International Journal of Computer Vision, 6:3, 227-243 (1991)

A Common Framework for Image Segmentation

DAVI GEIGER
Siemens Corporate Research, Inc., 755 College East Road, Princeton, NJ 08540

ALAN YUILLE
Division of Applied Sciences, Harvard University, Cambridge, MA 02138

Received November 2, 1989. Revised November 8, 1990 and April 1, 1991.

Abstract
We attempt to unify several approaches to image segmentation in early vision under a common framework. The Bayesian approach is very attractive since: (i) it enables the assumptions used to be explicitly stated in the probability distributions, and (ii) it can be extended to deal with most other problems in early vision. Here, we consider the Markov random field formalism, a special case of the Bayesian approach, in which the probability distributions are specified by an energy function.

We show that: (i) our discrete formulations for the energy function is closely related to the continuous formulation; (ii) by using the mean field (MF) theory approach, introduced by Geiger and Girosi [1991], several previous attempts to solve these energy functions are effectively equivalent; (iii) by varying the parameters of the energy functions we can obtain connections to nonlinear diffusion and minimal description length approaches to image segmentation; and (iv) simple modifications to the energy can give a direct relation to robust statistics or can encourage hysteresis and nonmaximum suppression.

1 Introduction

This article extends the work of Geiger and Girosi [1991] in an attempt to unify many methods of image segmentation under a common framework. First we study the weak membrane model that smooths the intensity field except at the discontinuities. It has been studied by many researchers including Blake and Zisserman [1987], Chou and Brown [1988], Gamble and Poggio [1987], Geman and Geman [1984], Koch et al. [1985], and Marroquin [1987].

We also address the problem of relating continuous space models [Mumford and Shah 1985] and models defined on lattices. This is done by finding a discrete formulation based on the continuous model of Ambrosio [1988] in which the boundaries are represented by a single line process. This contrasts with the usual lattice formulation [Geman & Geman 1984] which, by using both horizontal and vertical line processes, tends to bias lines toward these two directions. We show that using a single line process avoids some of these biases (see figure 1). In general it is preferable to formulate the energy function on the continuous space since it does not depend on the particular tesselation of the space; however, the lattice formulation proves to be more efficient for the computational point of view.

We study these energy function models in terms of statistical field theory. In the past years many researchers have investigated the use of statistical field theory, in particular Markov random fields, for early vision [Geman & Geman 1984; Marroquin 1987; Gamble & Poggio 1987].

Given an energy function model one can define a corresponding statistical model. If the energy $E(f, l)$ depends on two fields, f (the smoothed image) and l (the discontinuities), then (using the Gibbs distribution) the probability of a particular state of the system is defined by

$$P(f, l) = \frac{e^{-\beta E(f,l)}}{Z} \tag{1}$$

where β is the inverse of the temperature parameter and Z is the partition function (a normalization constant).

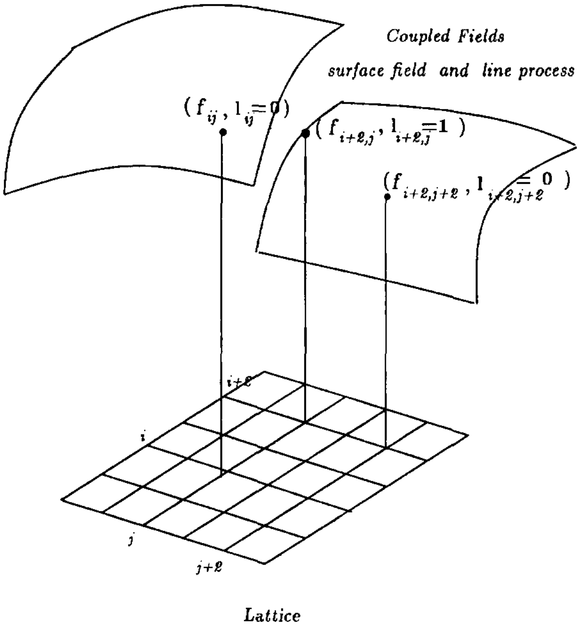
Coupled Fields
surface field and line process
$(f_{ij}, l_{ij}=0)$
$(f_{i+2,j}, l_{i+2,j}=1)$
$(f_{i+2,j+2}, l_{i+2,j+2}=0)$
i+2
i
j
j+2
Lattice

Fig. 1. The surface field f, and the line process l are represented at the sites, (i, j), $(i + 2, j)$ and $(i - 2, j + 2)$, of the lattice. Both are single quantities. Notice, however, that the line process is actually defined in a lattice displaced, from the image lattice, half pixel in each direction.

We can interpret the results in terms of Bayes' formula

$$P(f, l \mid g) = \frac{P(g \mid f, l)P(f, l)}{P(g)}$$

where $P(g \mid f, l)$ is the probability of the data g given a segmented and smoothed image f, l, $P(f, l)$ is the a priori probability of the segmented smoothed image and $P(g)$ is the a priori probability of the data.

Given these statistical models there are many quantities one would like to calculate, for example the maximum a posteriori estimates of the fields given the data. There are several different approaches, for example simulated annealing [Kirkpatrick et al. 1983; Geman & Geman 1984], for calculating these quantities.

