an approach based upon probabilistic image models, well-defined principals of inference, and a Monte Carlo computation theory. Exploiting this framework, we have recently obtained encouraging results in several areas of application, including tomography, texture analysis, and scene segmentation.

As an illustration of our approach, we shall discuss here the application to texture analysis. Other applications, and more complete discussions of the foundations, can be found in [**1**, **3**, **4**, **10**, **12**, **13**, **14**, **17**, **18**, **23**, **25**, and **27**]. In the section that follows, §2, we lay out, briefly, our paradigm in its general formulation. Then, in §3, the application to texture analysis is developed and illustrated by computer experiments. This application requires that we treat a somewhat unusual problem in parameter estimation, namely the estimation of parameters of a Markov random field from a single, large, sample. §4 details the estimation method used, and provides a proof of its consistency in the "large picture" limit, which is more appropriate than the usual "large sample size" limit.

2. Bayesian paradigm. In real scenes, neighboring pixels typically have similar intensities, boundaries are usually smooth and often straight, textures, although sometimes random locally, define spatially homogeneous regions, and objects, such as grass, tree trunks, branches and leaves, have preferred relations and orientations. Our approach to picture processing is to articulate such regularities mathematically, and then to exploit them in a statistical framework to make inferences. The regularities are rarely deterministic; instead, they describe correlations and likelihoods. This leads us to the Bayesian formulation, in which prior expectations are formally represented by a probability distribution. Thus we design a distribution (a "prior") on relevant scene attributes to capture the tendencies and constraints that characterize the scenes of interest. Picture processing is then guided by this prior distribution, which, if properly conceived, enormously limits the plausible restorations and interpretations.

The approach involves five steps, which we shall briefly review here (see [**13** and **18**] for more details). This will define the general framework, and then, in the following sections, we will concentrate on the analysis of texture, as an illustrative application.

Image models. These are probability distributions on relevant image attributes. Both for reasons of mathematical and computational convenience, we use *Markov random fields* (MRF) as prior probability distributions. Let us suppose that we index all of the relevant attributes by the index set S. S is application specific. It typically includes indices for each of the pixels (about 512×512 in the usual video digitization) and may have other indices for such attributes as boundary elements, texture labels, object labels on so on. Associated with each "site" $s \in S$ is a real-valued random variable X_s, representing the state of the corresponding attribute. Thus X_s may be the measured intensity at pixel s

(typically, $X_s \in \{0, \ldots, 255\}$) or simply 1 or 0 as a boundary element at location s is present or absent.

The kind of knowledge we represent by the prior distribution is usually "local," which is to say that we articulate regularities in terms of small local collections of variables. In the end, this leads to a distribution on $X = \{X_s\}_{s \in S}$ with a more or less "local neighborhood structure" (again, we refer to [**13** and **18**] for details). Specifically, our priors are Markov random fields: there exists a (symmetric) *neighborhood relation* $G = \{G_s\}_{s \in S}$, wherein $G_s \subseteq S$ is the set of neighbors of s, such that

$$\Pi(X_s = x_s | X_r = x_r, r \in S, r \neq s) = \Pi(X_s = x_s | X_r = x_r, r \in G_s).$$

$\Pi(a|b)$ is conditional probability, and, by convention, $s \notin G_s$. G symmetric means $s \in G_r \Leftrightarrow r \in G_s$. (Here, we assume that the range of the random vector X is discrete; there are obvious modifications for the continuous or mixed case.)

It is well known, and very convenient, that a distribution Π defines a MRF on S with neighborhood relation G if and only if it is Gibbs with respect to the same graph, (S, G). The latter means that Π has the representation

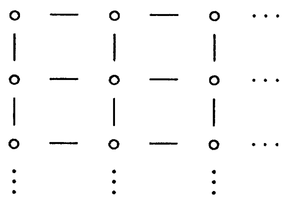
$$\Pi(x) = \tfrac{1}{z} e^{-U(x)} \tag{2.1}$$

where

$$U(x) = \sum_{c \in C} V_c(x). \tag{2.2}$$

C is the collection of all cliques in (S, G) (collections of sites such that every two sites are neighbors), and $V_c(x)$ is a function depending only on $\{x_s\}_{s \in c}$. U is known as the "energy," and has the intuitive property that the low energy states are the more likely states under Π. The normalizing constant, z, is known as the "partition function." The Gibbs distribution arises in statistical mechanics as the equilibrium distribution of a system with energy function U.

As a simple example (too simple to be of much use for real pictures) suppose the pixel intensities are known, a priori, to be one of two levels, minus one ("black") or plus one ("white"). Let S be the $N \times N$ square lattice, and let G be the neighborhood system that corresponds to nearest horizontal and vertical neighbors:



For picture processing, think of N as typically 512. Suppose that the only relevant regularity is that neighboring pixels tend to have the same intensities. An "energy" consistent with this regularity is the "Ising" potential:

$$U(x) = -\beta \sum_{\langle s,t \rangle} x_s x_t, \qquad \beta > 0,$$

where $\sum_{\langle s,t\rangle}$ means summation over all neighboring pairs $s, t \in S$. The minimum of U is achieved when $x_s = x_t$, $\forall s, t \in S$. Under (2.1), the likely pictures are therefore the ones that respect our prior expectations; they segment into regions of constant intensities. The larger β, the larger the typical region. Later we will discuss the issue of *estimating* model parameters such as β. (With energy (2.2), Π in (2.1) is called the Ising model. It models the equilibrium distribution of the spin states of the atoms in a ferromagnet. These spins tend to "line up," and hence the favored configurations contain connected regions of constant spins.)

One very good reason for using MRF priors is their Gibbs representations. Gibbs distributions are characterized by their energy functions, and these are more convenient and intuitive for modelling than working directly with probabilities. See, for example, [**12**, **13**, **14**, **18**, and **23**] for many more examples, and §3 below for a more complex and useful MRF model.

Degradation model. The image model is a distribution $\Pi(\cdot)$ on the vector of image attributes $X = \{X_s\}_{s\in S}$. *By design*, the components of this vector contain all of the relevant information for the image processing task at hand. Hence, the goal is to estimate X. This estimation will be based upon partial or corrupted observations, and based upon the prior information. In emission tomography, X represents the spacial distribution of isotope in a target region of the body. What is actually observed is a collection of photon counts whose probability law is Poisson, with a mean function that is an attenuated radon transform of X. In the texture labelling problem, X is the pixel intensity array and a corresponding array of texture labels. Each label gives the texture type of the associated pixel. The observation is only partial: we observe the pixels, which are just the digitized picture, but not the labels. The purpose is then to estimate the labels from the picture.

The observations are related to the image process (X) by a *degradation model.* This models the relation between X and the *observation process*, say $Y = \{Y_s\}_{s\in T}$. For texture analysis, we will define $X = (X^P, X^L)$, where X^P is the usual grey-level pixel intensity process, and X^L is an associated array of texture labels. The observed picture is just X^P, and hence $Y = X^P$: the degradation is a projection. More typically, the degradation involves a random component, as in the tomography setting where the observations are Poisson variables whose means are related to the image process X. A more simple, and widely studied (if unrealistic) example is additive "white" noise. Let $X = \{X_s\}_{s\in S}$ be just the basic pixel process. In this case, $T = S$, and for each $s \in S$ we observe $Y_s = X_s + \eta_s$ where, for example, $\{\eta_s\}_{s\in S}$ is Gaussian with independent components, having means 0 and variances σ^2.

Formally, the degradation model is a conditional probability distribution, or density, for Y given X: $\Pi(y|x)$. If the degradation is just added "white noise," as in the above example, then

$$\Pi(y|x) = \left(\frac{1}{2\pi\sigma^2}\right)^{|s|/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{s\in S}(y_s - x_s)^2\right\}.$$

For labelling textures, the degradation is deterministic: $\Pi(y|x)$ is concentrated on $y = x^P$, where $x = (x^P, x^L)$ has both pixel and label components.

Posterior distribution. This is the conditional distribution on the image process X given the observation process Y. This "posterior" or "a posteriori" distribution contains the information relevant to the image restoration or image analysis task. Given an observation $Y = y$, and assuming the image model $(\Pi(x))$ and degradation model $(\Pi(y|x))$, the posterior distribution reveals the likely and unlikely states of the "true" (unobserved) image X. Having constructed X to contain all relevant image attributes, such as locations of boundaries, labels of objects or textures, and so on, the posterior distribution comes to play the fundamental role in our approach to image processing.

The posterior distribution is easily derived from "Bayes's rule":

$$\Pi(x|y) = \frac{\Pi(y|x)\Pi(x)}{\Pi(y)}.$$

The denominator, $\Pi(y)$, is difficult to evaluate. It derives from the prior and degradation models by integration: $\Pi(y) = \int_x \Pi(y|x)\Pi(dx)$, but the formula is computationally intractable. Happily, our analysis of the posterior distribution will require only *ratios*, not absolute probabilities. Since y is fixed by observation, $1/\Pi(y)$ is a constant that can be ignored (see paragraph below on "computing").