This article concentrates on the mean [Geiger & Girosi 1991] quantities of the field (in the zero temperature limit, these are the quantities that minimize the energy function). A justification to use the mean field (MF) as a measure of the field f resides in the fact that it represents the minimum-variance Bayes estimator. More precisely, the variance of the field f is given by

$$\mathrm{var}_{\bar{f}} = \sum_{f,l} (f - \bar{f})^2 P(f, l)$$

where $\bar{f}$ is the center of the variance and the $\Sigma_{f,l}$ represents the sum over all the possible configurations of f and l. Minimizing $\mathrm{var}_{\bar{f}}$ with respect to all possible values of $\bar{f}$, we obtain

$$\frac{\partial}{\partial \bar{f}} \mathrm{var}_{\bar{f}} = 0 \rightarrow \bar{f} = \sum_{f,l} f P(f, l)$$

This implies that the minimum variance estimator is given by the MF value.

Recently Geiger and Girosi [1991] have shown that by applying mean field (MF) theory, as used in statistical physics, to coupled Markov random fields (MRFs), a set of deterministic equations are obtained, whose solutions correspond to the mean fields. In particular, while the values of the discontinuity field (the line process) can be statistically 0 or 1, the MF solution for the discontinuity field ranges from 0 to 1 throughout the continuous values. They also introduced the concept of an effective potential for the smoothed image field with the line process eliminated. The work proposed a link between the statistical algorithms [Geman & Geman 1984; Marroquin 1987] and the alternative deterministic graduated nonconvexity algorithm [Blake and Zisserman 1987].

The parameter β is external to the energy function and can be used to lower the temperature to zero (by increasing β to infinity) to find the zero temperature solution. This method can be thought of as a deterministic form of simulated annealing [Kirkpatrick et al. 1983] and has been used by many algorithms, for example [Hopfield & Tank 1985; Durbin & Willshaw 1987; Geiger & Girosi 1991]. It is also related to continuation methods [Wasserstrom 1973]. We find that several deterministic methods for image segmentation are special cases of these equations, namely (i) the graduated nonconvexity algorithm [Blake & Zisserman 1987] (previously shown in [Geiger & Girosi 1991]) (ii) the Hopfield network approach [Hopfield and Tank 1985; Koch et al. 1985; Yuille 1987], and (iii) the minimal length encoding [Leclerc 1988]. In addition, (iv) there are connections to smoothness with adaptive weights [Grimson & Pavlidis 1987; Terzopoulos 1986] and (v) we can show that some of the nonlinear diffusion approaches to image segmentation can be related to these equations.

The energy function contains several parameters specifying the relative importance of different terms. By varying these parameters we can extend the scale-space description of Witkin [1983]. This suggests a different strategy for obtaining the MF solution. Instead of calculating the solution by steepest descent (or some other method) for a specific set of the parameter values one can start with a known solution for a fixed set of parameter values and track the solution as these values change. This idea enables us to relate the energy function and statistical approach to alternative methods of image segmentation by nonlinear diffusion equations. In particular we show that the network suggested by Perona and Malik [1987] to implement an anisotropic diffusion equation is an approximation to this approach.

We then show how the cost function can be adapted to deal with salt-and-pepper noise where it is desirable to throw away outlying data. This gives a direct link to robust statistics [Huber 1981] and therefore provides a robust method for image segmentation.

Finally it is suggested that the interactions described in the energy function formalism can incorporate many of the basic features used in algorithms for image segmentation [Canny 1986]. These basic properties are (i) smoothing the field that represents the input, except at the discontinuities, (ii) encouraging the creation of lines in the direction perpendicular to the gradient of the field, this is an excitatory mechanism known as hysteresis [Canny 1986] for edge detection, and (iii) inhibiting the formation of multiple responses to edges. This corresponds to nonmaximum suppression in the direction defined by the gradient of the image.

This article is organized as follows: section 2 discusses the continuous and discrete formalization of energy function. In section 3 we introduce statistical mechanics, considering the winner-take-all problem, and use MF theory to obtain equations for the mean field solution. Section 4 applies MF theory to image segmentation. In section 5 we find dynamic equations to solve MF equations and show that this incorporates previous deterministic methods. Section 6 introduces a parameter space where each element is an MF solution for a different set of the parameters. This framework enables us to relate the nonlinear diffusion methods of smoothing to the minimization of an energy function. Section 7 shows that cost function can be easily extended to introduce robust statistics and/or incorporate nonmaximum suppression and hysteresis. Section 8 presents preliminary implementation results.

2 Continuous and Discrete Cost Functions

Consider the problem of smoothing the field and at the same time detecting discontinuities. This problem can be formulated in terms of a simple energy function that has the appropriate interaction between the data field and the line process (discontinuity field). A simplified version of the model is given by the weak membrane energy that has been proposed for the continuous case by Mumford and Shah [1985], and on the discrete lattice by [Geman & Geman 1984; Marroquin 1985; Blake & Zisserman 1987]. The continuous formulation of Mumford and Shah [1985] is

$$E = \iint_D (f(x) - g(x))^2 \, dx + \alpha \iint_{D-C} \nabla f(x) \cdot \nabla f(x) \, dx + \gamma \int_C ds$$

Here $f(x)$ is the smoothed data field, $g(x)$ is the data, D is the domain (a subset of R^2), and C is curves that divide the domain up into separate regions. The energy is to be minimized over all possible fields $f(x)$ and all possible edges C. The first term, called the *data term*, accounts for the error between the data and the field. The second term, called the *smoothness term*, imposes smoothness except across the edges C. The last term, called the *cost term*, establishes the price to establish the edges C.

Ambrosio [1988] proposes a continuous formulation of the problem which represents the edges using a line process $l(x)$.

$$E_\epsilon = \int_D \Big\{ (f(x) - g(x))^2 + \alpha\{\nabla f(x) \cdot \nabla f(x) + \nabla l(x) \cdot \nabla l(x)\} \times (1 - l(x))^{2/\epsilon} + \frac{\gamma l(x)^2}{4\epsilon^2} \Big\} \, dx \tag{2}$$

Here $f(x)$, $g(x)$, and D are as before. Ambrosio proves that this is equivalent to Mumford and Shah's formulation as $\epsilon \mapsto 0$, the lines where $l(x) \neq 0$ correspond to the curves C.

However, the lattice based approaches represent the edges by horizontal and vertical line processes. These occur on two lattices interposed with the lattice representing the data field. The concept of a single line process offers a way of going between the continuous and the discrete formulations. It suggests using a single line process on the lattice, instead of a process using a pair of horizontal and vertical lines. We argue that this helps reduce the anisotropy and the bias toward preferring horizontal and vertical edges.

2.1 The Discontinuity as a Single Field

Continuous formulations are desirable for mathematical reasons [Mumford & Shah 1985], including a lack of bias toward vertical and horizontal edges, but lattice formulations are more practical. In computer vision the image is either captured by a camera or synthetically produced. In both cases the image will be stored in the computer as an array and therefore the space will be discrete. However, different tesselations of the space may be used and they may have different properties. For instance, a hexagonal shaped lattice has different symmetry properties than a square lattice.

It would be desirable to obtain the lattice formulation by directly discretizing equation (2). For example, by setting $\epsilon = 2L$, where L is the lattice spacing. However, as Richardson [1990] points out, it is unclear if the limit of the resulting lattice formulation exists as $L \to 0$ ($\epsilon \to 0$). Our discrete formulation is inspired by Ambrosio but not directly equivalent.

We argue that a single field formulation couples better with any given tesselation of the space (not just the square lattice). Although it might seem to be natural to define the discontinuity field as a single field on the lattice this has not been previously used. Many authors, for example [Geman & Geman 1984; Koch et al. 1985; Marroquin et al. 1985; Geiger & Girosi 1991], have considered a horizontal and vertical line process which corresponds to have the line process as a 2-dimensional vector field. In a single-field formulation the direction of the contour curve is obtained by the derivatives (gradients) which have different shapes on different tesselations of the space.

We propose the discrete model to be

$$E = \Bigg[\sum_{i,j} (f_{i,j} - g_{i,j})^2 + \alpha(\Delta_j^2 f + \Delta_i^2 f)(1 - l_{i,j}) + \gamma l_{i,j} \Bigg] \tag{3}$$

where Δ_{if} is a discrete derivative of f in the x direction and $\Delta_i^2 f$ is short for $(\Delta_i f)^2$.

The choice of lattice approximation to the derivative is very important, since the localization of $\Delta_i f$ has to be the same as $\Delta_j f$ and consequently the same as l_{ij}. The "natural" choice, $\Delta_i f = f_{i,j} - f_{i-1,j}$, does not have this propriety. We define the derivative of f in the x-direction as $\Delta_i f = 0.5(f_{i,j} - f_{i-1,j} + f_{i,j-1} - f_{i-1,j-1})$ (see figure 2). Another way to write this is $\Delta_i f = K_{ij} + M_{ij}$ and $\Delta_j f = K_{ij} - M_{ij}$, where $M_{ij} = 0.5(f_{ij-1} - f_{i-1j})$, $K_{ij} = 0.5(f_{ij} - f_{i-1\,j-1})$. Notice however that this definition of discrete derivative decouples the lattice into two lattices (like black and white squares of a chessboard) for places where $l_{ij} = 0$. We point out that the field f does not coincide with the single-line process.

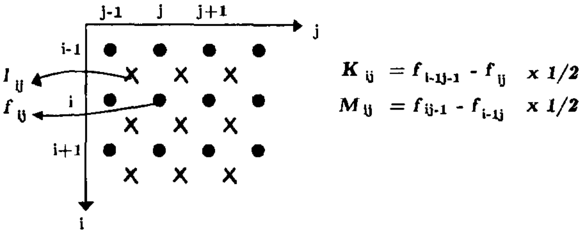
j-1
j
j+1
j
i-1
l_{ij}
f_{ij}
i
i+1
i
$K_{ij} = f_{i-1j-1} - f_{ij} \;\; x\ 1/2$
$M_{ij} = f_{ij-1} - f_{i-1j} \;\; x\ 1/2$

Fig. 2. This definition of derivative does not break the diagonal interactions.

3 Statistical Mechanics and MF Theory

The standard Markov random field approach to segmentation defines an energy function $E(f, l)$ of the line process and the field to be smoothed. The ensemble of this system is given by all possible states of the system. These states correspond to fields f and l defined everywhere on the grid. The energy function $E(f, l)$ defines the probability of a give solution for the fields by the Gibbs distribution [Parisi 1988].

$$P_E = \frac{1}{Z} e^{-\beta E(f,l)}$$

This implies that every state of the system has a finite probability of occurring. The more likely ones are those with low energy.

The parameter a, which can be interpreted as the inverse of the temperature $T = 1/\beta$, controls the sharpness of the distribution. It follows directly from the Gibbs distribution that the *ordering* of the relative likelihood of states depends only on their energies and is independent of β. The magnitude of the relative likelihood, however, varies with β. As $\beta \to 0$ all states become equally likely. Conversely, it can be shown as $\beta \to \infty$ that only the state, or states, with lowest energy has a nonzero probability of occurring. Thus $1/\beta$ can be thought of intuitively as a measure of the uncertainty of the model.