As an example we consider the simple "Ising model" prior, with observations corrupted by additive white noise. Then

$$\Pi(x) = \frac{1}{z} \exp\left\{ -\beta \sum_{\langle s,t \rangle} x_s x_t \right\}$$

and

$$\Pi(y|x) = \left(\frac{1}{2\pi\sigma^2}\right)^{|S|/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{s \in S} (y_s - x_s)^2 \right\}.$$

The posterior distribution is then

$$\Pi(x|y) = \frac{1}{z_p} \exp\left\{ -\beta \sum_{\langle s,t \rangle} x_s x_t - \frac{1}{2\sigma^2} \sum_{s \in S} (y_s - x_s)^2 \right\}.$$

We denote by z_p the normalizing constant for the posterior distribution. Of course, it depends upon y, but the latter is fixed. Notice that the posterior distribution is again a MRF. In the case of additive white noise, the neighborhood system of the posterior distribution is that of the prior, and hence local. For a wide class of useful degradation models, including combinations of blur, added or multiplicative "colored noise," and a variety of nonlinear transformations, the posterior distribution is a MRF with a more or less local graph structure. This is convenient for our computational schemes, as we shall see shortly. We should note, however, that exceptions occur. In tomography, for example, the posterior distribution is associated with a highly nonlocal graph. This situation incurs a high computational cost (see [**14**] for more details).

MAP estimate. In our framework, image processing amounts to choosing a particular image x, given an observation $Y = y$. A sensible, and suitably-defined optimal, choice is the "maximum a posteriori," or "MAP" estimate: choose x to maximize $\Pi(x|y)$. The MAP estimate chooses the most likely x, given the observation. In most applications, our goal is to identify the MAP estimate, or a suitable approximation. However, in some settings other estimators are more appropriate. We have found, for example, that the posterior mean $(\int x\Pi(dx|y))$ is more effective for tomography, at least in our experiments. Here, we concentrate on MAP estimation.

In most applications we cannot hope to identify the true maximum a posteriori image vector x. To appreciate the computational difficulty, consider again the Ising model with added white noise:

$$\Pi(x|y) = \frac{1}{z_p} \exp\left\{ -\beta \sum_{\langle s,t \rangle} x_s x_t - \frac{1}{2\sigma^2} \sum_{s \in S} (y_s - x_s)^2 \right\}. \tag{2.3}$$

This is to be maximized over all possible vectors $x = \{x_s\}_{s \in S} \in \{-1, 1\}^{|S|}$. with $|S| \sim 10^5$, brute force approaches are intractable; instead, we will employ a Monte Carlo algorithm which gives adequate approximations.

Maximizing (2.3) amounts to minimizing

$$U_p(x) = -\beta \sum_{\langle s,t \rangle} x_s x_t - \frac{1}{2\sigma^2} \sum_{s \in S} (y_s - x_s)^2$$

which might be thought of as the "posterior energy." (As with z_p, the fixed observation y is suppressed in the notation $U_p(x)$.) More generally, we write the posterior distribution as

$$\frac{1}{z_p} \exp\{-U_p(x)\} \tag{2.4}$$

and characterize the MAP estimator as the solution to the problem "choose x to minimize $U_p(x)$." The utility of this point of view is that it suggests a further analogy to statistical mechanics, and a computation scheme for approximating the MAP estimate, which we shall now describe.

Computing. Pretend that (2.4) is the equilibrium Gibbs distribution of a real system. Recall that MAP estimation amounts to finding a minimal energy state. For many physical systems the low energy states are the most ordered, and these often have desirable properties. The state of silicon suitable for wafer manufacturing, for example, is a low energy state. Physical chemists achieve low energy states by heating and then slowly cooling a substance. This procedure is called *annealing.* Cerný [**5**] and Kirkpatrick [**21**] suggest searching for good minimizers of $U(\cdot)$ by *simulating* the dynamics of annealing, with U playing the role of energy for an (imagined) physical system. In our image processing experiments, we often use "simulated annealing" to find an approximation to the MAP estimator.

Dynamics are simulated by producing a Markov chain, $X(1), X(2), \ldots$ with transition probabilities chosen so that the equilibrium distribution is the posterior (Gibbs) distribution (2.4). One way to do this is with the "Metropolis algorithm" [**24**]. More convenient for image processing is a variation we call *stochastic relaxation.* The full story can be found in [**13** and **18**]. Briefly, in stochastic relaxation we choose a sequence of sites $s(1), s(2), \ldots \in S$ such that each site in S is "visited" infinitely often. If $X(t) = x$, say, then $X_r(t+1) = x_r$, $\forall r \neq s(t)$, $r \in S$, and $X_{s(t)}(t+1)$ is a sample from

$$\Pi(X_{s(t)} = \cdot | X_r = x_r, r \neq s(t)),$$

the conditional distribution on $X_{s(t)}$ given $X_r = x_r$ $\forall r \neq s(t)$. By the Markov property,

$$\Pi(X_{s(t)} = \cdot | X_r = x_r, r \neq s(t)) = \Pi(X_{s(t)} = \cdot | X_r = x_r, r \in G^p_{s(t)})$$

where $\{G^p_s\}_{s \in S}$ is the *posterior* neighborhood system, determined by the posterior energy $U_p(\cdot)$. The prior distributions that we have experimented with have mostly had local neighborhood systems, and usually the posterior neighborhood system is also more or less local as well. This means that $|G^p_{s(t)}|$ is small, and this makes it relatively easy to generate, Monte Carlo, $X(t+1)$ from $X(t)$. In fact, if Ω is the range of $X_{s(t)}$, then

$$\Pi(X_{s(t)} = \alpha | X_r = x_r, r \in G^p_{s(t)}) = \frac{\Pi(\alpha,_{s(t)} x)}{\sum_{\hat{\alpha} \in \Omega} \Pi(\hat{\alpha},_{s(t)} x)} \tag{2.5}$$

where

$$(\alpha,_{s(t)} x)_r = \begin{cases} \alpha, & r = s(t), \\ x_r, & r \neq s(t). \end{cases}$$

Notice that (fortunately!) there is no need to compute the posterior partition function z_p. Also, the expression on the right-hand side of (2.5) involves only those potential terms associated with cliques containing $s(t)$, since all other terms are the same in the numerator and the denominator.

To simulate annealing, we introduce an artificial "temperature" into the posterior distribution:

$$\Pi_T(x) = \frac{\exp\{-U_p(x)/T\}}{Z_p(T)}.$$

As $T \to 0, \Pi_T(\cdot)$ concentrates on low energy states of U_p. To actually find these states, we run the stochastic relaxation algorithm while slowly lowering the temperature. Thus $T = T(t)$, and $T(t) \downarrow 0$. $\Pi_{T(t)}(\cdot)$ replaces $\Pi(\cdot)$ in computing the transition $X(t) \to X(t+1)$. In [**13**] we showed that, under suitable hypotheses on the sequence of site visits, $s(1), s(2), \ldots$:

> If $T(t) > c/(1 + \log(1+t))$, $T(t) \downarrow 0$, then for all c sufficiently large $X(t)$ converges weakly to the distribution concentrating uniformly on $\{x \colon U(x) = \min_y U(y)\}$.

More recently, our theorem has been improved upon by many authors. In particular, the smallest constant c which guarantees convergence of the annealing algorithm to a global minimum can be specified in terms of the energy function U_p (see [**15** and **19**]). Also, see Gidas [**16**] for some ideas about faster annealing via "renormalization group" methods.

In the experiments with texture to be described here, MAP estimates are approximated by using the annealing algorithm. This involves Monte Carlo computer-generation of the sequence $X(1), X(2), \dots$, terminating when the state ceases to change substantially.

3. Texture segmentation. Texture *synthesis* refers to computer generation of homogeneous patterns, usually intended to match a natural texture such as wood, grass, or sand. In many instances, Markov random fields provide good models, and Metropolis-like Monte Carlo methods yield respectable facsimiles of the real textures [**8**, **9**]. Here we combine MRF texture models, for the pixel process, with an Ising-like "texture label process," in order to segment and label a scene consisting of patches of natural textures. The image model thereby involves both a pixel process, of grey level intensities, and a label process, whose components identify the texture type of each picture element in the scene. Our approach is similar to those of Derin and Elliott [**9**] and Cohen and Cooper [**7**], especially in our use of the two-tiered image model.

Image model. The image process comprises a pixel process and a label process, $X = \{X^P, X^L\}$. As usual, the pixes sites form an $N \times N$ square lattice, say S^P. For each pixel site there is a corresponding label site, and thus the graph associated with the image model has sites $S = S^P \cup S^L$, where S^L is just a copy of S^P. The elements of S^P and S^L index the components of X^P and X^L, respectively, so that $X^P = \{X_s^P\}_{s \in S^P}$ and $X^L = \{X_s^L\}_{s \in S^L}$. In the experiments reported here, the pixels were allowed sixteen possible grey levels $X_s^P \in \{0, 1, \dots, 15\}$, $\forall s \in S^P$, whereas the range of the labels depended upon the actual number of textures in the scene, thus assuming this number to be known a priori. Let M be the number of textures that are to be modelled. Then $X_s^L \in \{1, 2, \dots, M\}$, $\forall s \in S^L$.

We shall develop the image model by first assuming that the texture type is fixed, say "l" and constant over the scene. *Conditioned* on $X_s^L = l \in \{1, 2, \dots, M\}$, $\forall s \in S^L$, the process X^P is a Markov random field:

$$\Pi(x^P | X_s^L = l, s \in S^L) = \frac{1}{z^{(l)}} \exp\{-U^{(l)}(x^P)\}$$

where $z^{(l)}$ is the usual normalizing constant $z^{(l)} = \sum_{x^P} \exp\{-U^{(l)}(x^P)\}$. Only pair-cliques appear in the energy $U^{(l)}$. There are six types of pair-cliques, as shown in Figure 1. These we index by $i \in \{1, 2, 3, 4, 5, 6\}$. We denote by $\langle s, t \rangle_i$ a pair of sites s, t which form a type i clique, and by $\sum_{\langle s,t \rangle_i}$ the summation over all such pairs. With these conventions, the (conditional) energy is

$$U^{(l)}(x^P) = -\sum_{i=1}^{6} \sum_{\langle s,t \rangle_i} \theta_i^{(l)} \Phi(x_s^P - x_t^P), \qquad \Phi(\Delta) \doteq (1 + (|\Delta|/\delta)^2)^{-1} \tag{3.1}$$

for some fixed $\delta > 0$. Notice that $\Phi(x_s^P - x_t^P)$ is larger when $x_s^P = x_t^P$, and is monotonic in $|x_s^P - x_t^P|$. Because of this, the texture-dependent parameters $\theta_1^{(l)}, \ldots, \theta_6^{(l)}$ determine the degree to which neighboring pixels, of a particular type of pair-clique, will tend to have similar grey-levels. In face, if $\theta_i^{(l)} > 0$, then for texture "l" we expect pixel pairs x_s and x_t, of clique type i, to typically have similar intensities. If $\theta_1^{(l)} < 0$ then the tendency is to be different. Of course, these simple rules are complicated by the actions of the other five types of pair-cliques.

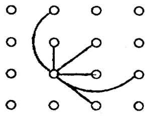


FIGURE 1. Pair-cliques for texture model.

The parameters $\theta_i^{(l)}, i = 1, 2, \ldots, 6$, $l = 1, 2, \ldots, M$, are estimated from pictures of the M textures, as explained in the following section (§4). On the other hand, Φ, and indeed the neighborhood structure, is ad hoc. We have used Φ extensively in other applications in which our main concern is with the difference of intensities between neighboring pixels. Of course the quadratic, $\Phi(\Delta) = \Delta^2$, is simpler, but it unduly penalizes large differences. Having modeled the M textures, we now construct a composite Markov random field which accounts for both texture labels, $X^L = \{X_s^L, s \in S^L\}$, and grey-levels, $X^P = \{X_s^P, s \in S^P\}$. The joint distribution is

$$\Pi(X^P = x^P, X^L = x^L) = \frac{\exp\{-U_1(x^P, x^L) - U_2(x^L)\}}{z} \tag{3.2}$$

in which U_2 promotes label bonding (we expect the textures to appear in patches rather than interspersed) and U_1 specifies the interaction between labels and intensities. Specifically, we employ a simple Ising-type potential for the labels:

$$U_2(x^L) = -\beta \sum_{[s,t]} 1_{x_s^L = x_t^L} + \sum_{s \in S} w(x_s^L), \qquad \beta > 0. \tag{3.3}$$

Here β determines the degree of clustering, $[s,t]$ indicates a pair of nearest horizontal or vertical neighbors, and $w(\cdot)$ is adjusted to eliminate bias in the label probabilities (more on the choice of $w(\cdot)$ later).

To describe the interaction between labels and pixels we introduce the symbols $\tau_1, \tau_2, \ldots, \tau_6$ to represent the lattice vectors associated with the 6 pair-cliques (Figure 1). Thus s and $s + \tau_i$ are neighbors, constituting a pair with clique type i. The interaction is then given in terms of pixel-based contributions,

$$H(x^P, l, s) \doteq -\sum_{i=1}^{6} \theta_i^{(l)} \{\Phi(x_s^P - x_{s+\tau_i}^P) + \Phi(x_s^P - x_{s-\tau_i}^P)\} \tag{3.4}$$

and local sums of these called block-based contributions,

$$Z(x^P, l, s) \doteq \frac{1}{a} \sum_{t \in N_s} H(x^P, l, t). \tag{3.5}$$

Here, N_s is a block of sites centered at s (5 by 5 in all of our experiments), and the constant a is adjusted so that the sum of all block-based contributions reduces to $U^{(l)}$ (see (3.1)):

$$U^{(l)}(x^P) = \sum_{s\in S} Z(x^P, l, s). \tag{3.6}$$

This amounts to ensuring that each pair-clique appears exactly once ($a = 50$, for example, when N_s is 5 by 5). In terms of (3.4) and (3.5), the "interaction energy," $U_1(x^P, x^L)$, is written

$$U_1(x^P, x^L) = \sum_{s\in S} Z(x^P, x_s^L, s). \tag{3.7}$$

Because of (3.6), the model is consistent with (3.1) for homogeneous textures, $X_s^L = l$, $\forall s \in S$. The idea is that each local texture label, X_s^L, is influenced by the pixel grey levels in a neighborhood of s.

Finally, to clarify the bias correction term $w(\cdot)$, we briefly examine the local characteristics of the field, specifically the conditional distributions for the labels given all the intensity data and the values of the neighboring labels. (The actual neighborhoods of the Markov random field corresponding to (3.2) can be easily inferred from (3.3) and (3.7).) The log odds of texture type k to type j is

$$\begin{aligned}
&\log\left\{\frac{\Pi(X_r^L = k \mid X_s^L = x_s^L,\ s \neq r;\ X_s^P = x_s^P,\ s\in S)}{\Pi(X_r^L = j \mid X_s^L = x_s^L,\ s \neq r;\ X_s^P = x_s^P,\ s\in S)}\right\}\\
&\quad = Z(x^P, j, r) - Z(x^P, k, r) + \beta \sum_{t:\ [t,r]} (1_{x_t^L = k} - 1_{x_t^L = j}) + w(j) - w(k)\\
&\quad = \frac{1}{z}\sum_{i=1}^{6}\sum_{s\in N_r} (\theta_i^{(k)} - \theta_i^{(j)})\{\Phi(x_s^P - x_{s+r_i}^P) + \Phi(x_s^P - x_{s-r_i}^P)\}\\
&\qquad + \beta \sum_{t:\ [t,r]} (1_{x_t^L = k} - 1_{x_t^L = j}) + w(j) - w(k).
\end{aligned}$$

The first term imposes fidelity to the "data" x^P, and the second bonds the labels. The efficacy of the model depends on the extent to which the first term separates the two types k and j, which can be assessed by plotting histograms for the values of this quantity both for pure k and pure j data. A clean separation of the histograms signifies a good discriminator. However, since we are looking at log odds, we insist that the histograms straddle the origin, with positive (resp. negative) values associated with texture type k (resp. j). The function $w(\cdot)$ makes this adjustment.

Degradation model. The degradation is deterministic. The observation process is the pixel process $Y = X^P$, and hence the degradation is just the *projection* $(X^P, X^L) \to X^P$.