This intuition can be made more precise by interpreting the probability in terms of Bayes' theorem, see section 1. This theorem expresses the probability P_E in terms of the probability $P(d \mid f, l)$ of the data given fields (f, l) times the prior probability $P(f, l)$ of the fields (f, l). For the image segmentation energy function, equation (2), the $(f - g)^2$ term ensures that $P(d \mid f, l)$ is a Gaussian distribution with variance $1/2\beta$. Thus maximizing P_E with respect to the fields (f, l) corresponds to maximum a posteriori estimation of (f, l) with a model that assumes Gaussian noise in the data with variance $1/2\beta$ [Geman & Geman 1984]. Thus $1/\beta$ is a measure of the noise assumed to be in the data, and hence is a measure of the uncertainty of the model.

Minimizing the energy function will correspond to finding the most probable state, independent of the value of β. The MF solution is more general and reduces to the most probable solution as $T \rightarrow 0$. It corresponds to defining the solution to be the mean fields, that is, the averages of the f and l fields over the probability distribution. This enables us to obtain different solutions depending on the uncertainty. Statisticians refer to it as the minimal variance estimator.

One can now estimate the most probable states, or the mean states, of this probability distribution by, for example, using Monte Carlo techniques [Metropolis et al. 1953]. The drawback of these methods is the amount of computer time needed for the implementation.

The β parameter may now be used to fulfill a different purpose. It is often easier to compute the solutions at higher temperatures (low β) and slowly reduce the temperature at the same time as updating the solution. This is the essence of the extremely useful simulated annealing [Kirkpatrick et al. 1983] approach and the deterministic annealing approach that we describe next.

It should be emphasized that the a parameter has two independent purposes. Firstly, it is used as a parameter to specify the probability distributions assumed by the model. Secondly, it can be used as a continuation parameter, as in simulated annealing, to help compute the most probable states or the mean field values.

3.1 MF Theory and the Winner Take All (W.T.A)

MF theory gives methods for obtaining fast deterministic algorithms that make use of the richness of the statistical formulation to find the MF solution. The MF solution corresponds to the average solution of the problem and is obtained by averaging all possible states according to their probability. So the MF solution will not necessarily correspond to the solution that minimizes the energy, especially if that state is very isolated from the others. However, if there is uncertainty in the model, then the MF solution may be more reliable [Berger 1985] than the minimum energy solution. Moreover for the limit of $\beta \rightarrow \infty$ (no uncertainty) the MF solution becomes the minimum of the energy.

We introduce MF theory by using it to solve the problem of W.T.A. The W.T.A. problem can be posed as: given a set $\{T_i\}$ of N inputs to a system how does one choose the maximum input and suppress the others. For simplicity, we assume all of the T_i's to be positive. We introduce the set of binary variables $\{V_i\}$, such that $V_w = 1$ selects the winner and $V_i = 0$ for $i \neq w$.

We will calculate the partition function in two separate ways for comparison. The first method uses the mean field approximation and gives an approximate answer. The second method is exact and is best for this application, but cannot be used for all problems. It involves calculating the partition function for a subset of the possible V_i's, a subset chosen to ensure that only one V_i is nonzero.

3.1.1 W.T.A. with the Mean-Field Approximation.

Define the energy function

$$E_1^{WtA}[\vec{V}] = \sum_{i=1}^{N} \left(\sum_{j \neq i, =1}^{N} V_i V_j \right) - \lambda \sum_{i=1}^{N} T_i V_i \qquad (4)$$

where λ is a parameter to be specified and $\vec{V} = (V_1, \ldots, V_N)$ denotes the state of the system. The solution of the W.T.A. will have all the V_i to be zero except for the one corresponding to the maximum T_i. This constraint is imposed implicitly by the first term on the right-hand side of (4) (note that the constraint is encouraged rather than explicitly enforced).

Now we formulate the problem statistically. The energy function above defines the probability of a state $\vec{V}$ to be $P[\vec{V}] = 1/Z\, e^{-\beta E_1^{WtA}[\vec{V}]}$, where β is the inverse of the temperature parameter and and Z is the normalization constant, called the *partition function*, given by

$$Z = \sum_{\vec{V}} e^{-\beta E_1^{WtA}[\vec{V}]}$$

where the sum is taken over all admissible configurations $\vec{V}$.

The mean values, $\{\bar{V}_i\}$, of the fields can be computed directly from the partition function,

$$\bar{V}_k = \sum_{\vec{V}} V_k \frac{e^{-\beta E_1^{WtA}[\vec{V}]}}{Z}$$

$$= \frac{1}{Z\beta\lambda} \frac{\partial}{\partial T_k}$$

$$\sum_{\vec{V}} \exp\left(-\beta \sum_{i=1}^{N} \left[\left(\sum_{j \neq i,=1}^{N} V_i V_j\right) - \lambda T_i V_i\right]\right)$$

$$= \frac{1}{\beta\lambda} \frac{1}{Z} \frac{\partial Z}{\partial T_k}$$

$$= \frac{1}{\beta\lambda} \frac{\partial \ln Z}{\partial T_k}$$

Using the mean field approximation [Geiger & Girosi 1991] we can now approximate the partition function Z. We can write Z as

$$Z = \sum_{\vec{V}} e^{-\beta E_1^{WtA}[\vec{V}]}$$

$$= \sum_{\vec{V}} \prod_i \exp\left[-\beta V_i \left(\sum_{j \neq i} V_j - \lambda T_i\right)\right]$$