Posterior distribution. In this special case, the posterior energy is the same as the prior energy, but some of the components are fixed. In particular,

$$\Pi((x^P, x^L)|y) = \frac{1}{z_p}\exp\{-U_1(x^P, x^L) - U_2(x^L)\} 1_{x^P = y}.$$

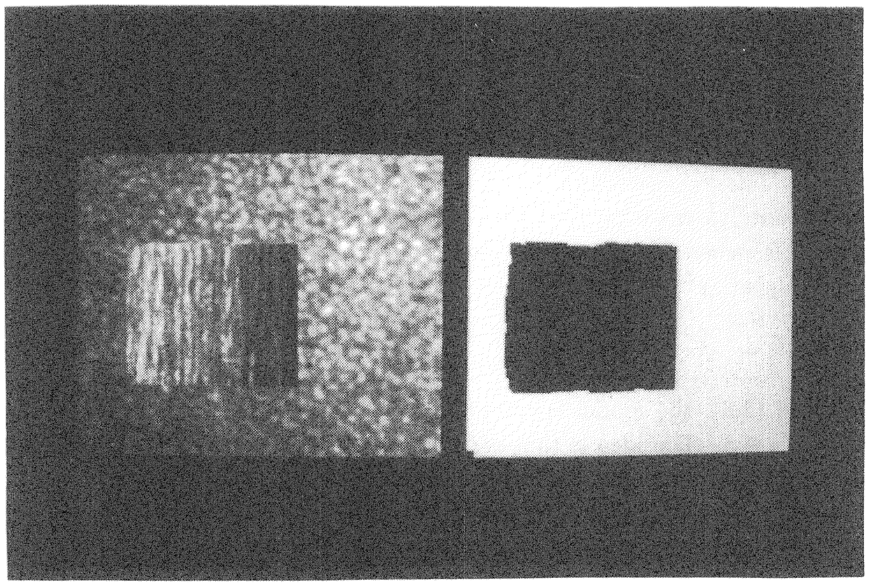


FIGURE 2. Wood on plastic background.

Equivalently, we simply use $\Pi(x^L|x^P)$ as the posterior distribution:

$$\Pi(x^L|x^P) = \frac{1}{z_p}\exp\{-U_1(x^P, x^L) - U_2(x^L)\}.$$

MAP estimate. Given an observation, $X^P = x^P$, we shall seek x^L to minimize $U_1(x^P, x^L) + U_2(x^L)$.

Computing. We use stochastic relaxation, with simulated annealing, as described in §2. A convenient starting point is arrived at by "turning off" the Ising term in the label model (3.3): we set $\beta = 0$. Since this is the only label/label interaction term in the model, the MAP estimate of x^L, with $\beta = 0$, is determined by (locally) optimizing x_s^L at each $s \in S^L$. The computation time is negligible. Thereafter, we set β to the model value (see §4) and begin stochastic relaxation. In the experiments, each site was visited about 150 times.

Experimental results. Three experiments were done on texture discrimination, based on two images with two textures each and one with four. There are four textures involved: wood, plastic, carpet, and cloth. As mentioned above, the parameters were estimated from the pure types (see §4). There was no pre- or post-processing. In particular, no effort was made to "clean-up" the boundaries, expecting smooth transitions. The results are shown in Figures 2, 3, and 4; these correspond to (i) wood on plastic, (ii) carpet on plastic, and (iii) wood, carpet, and cloth on plastic background. In each figure, the left panel is the textured scene, and the right panel shows the segmentation, with texture labels coded by grey level. It is interesting to note that the grey-level histograms of the four textures are very similar (Figure 5); in particular, discrimination based on shading alone is virtually impossible.

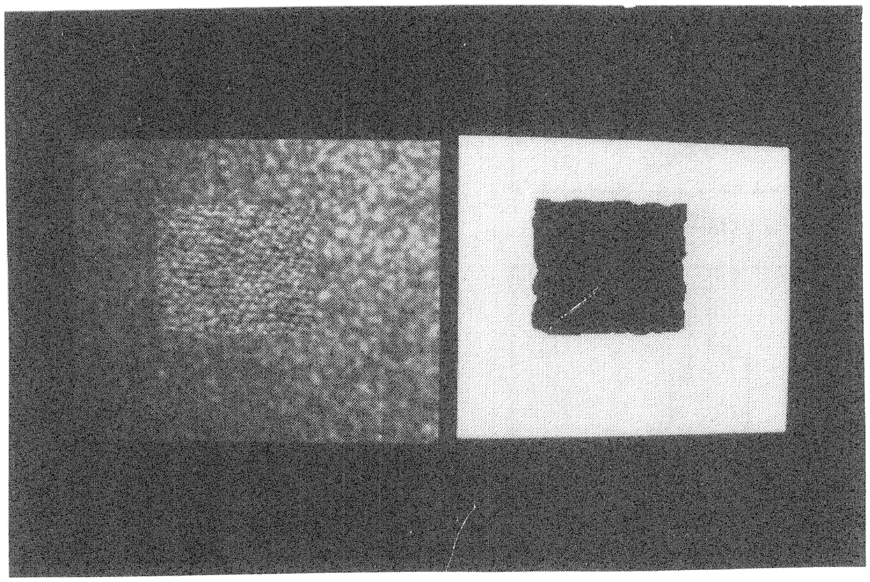


FIGURE 3. Carpet on plastic background.

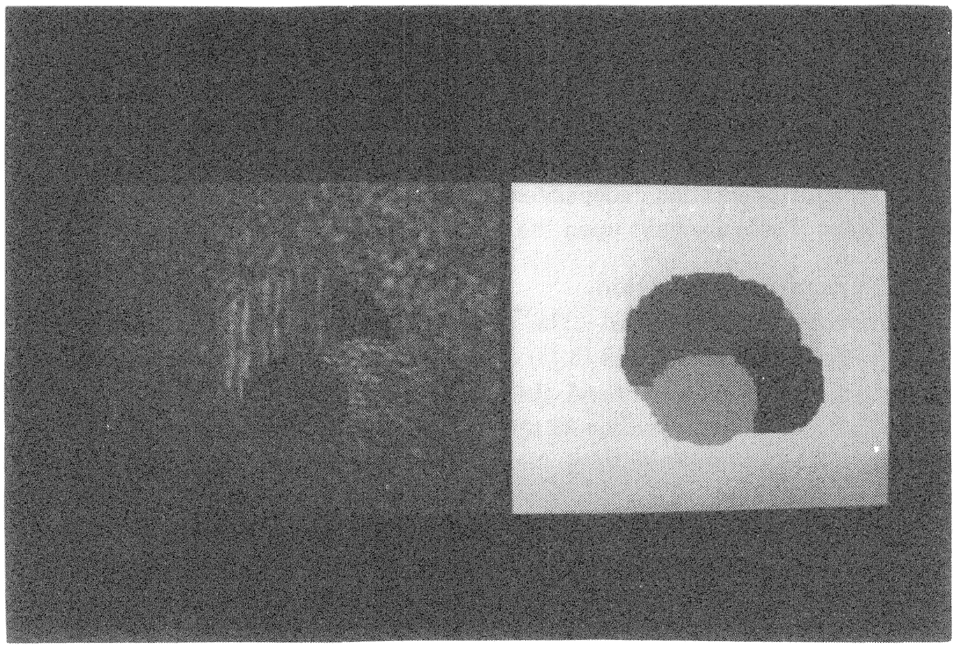
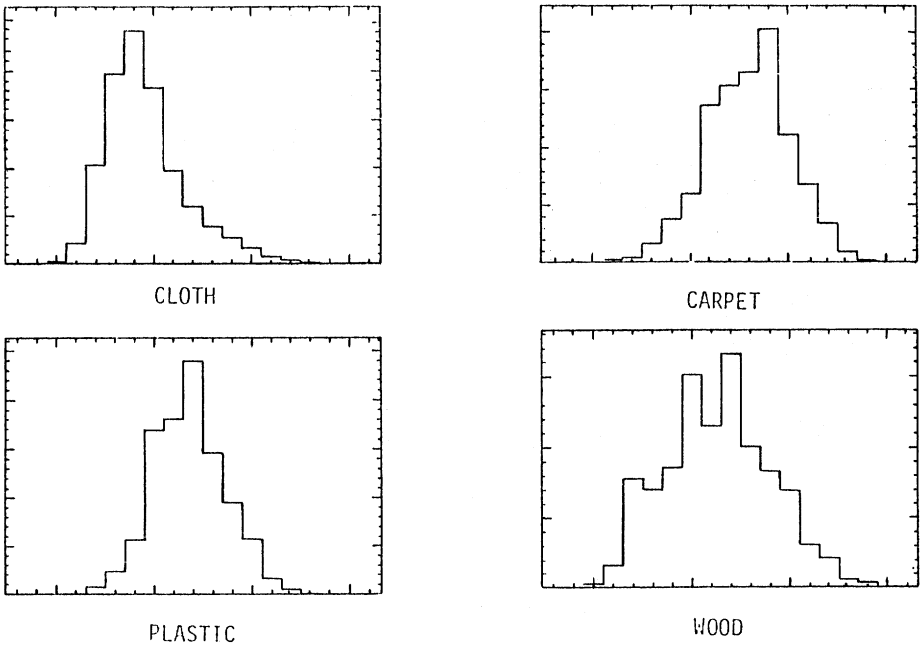


FIGURE 4. Wood, carpet, and cloth on plastic background.



CLOTH

CARPET

PLASTIC

WOOD

FIGURE 5. Grey-level histograms.

The model is not really adequate for texture *synthesis*; samples generated from the model do not resemble the texture very well. Evidently, the utility of Markov random field models does not depend on their capacity for simulating real-world imagery. A more serious drawback of our model is that it is dedicated to a fixed repertoire of textures, viewed at a particular orientation and at a particular magnification, or range. The problem is easier if the goal is merely *segmentation*, without *recognition*. We are experimenting with segmentation algorithms that are scale and orientation independent. Indeed, there are no texture-specific parameters. These are built upon the same modelling/computing framework.

4. Parameter estimation.

Maximum pseudolikelihood. The performance of the model is not unduly sensitive to the choice of δ (see (3.1)) or β (see (3.3)), which were determined by trial and error. On the other hand, the pair-clique parameters $\theta_i^{(l)}$, $i = 1, 2, \ldots, 6$, $l = 1, 2, \ldots, M$, characterize the M textures, and critically determine the ability of the model to segment and label. Needless to say, these must be systematically estimated. Trial and error is not feasible.

We have estimated the parameters from samples of the M textures. These "training samples" contain only one texture each, and we used just one sample for each texture. For a fixed texture, say wood, and from a single sample, say $\tilde{x}^p$, the problem then is to estimate $\theta_1, \theta_2, \ldots, \theta_6$ in the model

$$\Pi(X^P = x^P; \theta) = \frac{\exp\{-U(x^P; \theta)\}}{z(\theta)}$$

where

$$U(x^P;\theta) = -\sum_{i=1}^{6} \sum_{\langle s,t\rangle_i} \theta_i \Phi(x_s^P - x_t^P)$$

and

$$z(\theta) = \sum_{x^P} \exp\{-U(x^P;\theta)\}.$$

(We include $\theta \doteq (\theta_1, \ldots, \theta_6)$ in Π, U, and Z to emphasize the dependencies on the unknown parameters.) The standard approach is to maximize the "likelihood": choose θ to maximize $\Pi(\tilde{x}^P;\theta)$. Of course, maximizing Π is equivalent to maximizing $\log \Pi$. It is easily demonstrated that the latter is *concave* in θ with gradient

$$\nabla \log \Pi(\tilde{x}^P;\theta) = \left\{ \sum_{\langle s,t\rangle_i} \Phi(\tilde{x}_s^P - \tilde{x}_t^P) - E_\theta \left[\sum_{\langle s,t\rangle_i} \Phi(X_s^P - X_t^P) \right] \right\}_{i=1,\ldots,6} \tag{4.1}$$

where $E_\theta[\cdot]$ is expectation with respect to $\Pi(\cdot;\theta)$. This suggests a gradient ascent procedure, but the expectation $E_\theta[\cdot]$ is intractable, involving summation over the entire range of X^P. In our experiments, we used a 16 grey-level scale for the pixels, and 204 × 204 lattices: the expectation in (4.1) has 16^{204^2} terms. An alternative to brute force evaluation is to use stochastic relaxation (see §2), which produces an (asymptotically) *ergodic* sequence $X^P(1), X^P(2), \ldots$ for any given θ, and from which expectations can be approximated by appropriate time-averages. This, too, is computationally intensive, but feasible. In some settings we have found no alternative, and this Monte Carlo procedure has worked well, albeit slowly (see [**22**]). See also Hinton and Sejnowski [**20**] for a closely related algorithm, used to model learning in a theory of neuron dynamics.

For homogeneous random fields, such as our image models, Besag [**2**, **3**] has proposed an ingenious alternative to maximum likelihood, known as "maximum pseudolikelihood." The pseudolikelihood function is

$$PL(\tilde{x}^P;\theta) \doteq \prod_{s \in S^P \backslash \partial S^P} \Pi(X_s^P = \tilde{x}_s^P | X_r^P = \tilde{x}_r^P, r \neq s; \theta)$$

where ∂S^P is the boundary of S^P under the neighborhood system determined by the energy U, and $S^P \backslash \partial S^P$ is the complement of ∂S^P relative to S^P. The "pseudolikelihood estimator" is the θ that maximizes $PL(\tilde{x}^P;\theta)$. In the next few pages we shall lend some analytic support, by establishing consistency of pseudolikelihood in the "large graph" limit. But first, we emphasize the overwhelming computational advantage. As with the log likelihood function, the log pseudolikelihood function, log $PL(\tilde{x}^P;\theta)$, is concave, but this time the gradient is directly computable:

$$\begin{aligned}\nabla \log PL(\tilde{x}^P;\theta) \\ = \nabla \sum_{s\in S^P\setminus\partial S^P} \left\{ \sum_{i=1}^{6} \theta_i \{\Phi(\tilde{x}^P_s - \tilde{x}^P_{s+\tau_i}) + \Phi(\tilde{x}^P_s - \tilde{x}^P_{s-\tau_i})\} \right. \\ \left. - \log \sum_{\alpha} \exp \left\{ \sum_{i=1}^{6} \theta_i \{\Phi(\alpha - \tilde{x}^P_{s+\tau_i}) + \Phi(\alpha - \tilde{x}^P_{x-\tau_i})\} \right\} \right\}\end{aligned}$$

(where $\sum_\alpha$ is summation over pixel grey levels, zero through fifteen in our experiments)

$$\begin{aligned} = \left\{ \sum_{s\in S^P\setminus\partial S^P} \{\Phi(\tilde{x}^P_s - \tilde{x}^P_{s+\tau_i}) + \Phi(\tilde{x}^P_s - \tilde{x}^P_{s-\tau_i}) \right. \\ \left. - E_\theta[\Phi(X^P_s - X^P_{s+\tau_i}) + \Phi(X^P_s - X^P_{s-\tau_i}) | X^P_r = \tilde{x}^P_r, r \neq s]\} \right\}_{i=1,\dots,6} .\end{aligned}$$

This time, the expectation is tractable. The conditional distribution on X^P_s, given $X^P_r = \tilde{x}^P_r$, $r \neq s$, involves only those variables $\tilde{x}^P_r$ in the neighborhood of s. Furthermore, this time summation is over the range of X^P_s only, which has only sixteen values. In short, the gradient of the log pseudolikelihood is directly computable, and therefore gradient ascent is feasible without resorting to time-consuming Monte Carlo methods. For the experiments discussed in the previous section, the pair-clique parameters were estimated, for each texture type, by gradient ascent of the pseudolikelihood function.

Some modifications of maximum and pseudolikelihood have been recently introduced by Chalmond [**6**]. A third alternative was suggested by Derin and Elliott [**9**, **11**], and has been studied and analyzed extensively by Possolo [**26**]. This involves a regression fit of the log of the local conditional probabilities, and works best when there are a small number of values in the range of the random variables. For example, the method is very effective for Ising-like models.

Consistency of pseudolikelihood. We will study parameter estimation from a single realization of a finite-graph Markov random field. The typical framework for establishing consistency is in the limit as the number of samples increases. But we have in mind estimation from a single sample of the random field, with the *number of sites* large (e.g., 512 × 512). To study estimation in this "large graph" setting, we will imagine a sequence of samples, $X(1), X(2), \dots$, from a sequence of Markov random fields, $\Pi_1, \Pi_2, \dots$, in which the latter are associated with an expanding sequence of regular graphs. We will assume that the sequence of distributions of these random fields has a common unknown parameter vector $\theta_0 \in R^m$. We will define the pseudolikelihood estimate, $\hat{\theta}_n = \hat{\theta}_n(X(n))$, for each sample, $X(n)$, and show that $\hat{\theta}_n \to \theta_0$ with probability one.

The samples $X(1), X(2), \dots$ need not be independent. For example, we may wish to model the observations as subsamples from a single *infinite volume*

Gibbs state. Then, there is one infinite-volume process X, e.g., $X = \{X_s\}_{s\in S}$, $S \doteq \{(i,j)\colon -\infty < i,j < \infty\}$, and the observations are associated with increasing subsets: $X(k) = \{X_s\}_{s\in S_k}$ with, e.g. $S_1 \subseteq S_2 \subseteq \cdots, \bigcup_{k=1}^{\infty} S_k = S$. The sequence of distributions, $\Pi_1, \Pi_2, \ldots$, is the sequence of *conditional distributions*, on $\{X_s\}_{s\in S_k}$, conditioned on $\{X_s\}_{s\in S\setminus S_k}, k = 1,2,\ldots$. Under a suitable "homogeneity" (translation invariance) assumption for the Gibbs potential, the theorem applies, guaranteeing consistency of the pseudolikelihood estimate. This is regardless of *critical phenomena*, or *lack of spatial stationarity*, both of which can occur with infinite volume Gibbs states having translation-invariant potentials [**28**].