When calculating the contribution of Z from a specific element V_i, the mean field approximation replaces the values of the other elements V_j by their mean values $\bar{V}_j$. This assumes that only low-order correlations between elements are important [Parisi 1988]. This yields an approximate partition function Z_{approx}

$$Z_{\text{approx}} = \sum_{\vec{V}} \prod_i \exp\left[-\beta V_i \left(\sum_{j \neq i} \bar{V}_j - \lambda T_i\right)\right]$$

The expression for Z_{approx} is now a product of independent terms, which can be summed over independently.

$$Z_{\text{approx}} =$$

$$\prod_i \left\{ \sum_{V_i=\{0,1\}} \exp\left[-\beta V_i \left(\sum_{j \neq i} \bar{V}_j - T_i\right)\right]\right\}$$

$$= \prod_i \left\{1 + \exp\left[-\beta\left(\sum_{j \neq i} \bar{V}_j - \lambda T_i\right)\right]\right\}$$

We can now differentiate Z_{approx} with respect to the $\{T_i\}$ to obtain consistency conditions for the $\{\bar{V}_i\}$, the mean field equations

$$\bar{V}_i = \frac{1}{1 + \exp\left[\beta\left(\sum_{j \neq i,=1}^{N} \bar{V}_j - \lambda T_i\right)\right]} \quad (5)$$

To solve (5) we can arrange for it to be the fixed point of a dynamical equation, for example

$$\frac{d\bar{V}_i}{dt} = -\bar{V}_i + \frac{1}{1 + \exp\left[\beta\left(\sum_{j \neq i,=1}^{N} \bar{V}_j - \lambda T_i\right)\right]}$$

The problem with this method is that we cannot guarantee that it will converge to the correct solution (there may be several solutions to (5)). Though, provided that $\lambda < 1/T_i^{\max}$ (so $\lambda T_i < 1$ for all i), then as $\beta \rightarrow \infty$ the correct solution will satisfy (5), that is, $\bar{V}_i = 1$ for the maximum T_i and $\bar{V}_i = 0$ otherwise. For the finite values of a the solution is more general and $\bar{\ }_i$ assigns a weight for each value of T_i that can be used to enhance the signal.

Because this method is not guaranteed to give the correct solution we now discuss a more efficient way of dealing with the winner-take-all problem.

3.1.2 W.T.A Without the Mean Field Approximation. We now impose the constraint that, for each admissible state, the V_i sum to 1 explicitly during the computation of the partition function. The first term on the right-hand side of (4) is now unnecessary and we use an energy function

$$E_2^{WtA}[\vec{V}] = -\sum_{i=1}^{N} T_i V_i$$

We compute Z by summing over all possible $\vec{V}$ under the constraint that the components sum to one (i.e., we sum over the states $\vec{V} = (1, 0, 0, \ldots, 0), (0, 1, 0, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)$. This gives

$$Z = \sum_{\{V_i=0,1\}}^{\Sigma_k V_k = 1} e^{-\beta E_2^{WtA}[V_i]} = \sum_i e^{\beta T_i}$$

In this case no approximation is needed and we obtain

$$\bar{V}_j = \frac{1}{\beta}\frac{\partial \ln Z}{\partial T_j} = \frac{e^{\beta T_j}}{\Sigma_i\, e^{\beta T_i}} \tag{6}$$

Thus as $\beta \to \infty$ the V_i corresponding to the largest T_i will be switched on and the other V_j will be off. This method is guaranteed to converge to the correct solution.

4 MF Theory and the Effective Energy

The tools of statistical mechanics enable us to calculate the MF solution directly, provided the partition function is known. Unfortunately, however, it is often impossible to directly compute the partition function and saddle-point approximations must be used [Parisi 1988]. This gives rise to a set of consistency equations for an approximation to the mean fields. To obtain the correct solution to these equations, a heuristic continuation method known as deterministic annealing is used, which involves the equations at high temperature and tracking the solution to lower temperatures.

We now compute the partition function associated to the energy function by summing the probability over all possible states of the fields. Following [Geiger and Girosi 1991], we first sum over all possible l_{ij}, and an effective energy is obtained. The analysis of this effective energy allows us to better understand the system. Moreover by using a saddle point approximation we can find the mean value of f_{ij} by minimizing the effective energy with respect to this field.

4.1 MF and the Effective Potential

For the weak membrane model (given by (3)) the partition function is given by

$$Z = \sum_{\{f\}} \sum_{\{l=0,1\}} \exp\left\{-\beta\left[\sum_{i,j}(f_{i,j} - g_{i,j})^2 + \gamma + (\alpha_{L^2}(\Delta_j^2 f + \Delta_i^2 f) - \gamma)(1 - l_{i,j})\right]\right\}$$

where the $\Sigma_{\{f\}}\,\Sigma_{\{l=0,1\}}$ represents the sum over all the possible configurations of the field f and l. Computing the sum over all the possible states l we obtain

$$Z = \sum_{\{f\}} \exp\left\{-\beta \sum_{i,j}\left[(f_{i,j} - g_{i,j})^2 + \gamma - \frac{1}{\beta}\ln(1 + e^{\beta H_{ij}})\right]\right\} \tag{7}$$

where

$$H_{ij} = \gamma - \alpha_{L^2}(\Delta_i^2 f + \Delta_j^2 f)$$

The interaction of the field f with itself has changed after the line process has been averaged in the partition function. From (7) we notice that the partition function can be rewritten as

$$Z = \sum_{\{f\}} e^{-\beta E_{\text{eff}}(f)}$$

where

$$E_{\text{eff}}(f) = \sum_{i,j}(f_{ij} - g_{ij})^2 + \gamma - \frac{1}{\beta}\ln\left[(1 + e^{\beta \bar{H}_{i,j}})\right] \tag{8}$$