Henceforth, we specialize to regular square lattices: S will represent the d-dimensional infinite square lattice. (Generalizations are straightforward.) For each $n, S_n \subset S$ is a d-dimensional cube with sides length n. On S is a translation-invariant neighborhood system $G = \{G_s\}_{s\in S}$ ($s \notin G_s; s \in G_r \Leftrightarrow r \in G_s; s \in G_r \Leftrightarrow s + \tau \in G_{r+\tau}\ \forall s, r, \tau \in S$). We will assume "finite" interactions: $\exists R \ni s \in G_r \Rightarrow |s - r| \leq R$. We will denote the subgraph of (S,G) with sites S_n by (S_n, G). Associated with each n is a Markov random field, Π_n, on (S_n, G). The site variables, $\{X_s\}_{s\in S_n}$, are assumed to have common range Ω, with $|\Omega| < \infty$.

The distributions $\Pi_1, \Pi_2, \ldots$ are related by their dependencies on a common unknown parameter $\theta_0 \in R^m$. Pseudolikelihood exploits the dependencies of local conditional probabilities on this parameter. In particular, fix n and let $x \in \Omega^{S_n}$, the range of the random field with distribution Π_n. For each s, let ${}_s x = \{x_r\colon r \in G_s \cap S_n\}$. Actually, ${}_s x$ will be treated as a vector, in which the components are placed in some arbitrary order. "Local characteristics" of Π_n refers to the conditional probabilities $\Pi_n(X_s = x_s | {}_s X = {}_s x; \theta_0)$ for each $s \in S_n$, $x \in \Omega^{S_n}$. The distributions $\Pi_1, \Pi_2, \ldots$ are tied together by the assumption that these local characteristics, which depend upon θ_0, are independent of s and n, for all s in the interior of S_n. More precisely, letting $S_n^0 = S_n \setminus \partial S_n$ under G, we assume that there exists $\Psi(\cdot) = (\Psi_1(\cdot), \ldots, \Psi_m(\cdot))$ such that

$$\Pi_n(X_s = x_s | {}_s X = {}_s x; \theta_0) = \frac{\exp\{\theta_0 \cdot \Psi(x_s, {}_s x)\}}{\sum_{\alpha\in\Omega} \exp\{\theta_0 \cdot \Psi(\alpha, {}_s x)\}} \tag{4.2}$$

for all n, $s \in S_n^0$, x_s, and ${}_s x$. Any homogeneous field with finite interactions is suitable, regardless of boundary conditions. Examples include the Ising model, and the texture model (for a single, homogeneous texture) developed in §3.

Whenever $s \in S_n^0, \Pi_n(X_s = x_s | {}_s X = {}_s x; \theta)$ does not depend on n. Since we will only be interested in local characteristics at interior sites, we henceforth drop the subscript n when writing conditional probabilities. Given $X = x$, a sample from Π_n, the pseudolikelihood function of $\theta \in R^m$ is

$$\begin{aligned} PL_n(x;\theta) &\doteq \prod_{s\in S_n^0} \Pi(x_s | {}_s x; \theta) \\ &= \prod_{s\in S_n^0} \frac{\exp\{\theta \cdot \Psi(x_s, {}_s x)\}}{\sum_{\alpha\in\Omega} \exp\{\theta \cdot \Psi(\alpha, {}_s x)\}}. \end{aligned}$$

The pseudolikelihood estimate is the *set* $M_n(x)$, of θ that maximize $PL_n(x;\theta)$:

$$M_n(x) = \left\{ \theta \in R^m \colon PL_n(x,\theta) = \sup_{\phi \in R^m} PL_n(x,\phi) \right\}.$$

In establishing consistency for pseudolikelihood estimation we will assume *identifiability*, in the following sense:

DEFINITION. We will say that $\theta_0 \in R^m$, is *identifiable* if $\theta \neq \theta_0 \Rightarrow \exists x_{s,s}x$, such that $\Pi(x_s|_s x;\theta) \neq \Pi(x_s|_s x;\theta_0)$.

THEOREM (CONSISTENCY OF PSEUDOLIKELIHOOD). *For each* $n = 1, 2, \ldots$, *let* $X(n)$ *be a sample from the Markov random field* Π_n, *with local characteristics* (4.2). *If* θ_0 *is identifiable, then*

(a) $P(\log PL_n(X(n);\theta)$ *is strictly concave for all n sufficiently large* $) = 1$;

(b) $P(M_n(X(n))$ *is a singleton for all n sufficiently large* $) = 1$;

(c) $P(\sup_{\theta \in M_n(X(n))} |\theta - \theta_0| \to 0) = 1$.

REMARKS. (1) Extensions to more general graph structures and interaction potentials are possible, and mostly routine.

(2) More relevant to the problem of estimating θ_0 from a sample $X(n)$, with n large, is the following immediate corollary:

$$\lim_{n\to\infty} P\left(\sup_{\theta \in M_n(X(n))} |\theta - \theta_0| > \varepsilon \right) = 0 \quad \forall \varepsilon > 0.$$

PROOF OF THEOREM. Let $N_n = |S_n^0|$,

$$N_n(\beta) = \#\{s \in S_n^0 \colon {}_sX(n) = \beta\},$$

and

$$N_n(\alpha,\beta) = \#\{s \in S_n^0 \colon X_s(n) = \alpha,\ {}_sX(n) = \beta\},$$

using α and β as generic elements of Ω and $\Omega^{|G_s|}$, respectively. The proof can be divided into five steps, which we now state as lemmas.

LEMMA 1. $\liminf_{n\to\infty} N_n(\beta)/N_n > 0$ *a.s.*, $\forall \beta$.

LEMMA 2. $\lim_{n\to\infty} N_n(\alpha,\beta)/N_n(\beta) = \Pi(\alpha|\beta;\theta_0)$ *a.s.*, $\forall \alpha, \beta$.

LEMMA 3. *Let*

$$\begin{aligned} F_n(\theta) &= \frac{1}{N_n}\{\log\ PL_n(X(n);\theta) - \log\ PL_n(X(n);\theta_0)\} \\ &= \sum_\beta \frac{N_n(\beta)}{N_n} \sum_\alpha \frac{N_n(\alpha,\beta)}{N_n(\beta)} \log \frac{\Pi(\alpha|\beta;\theta)}{\Pi(\alpha|\beta;\theta_0)}. \end{aligned}$$

$P(F_n(\cdot)$ *is strictly concave for all n sufficiently large* $) = 1$.

LEMMA 4. *Let*

$$G_n(\theta) = \sum_\beta \frac{N_n(\beta)}{N_n} \sum_\alpha \Pi(\alpha|\beta;\theta_0) \log \frac{\Pi(\alpha|\beta;\theta)}{\Pi(\alpha|\beta;\theta_0)}.$$

(a) *With probability one*, $\forall \varepsilon > 0\ \exists \delta > 0 \ni$

$$\limsup_{n\to\infty} \sup_{|\theta-\theta_0|\le\varepsilon} \sup_{\phi\in R^m,\ |\phi|=1} \phi^t H(G_n(\theta))\phi < -\delta$$

where $H(G_n(\theta))$ *is the matrix of second derivatives* (*Hessian*) *of* $G_n(\theta)$ *with respect to* θ.

(b) $G_n(\theta) \le 0\ \forall\theta, n$.

(c) $G_n(\theta_0) = 0\ \forall n$.

LEMMA 5. $\forall\varepsilon > 0$,

$$\lim_{n\to\infty} \sup_{|\theta-\theta_0|\le\varepsilon} |F_n(\theta) - G_n(\theta)| = 0 \quad a.s.$$

With these pieces in place, we complete the proof as follows.

Fix $\varepsilon > 0$. From Lemma 4, conclude that

$$\liminf_{n\to\infty} \inf_{|\theta-\theta_0|=\varepsilon} (G_n(\theta_0) - G_n(\theta)) > 0 \quad \text{a.s.} \tag{4.3}$$

Since F_n is uniformly approximated by G_n (in the sense of Lemma 5), (4.3) also holds for F_n:

$$\liminf_{n\to\infty} \inf_{|\theta-\theta_0|=\varepsilon} (F_n(\theta_0) - F_n(\theta)) > 0 \quad \text{a.s.}$$

Since F_n is eventually strictly concave (Lemma 3), it eventually achieves its maximum, uniquely, in $\{\theta\colon |\theta - \theta_0| < \varepsilon\}$. Finally, since $\log PL_n(X(n);\theta) = N_nF_n(\theta) + \log PL_n(X(n);\theta_0)$, these same statements apply to $\log PL_n(X(n);\theta)$.

We now proceed to prove Lemmas 1–5.

PROOF OF LEMMA 1. The first two lemmas are based on the following version of the "strong law of large numbers":

PROPOSITION. *For each* $n = 1, 2, \ldots$, *let* $Z_1(n), Z_2(n), \ldots, Z_{m_n}(n)$ *be random variables and* $Y(n)$ *be a random vector. Assume*

(1) $\liminf_{n\to\infty} m_n/n > 0$.

(2) $Z_1(n), \ldots, Z_{m_n}(n)$ *are conditionally independent, given* $Y(n)$.

(3) $|Z_i(n)| \le B < \infty\ \forall i, n$.

Then

$$\left|\frac{1}{m_n}\sum_{i=1}^{m_n}(Z_i(n) - E[Z_i(n)|Y(n)])\right| \to 0 \quad a.s.$$

PROOF. The methods here are standard. We will provide an outline only. Fix $\varepsilon > 0$ and let A_n be the event

$$A_n = \left\{\left|\frac{1}{m_n}\sum_{i=1}^{m_n}(Z_i(n) - E[Z_i(n)|Y(n)])\right| > \varepsilon\right\}.$$

Then the usual exponential bounds (but derived by first conditioning on $Y(n)$) give $P(A_n) = o(1/C^{m_n})$ for some $C > 1$. The rest follows from the Borel-Cantelli lemma: $P(A_n$ infinitely often $) = 0$.

Now back to the proof of Lemma 1: For any $s \in S$, let

$$B_s = \partial\{(s \cup G_s)^c\} = \{r\colon \exists t \in (s \cup G_s) r \in G_t,\ r \notin (s \cup G_s)\},$$

i.e., the neighborhood of $s \cup G_s$. For each n, choose $s_1, s_2, \ldots, s_{m_n} \in S_n$ such that

(1) $\liminf_{n\to\infty} m_n/N_n > 0$,
(2) $B_{s_i} \subseteq S_n$, $i = 1, \ldots, m_n$,
(3) $i \neq j \rightarrow (s_i \cup G_{s_i}) \cap B_{s_j} = \varnothing$,

(e.g., regularly partition S_n into large cubes, with sizes independent of n, and big enough to accommodate $s \cup G_s \cup B_s$, for some s).

Fix β and let $Y(n) = \{X_s(n)\colon s \in \bigcup_{i=1}^{m_n} B_{s_i}\}$, and $Z_i(n) = 1_{{}_{s_i}X(n)=\beta}$. By the Markov property, $Z_1(n), \ldots, Z_{m_n}(n)$ are conditionally independent, given $Y(n)$. Hence, by the proposition,

$$\left| \frac{1}{m_n} \sum_{i=1}^{m_n} (Z_i(n) - E[Z_i(n)|Y(n)]) \right| \rightarrow 0 \quad \text{a.s.}$$

Using again the Markov property,

$$E[Z_i(n)|Y(n)] = \Pi({}_{s_i}X(n) = \beta | X_s(n), s \in B_{s_i}; \theta_0) \text{ *}$$

which can have only a finite number of possible values (corresponding to the $|\Omega|^{|B_{s_i}|}$ configurations of $\{X_s(n)\}_{s \in B_{s_i}}$), all of which are positive. Hence, for some $\varepsilon > 0$,

$$\frac{1}{m_n} \sum_{i=1}^{m_n} E[Z_i(n)|Y(n)] > \varepsilon, \quad \forall n,$$

and

$$\liminf \frac{1}{m_n} \sum_{1}^{m_n} Z_i(n) \geq \varepsilon \quad \text{a.s.}$$

Since $N_n(\beta) \geq \sum_1^{m_n} Z_i(n)$, it also follows that $\liminf N_n(\beta)/m_n \geq \varepsilon$ a.s. Finally, since $\liminf m_n/N_n > 0$, $\liminf N_n(\beta)/N_n \geq \liminf N_n(\beta)/m_n \cdot \liminf m_n/N_n > 0$.

PROOF OF LEMMA 2. Let $C = \{c_i\colon i = 1, \ldots, n_c\}$ be a coloring of (S, G). In other words, $c_1, c_2, \ldots, c_{n_c}$ partition S, and $r, s \in c_i \rightarrow r \notin G_s$. Because (S, G) is regular, we can assume that C is chosen so that $\liminf |S_n^0 \cap c_i|/N_n > 0$, $i = 1, \ldots, n_c$.

For each $i \in \{1, \ldots, n_c\}$ define

$$N_n(\beta; c_i) = \#\{s \in S_n^0 \cap c_i\colon {}_sX(n) = \beta\},$$
$$N_n(\alpha, \beta; c_i) = \#\{s \in S_n^0 \cap c_i\colon X_s(n) = \alpha, {}_sX(n) = \beta\}.$$

Fix $i \in \{1, \ldots, n_c\}$, α and β, and let

$$Z_s(n) = 1_{X_s(n)=\alpha; {}_sX(n)=\beta} \quad \text{for each } s \in S_n^0 \cap c_i.$$

Let $B_n = \partial\{(S_n^0 \cap c_i)^c\}$ (the neighborhood of $S_n^0 \cap c_i$) and let $Y(n) = \{X_s(n)\colon s \in B_n\}$. Given $Y(n)$, the random variables $Z_s(n), s \in S_n^0 \cap c_i$, are independent

*It is well known that the local characteristics (4.2) determine these conditional probabilities as well. Hence, this conditional distribution is independent of n.

(Markov property). By the proposition

$$\left| \frac{1}{|S_n^0 \cap c_i|} \sum_{s \in S_n^0 \cap c_i} (Z_s(n) - E[Z_s(n)|Y(n)]) \right| \to 0 \quad \text{a.s.}$$

Using again the Markov property: $E[Z_s(n)|Y(n)] = \Pi(\alpha|\beta;\theta_0)1_{{}_sX(n)=\beta}$. Since

$$\sum_{s \in S_n^0 \cap c_i} Z_s(n) = N_n(\alpha,\beta;c_i) \quad \text{and} \quad \sum_{s \in S_n^0 \cap c_i} 1_{{}_sX(n)=\beta} = N_n(\beta;c_i),$$

$$\frac{1}{|S_n^0 \cap c_i|} |N_n(\alpha,\beta;c_i) - \Pi(\alpha|\beta;\theta_0) \cdot N_n(\beta;c_i)| \to 0 \quad \text{a.s.}$$

Finally, recalling that $\liminf N_n(\beta)/N_n > 0$, a.s.:

$$\begin{aligned} &\left| \frac{N_n(\alpha,\beta)}{N_n(\beta)} - \Pi(\alpha|\beta;\theta_0) \right| \\ &\quad = \left| \sum_{i=1}^{n_c} \frac{N_n(\alpha,\beta;c_i)}{N_n(\beta)} - \Pi(\alpha|\beta;\theta_0) \sum_{i=1}^{n_c} \frac{N_n(\beta;c_i)}{N_n(\beta)} \right| \\ &\quad \le \sum_{i=1}^{n_c} \frac{1}{N_n(\beta)} |N_n(\alpha,\beta;c_i) - \Pi(\alpha|\beta;\theta_0) N_n(\beta;c_i)| \\ &\quad = \sum_{i=1}^{n_c} \frac{N_n}{N_n(\beta)} \frac{|S_n^0 \cap c_i|}{N_n} \frac{1}{|S_n^0 \cap c_i|} |N_n(\alpha,\beta;c_i) - \Pi(\alpha|\beta;\theta_0) N_n(\beta;c_i)| \\ &\quad \to 0 \quad \text{a.s.} \end{aligned}$$

PROOF OF LEMMA 3. Let $H(F_n(\theta))$ be the Hessian (matrix) of $F_n(\theta)$, and let $\phi \in R^m$. By routine calculation, we derive

$$\begin{aligned} &\phi^t H(F_n(\theta))\phi \\ &\quad = -\sum_\beta \frac{N_n(\beta)}{N_n} \frac{\sum_{\tilde\alpha \in \Omega} (\phi \cdot (\psi(\tilde\alpha,\beta) - E_\theta[\psi(\alpha,\beta)|\beta]))^2 \exp\{\theta \cdot \psi(\tilde\alpha,\beta)\}}{\sum_{\tilde\alpha \in \Omega} \exp\{\theta \cdot \psi(\tilde\alpha,\beta)\}} \end{aligned}$$

where $E_\theta[\cdot|\beta]$ is expectation on Ω with respect to $\Pi(\cdot|\beta;\theta)$. Obviously,

$$\phi^t H(F_n(\theta))\phi \le 0, \quad \forall \phi,$$

and hence $F_n(\theta)$ is concave. By Lemma 1, with probability one, $\inf_\beta N_n(\beta)/N_n > 0$ for all n sufficiently large. Suppose $\inf_\beta N_n(\beta)/N_n > 0$ and $\phi^t H(F_n(\theta))\phi = 0$ for some θ and $\phi \ne 0$. Then, for all $\tilde\alpha$ and $\beta, \phi \cdot \psi(\tilde\alpha,\beta) = E_\theta[\psi(\alpha,\beta)|\beta]$. In particular, for every β, $\phi \cdot \psi(\alpha,\beta)$ is independent of α. This implies that $\Pi(\alpha|\beta;\theta+\phi) = \Pi(\alpha|\beta;\theta_0)$ for all α and β, which contradicts the identifiability assumption. Hence $F_n(\theta)$ is strictly concave whenever $\inf_\beta N_n(\beta)/N_n > 0$.