We then plot the effective potential (without the data term) as a function of the gradient of f (see figure 3). This is equivalent to the result obtained by Geiger and Girosi, but for the scalar line process. As they point

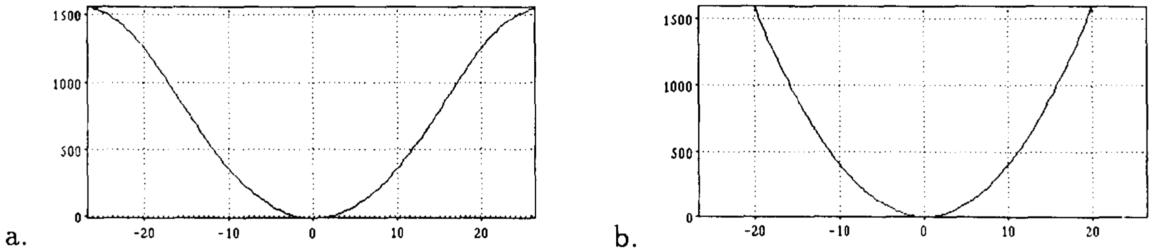

Fig. 3. The effective potential is shown as a function of $\Delta_i^2 f$. a. For $\beta = 0.002$. b. Zero temperature limit ($\beta \to \infty$). Taken from Geiger and Girosi [1991].

out, an exact computation of Z by means of the transfer matrix method shows that this potential is highly nonlocal and it becomes hard to solve the problem. It is possible however, by using the saddle-point approximation, to substitute the sum over all possible states of the field f by its maximum value. This approximation holds if the fluctuations of the values of f are small and we make this assumption. The partition function then becomes (up to a scaling constant)

$$Z_A = \exp\{-\beta \min_f [E_{\text{eff}}(f)]\} \tag{9}$$

Alternatively, we can choose our estimate of f to be the maximum a posteriori estimate of (1) after summing out the line process fields. This corresponds to minimizing $E_{\text{eff}}(f)$ with respect to f, and gives identical results without need of the saddle-point approximation. The MF equation for the line process field is given by

$$\bar{l}_{i,j} = -\frac{1}{\beta}\frac{\partial \ln Z_A}{\partial H_{ij}} = \frac{1}{1 + e^{\beta \bar{H}_{i,j}}} \tag{10}$$

where $\bar{H}_{i,j}$ is computed with the mean value of f and $H_{ij} = [\gamma - \alpha_{L^2}[(\Delta_i \bar{f})^2 + (\Delta_j \bar{f})^2)]$. It is interesting to notice that the mean value of the line process can vary continuously from 0 to 1. For the zero temperature limit ($\beta \rightarrow \infty$) equation (10) becomes $\bar{l}_{i,j} = \theta(-\bar{H}_{i,j})$ where $\theta(x)$ is the step function.

5 Deterministic Solutions for f

The aim of this section is to show that several existing algorithms using line processes are essentially the same.

There are several dynamic methods to compute the MF solution for f. The first one, discussed in this section, defines a dynamic equation with respect to time such that the fixed point of the equation (as time → ∞) is the MF solution. The temperature of the system can be gradually reduced to provide deterministic annealing. It was previously shown [Geiger & Girosi 1991] that the graduated non convexity algorithm [Blake & Zisserman 1987] is an approximation to this. We now show that another deterministic approach based on Hopfield networks [Koch et al. 1985] is also equivalent. Moreover several approaches based on adaptive smoothness are also related [Grimson & Pavlidis 1985; Lee & Pavlidis 1987; Terzopoulos 1986].

The second method, described in the next section, considers variations of the MF solution over the parameter space. We derive the mean field equations with respect to a particular parameter and provide suitable initial conditions.

5.1 Dynamic Equations

As we discussed before, the MF solution minimizes $E_{\text{eff}}(f)$. A possible way to introduce dynamics is to make the field f time dependent and perform steepest descent on the energy function.

$$\frac{df_{ij}(t)}{dt} = -\frac{\partial E_{\text{eff}}(f)}{\partial f_{ij}(t)} \tag{11}$$

In this case the fixed point, that is, f_{fix} such that $(df/dt)(f_{\text{fix}}) = 0$, is the solution of the static MF equation. For discrete time, (11) becomes the gradient descent algorithm

$$\frac{\partial E_{\text{eff}}(f)}{\partial f_{ij}}(f_{ij}^n) = -\frac{1}{\omega}(f_{ij}^{n+1} - f_{ij}^n)$$

where ω is the time step and n counts the number of steps. For the effective energy equation (8), the dynamic equation becomes

$$\begin{aligned} f_{ij}^{n+1} &= f_{ij}^n - 2\omega\{\lambda_{ij}(f_{ij}^n - g_{ij}) \\ &\quad + \alpha_{L^2}[K_{ij}(1 - \bar{l}_{ij}) - K_{i+1\,j+1}(1 - \bar{l}_{i+1\,j+1}) \\ &\quad - M_{i+1j}(1 - \bar{l}_{i+1j}) + M_{ij+1}(1 - \bar{l}_{ij+1})]\} \end{aligned} \tag{12}$$

where the index n indicates the step of the evaluation procedure and $M_{ij} = 0.5(f_{ij-1} - f_{i-1j})$, $K_{ij} = 0.5\,(f_{ij} - f_{i-1\,j-1})$.

We can then find the fixed point by updating f recursively according to (12), coupled with the updating rule for the line process given by (10). For dense data the fixed point $\bar{f}$ will satisfy

$$\begin{aligned} \bar{f}_{ij} = g_{ij} &- \alpha_{L^2}[K_{ij}(1 - \bar{l}_{ij}) \\ &- K_{i+1\,j+1}(1 - \bar{l}_{i+1\,j+1}) \\ &- M_{i+1j}(1 - \bar{l}_{i+1j}) \\ &+ M_{ij+1}(1 - \bar{l}_{ij+1})] \end{aligned} \tag{13}$$

From now on, for simplicity, we drop the index L^2 of α.

5.2 The Hopfield Network Approach

An alternative approach to minimizing energy functions was developed by Koch et al. [1985] and Yuille [1987] by adapting a technique used by Hopfield [1984]. This led to a method that was similar to graduated nonconvexity [Blake & Zisserman 1987] and had analogies to the MF theory. Indeed it was shown by Yuille [1987] that this method generated solutions to the MF theory equations. We now show the correspondence to MF theory.

In Koch et al. [1985] and Yuille [1987], the binary line process fields l_{ij} become continuous variables in the range [0, 1]. They are related to variables m_{ij} by $l_{ij} = g(m_{ij}, \lambda)$ where

$$g(m_{ij}, \lambda) = \frac{1}{1 + e^{-\lambda m_{ij}}}$$

is a sigmoid function. The energy is modified by the addition of a term (this is equivalent to the inverse gain function term used by Hopfield [1984] which can be thought of as entropy

$$E_G = \frac{c}{2\lambda} \sum_{i,j} \{l_{ij} \log (l_{ij}) + (1 - l_{ij}) \log (1 - l_{ij})\}$$

The dynamics of the system are given by

$$\frac{df_{ij}}{dt} = - \frac{\partial}{\partial f_{ij}} \{E + E_G\}$$

$$\frac{dm_{ij}}{dt} = - \frac{\partial}{\partial l_{ij}} \{E + E_G\} \qquad (14)$$

where E is given by equation (3). It is straightforward to show [Koch et al. 1985], following Hopfield [1984], that the system converges to a minimum of $E + E_G$. As $\lambda \mapsto 0$ the energy $E + E_G$ becomes convex and there is a unique minima. As $\lambda \mapsto \infty$ the additional energy $E_G \mapsto 0$ and we recover the original energy function. This suggests minimizing the energy for small λ and tracking the solution as λ increases.

Equations (14) converge to solutions of

$$\frac{\partial}{\partial f_{ij}} \{E + E_G\} = 0$$

$$\frac{\partial}{\partial l_{ij}} \{E + E_G\} = 0 \qquad (15)$$

Equation (15) gives for the weak membrane model

$$-\alpha(\Delta_i^2 f + \Delta_j^2 f) + \gamma + \frac{1}{\lambda} (\log l_{ij} - \log (1 - l_{ij}))$$

$$= 0 \rightarrow l_{i,j} = \frac{1}{1 + e^{\lambda(\gamma - \alpha(\Delta_i^2 f + \Delta_j^2 f))}} \qquad (16)$$

This is identical to the MF solution in (10) provided we set $\lambda = \beta$. By substituting l_{ij}, given by (10), into the effective energy we obtain the same energy $E + E_G$ as above. We conclude that the methods using the Hopfield technique [Koch et al. 1985; Yuille 1987] are equivalent to those using mean field theory [Geiger & Girosi 1991]. MF theory applied to the fields f (input) and l (output) naturally accounts for the continuous values of l and the gain term introduced by Hopfield.

5.3 Smoothing with Adaptive Weights

By adapting recent work from Durbin [private communication] on the elastic-net algorithm for the traveling salesman problem [Durbin & Willshaw 1987], we can propose an alternative minimization algorithm. This algorithm minimizes $E[f, l]$ with respect to the f variables with l fixed, then calculates the most probable estimate of the l's analytically using (10), and repeats the minimization-estimation process. It is closely related to the EM algorithms [Dempster et al. 1977]. An advantage of this algorithm is an increased speed-up in time since, when the l's are fixed, the energy function is quadratic in the f's and quicker algorithms than steepest descent can be used for minimizing it. Since the l's are computed analytically, both stages of the EM strategy are very fast.

This new algorithm is also related to smoothing algorithms using variable weights. These typically do smoothing with a quadratic energy function, readjust the weights of the smoothing terms based on some fitness criterion and smooth again. Terzopoulos [1986] proposes smoothing over discontinuities and then breaking the surface at places where the tension is too high. This breaking can be achieved by adjusting the weights of the smoothness terms and would correspond to setting the l's to be 1 at such places. Grimson and Pavlidis [1985] and Lee and Pavlidis [1987] discuss ways to readjust these weights iteratively on the basis of the residual difference between the data and the interpolated surface.

6 Parameter Space: Generalized Scale Space

The energy function is specified by a set of parameters. As we vary these parameters we obtain different MF solutions. This generates a parameter space of solutions where each point in the space is the MF solution for the corresponding set of parameters. This can be thought of as a generalization of scale space [Witkin 1983; Yuille & Poggio 1983]. Several authors have suggested varying the values of the parameters of the energy function either to obtain a scale space description [Blake & Zisserman 1987] or to relate the discontinuities of this model to zero crossings of certain linear operators [Mumford & Shah 1985].