PROOF OF LEMMA 4. By the same argument used for Lemma 3, $G_n(\theta)$ is strictly concave, whenever $\inf_\beta N_n(\beta)/N_n > 0$. By Lemma 1, with probability one, there is a $\varsigma > 0$ such that $\inf_\beta N_n(\beta)/N_n \ge \varsigma$ for all n sufficiently large. Since $\phi^t H(G_n(\theta))\phi$ is jointly continuous in ϕ, θ, and the finite collection of variables $N_n(\beta)/N_n$, it must achieve its maximum on the compact set $|\phi| = 1$, $|\theta - \theta_0| \le \varepsilon$, and $N_n(\beta)/N_n \in [\varsigma, 1]$ for all β. Part (a) of Lemma 4 now follows from the strict concavity of $G_n(\theta)$.

For part (b), apply Jensen's inequality:

$$\begin{aligned}\sum_\alpha \Pi(\alpha|\beta;\theta_0)\log\frac{\Pi(\alpha|\beta;\theta)}{\Pi(\alpha|\beta;\theta_0)} &\leq \log\sum_\alpha \Pi(\alpha|\beta;\theta_0)\frac{\Pi(\alpha|\beta;\theta)}{\Pi(\alpha|\beta;\theta_0)}\\ &= \log\sum_\alpha \Pi(\alpha|\beta;\theta) = \log 1 = 0.\end{aligned}$$

Part (c) follows immediately from the expression for $G_n(\theta)$.

PROOF OF LEMMA 5.

$$\begin{aligned}&\limsup_{n\to\infty}\sup_{|\theta-\theta_0|\leq\varepsilon}|F_n(\theta)-G_n(\theta)|\\ &\quad=\limsup_{n\to\infty}\sup_{|\theta-\theta_0|\leq\varepsilon}\left|\sum_\beta\frac{N_n(\beta)}{N_n}\sum_\alpha\left(\frac{N_n(\alpha,\beta)}{N_n(\beta)}-\Pi(\alpha|\beta;\theta_0)\right)\log\frac{\Pi(\alpha|\beta;\theta)}{\Pi(\alpha|\beta;\theta_0)}\right|\\ &\quad\leq|\Omega|\sup_{\alpha,\beta,|\theta-\theta_0|<\varepsilon}\left|\log\frac{\Pi(\alpha|\beta;\theta)}{\Pi(\alpha|\beta;\theta_0)}\right|\limsup_{n\to\infty}\sup_{\alpha,\beta}\left|\frac{N_n(\alpha,\beta)}{N_n(\beta)}-\Pi(\alpha|\beta;\theta_0)\right|.\end{aligned}$$

By Lemma 2,

$$\limsup_{n\to\infty}\sup_{\alpha,\beta}\left|\frac{N_n(\alpha,\beta)}{N_n(\beta)}-\Pi(\alpha|\beta;\theta_0)\right| = 0 \quad \text{a.s.}$$

Since $\Pi(\alpha|\beta;\theta)\neq 0$ for any $\alpha,\beta,\theta\in R^m$, and is continuous in θ for each of the finite numbers of $\alpha\in\Omega$, $\beta\in\Omega^{|G_s|}$,

$$\sup_{\alpha,\beta,|\theta-\theta_0|<\varepsilon}\left|\log\frac{\Pi(\alpha|\beta;\theta)}{\Pi(\alpha|\beta;\theta_0)}\right|$$

is finite.

ACKNOWLEDGEMENTS. This work has benefited from a host of suggestions by D. Geman and B. Gidas. D. Geman has experimented extensively with MRF texture models, and with pseudolikelihood parameter estimation, and has shared his conclusions. B. Gidas has shared his insight and knowledge of Markov random fields, especially concerning the vagaries of infinite volume Gibbs states.

REFERENCES

1. E. Aarts and P. van Laarhoven, *Simulated annealing: a pedestrian review of the theory and some applications*, NATO Advanced Study Institute on Pattern Recognition: Theory and Applications, Spa, Belgium, June 1986.

2. J. Besag, *Spatial interaction and the statistical analysis of lattice systems (with discussion)*, J. Roy. Statist. Soc. Ser. B **36** (1974), 192–236.

3. ——, *Statistical analysis of non-lattice data*, The Statistician **24** (1975), 179–195.

4. ——, *On the statistical analysis of dirty pictures (with discussion)*, J. Roy. Statist. Soc. Ser. B **48** (1986).

5. V. Cerný, *A thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm*, Preprint, Inst. Phys. and Biophys., Comenius Univ., Bratislava, 1982.

6. B. Chalmond, *Image restoration using an estimated Markov model*, Preprint, Mathematics Dept., University of Paris-Sud, Orsay, 1986.

7. F. S. Cohen and D. B. Cooper, *Simple parallel hierarchical and relaxation algorithms for segmenting noncausal Markovian random fields*, IEEE Trans. Pattern Anal. Machine Intell., PAMI-9 (1987), pp. 195–219.

8. G. R. Cross and A. K. Jain, *Markov random field texture models*, IEEE Trans. Pattern Anal. Machine Intell., PAMI-5 (1983), pp. 25–40.

9. H. Derin and H. Elliott, *Modelling and segmentation of noisy and textured images using Gibbs random fields*, IEEE Trans. Pattern Anal. Machine Intell., PAMI-9 (1987), pp. 39–55.

10. P. A. Devijver, *Hidden Markov models for speech and images*, Nato Advanced Study Institute on Pattern Recognition: Theory and Applications, Spa, Belgium, June 1986.

11. H. Elliott and H. Derin, *Modeling and segmentation of noisy and textured images using Gibbs random fields*, Tech. Report ECE-UMASS-SE84-15, Dept. Elec. Comput. Eng., Univ. of Massachusetts, Amherst, Mass.

12. D. Geman, S. Geman, and C. Graffigne, *Locating texture and object boundaries*, Pattern Recognition Theory and Application (P. Devijver, ed.), NATA ASI Series, Springer-Verlag, Heidelberg, 1986.

13. S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Trans. Pattern Anal. Machine Intell. **6** (1984), 721–741.

14. S. Geman and D. E. McClure, *Bayesian image analysis: an application to single photon emission tomography*, 1985, Statistical Computing Section, Proceedings of the American Statistical Association, 1985, pp. 12–18.

15. B. Gidas, *Non-stationary Markov chains and convergence of the annealing algorithm*, J. Statist. Phys. **39** (1985), 73–131.

16. ——, *A renormalization group approach to image processing problems*, Preprint, Division of Applied Mathematics, Brown University, 1986.

17. U. Grenander, *Lectures in pattern theory*, vols. I,II,III, Springer-Verlag, New York, 1976.

18. ——, *Tutorial in pattern theory*, Division of Applied Mathematics, Brown University, 1983.

19. B. Hajek, *Cooling schedules for optimal annealing*, Math. Oper. Res. (to appear).

20. G. E. Hinton and T. J. Sejnowski, *Optimal perceptual inference*, Proc. IEEE Conf. Comput. Vision Pattern Recognition, 1983.

21. S. Kirkpatrick, C. D. Gellatt, and M. P. Vecchi, *Optimization by simulated annealing*, Science **220** (1983), 671–680.

22. A. Lippman, *A maximum entropy method for expert system construction*, Ph.D. Thesis, Division of Applied Mathematics, Brown University, 1986.

23. J. Marroquin, S. Mitter, and T. Poggio, *Probabilistic solution of ill-posed problems in computational vision*, Artif. Intell. Lab. Tech. Report, M.I.T., 1985.

24. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *Equations of state calculations by fast computing machines*, J. Chem. Phys. **21** (1953), 1087–1091.

25. D. W. Murray, A. Kashko, and H. Buxton, *A parallel approach to the picture restoration algorithm of Geman and Geman on an SIMD machine*, Preprint, 1986.

26. A. Possolo, *Estimation of binary Markov random fields*, Preprint, Department of Statistics, Univ. of Washington, Seattle, 1986.

27. B. D. Ripley, *Statistics, images, and pattern recognition*, Canad. J. Statist. **14** (1986), 83–111.

28. D. Ruelle, *Thermodynamic formalism*, Addison-Wesley, Reading, Mass., 1978.

Brown University, Providence, Rhode Island 02912, USA

Brown University, Providence, Rhode Island 02912, USA