The techniques described above will obtain the MF solution. Altering the parameter values will change the probability distributions and lead to different solutions. In some situations the optimal values of these parameters can be estimated [Geiger & Poggio 1987], but in general they are unknown.

The β parameter is used to take into account the uncertainty of the model. Changing β will not alter the *ordering* of the probabilities of states, since β multiplies all the terms of the energy function, but it will affect the *relative* probabilities of states. The more confident we are the bigger we make β. In the limit as $\beta \mapsto \infty$ the probability distribution is infinitely strongly peaked about the state with least energy. In this case the mean solutions of the field will correspond to values of the field in the least-energy configuration. Increasing β during the computation gives rise to annealing.

The parameters α and γ are more closely related to the scale of the structure in the image that we wish to extract. Increasing α (with the other parameters fixed) will correspond to increasing the smoothness. Increasing γ (with the other parameters fixed) reduces the number of lines.

We now investigate what happens as we vary the parameters (see figure 4). For any specific values of the parameters we can use the steepest descent techniques to obtain solutions of the mean field theory. We can now track a path through parameter space by seeing how the solutions vary as the parameters change. This could be done by two methods: (i) we use the solution for one set of parameter values as initial condition for steepest descent for neighboring values, (ii) we differentiate the MF equations with respect to the parameters to obtain first-order equations for how the solutions change as the parameters vary.

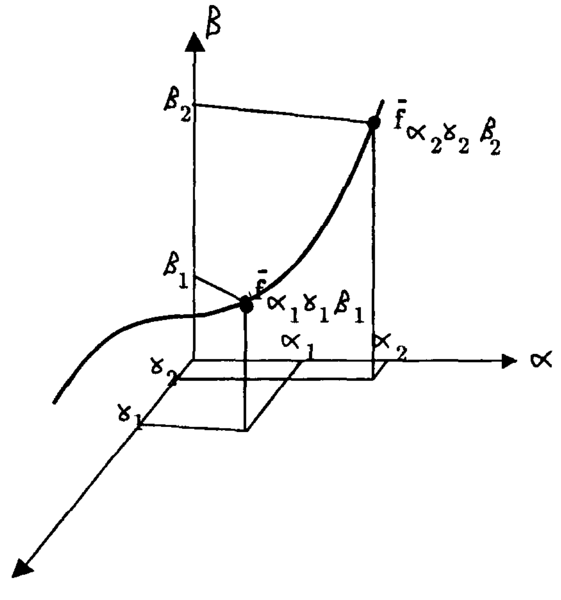

Fig. 4. The parameter space for the parameters α, β, and γ. Each point in space corresponds to an MF solution for that set of parameters.

The second approach by-passes the need to perform steepest descent at all *provided* we have an initial solution to start from. One method of obtaining such a solution is to observe that for $\alpha = 0$ the global minimum of the energy function (and the MF solutions) correspond to $f = g$, so the solution is known for this case. Starting with a solution $f = g$ for $\alpha = 0$ and then increasing α leads to a method very similar to the nonlinear diffusion methods for edge detection. In particular we show in the next section that the method proposed by Perona and Malik [1987] is an approximation to these techniques.

6.1 Nonlinear Diffusion: Varying α, β, and γ

We now consider finding a solution of the system and tracking it as the amount of smoothness increases. This gives a direct connection between methods of image segmentation using Markov random field techniques and approaches using nonlinear diffusion. We study properties of the MF solution when the scale parameter varies.

For $\alpha = 0$ (no smoothness imposed) the MF solution is $f = g$ since the line-process fields are decoupled from the data fields, and for quadratic fields the mean solution corresponds to the minimum of the energy function. We can now track the solution as the amount of smoothness increases. Intuitively, as we increase the amount of smoothness we will blur the image and obtain a scale space description. By using the equations for the MF solution we can obtain an equation for how the solution varies with α.

The continuous version of the discrete MF equations given by (13) and (10) are

$$\bar f(x) = g(x) + \alpha\nabla\cdot(\nabla\bar f(x)(1 - \bar l(x)))$$

$$\bar l(x) = \frac{1}{1 + \exp\{\beta[\gamma - \alpha\nabla\bar f(x)\cdot\nabla\bar f(x))]\}} \qquad (17)$$

We can take the derivative of $\bar f(x)$ with respect to α to see how the solution varies in the parameter space.

$$\frac{\partial \bar{f}(x)}{\partial \alpha} = \nabla \cdot [\nabla \bar{f}(x)(1 - \bar{l}(x))] + \alpha \frac{\partial \nabla \cdot [\nabla \bar{f}(x)(1 - \bar{l}(x))]}{\partial \alpha} \qquad (18)$$

If we use (10) to substitute for $\bar{l}(x)$ we obtain an initial value equation for $\bar{f}$ in α with initial conditions $\bar{f} = g$ (for $\alpha = 0$). For small values of the parameter α ($\alpha \mapsto 0$), equation (18) becomes

$$\frac{\partial \bar{f}(x)}{\partial \alpha} = \nabla \cdot \left(\frac{\nabla \bar{f}(x)}{1 + \exp\{-\beta[\gamma - \alpha \nabla f(x) \cdot \nabla f(x)]\}} \right) \qquad (19)$$

This is a diffusion equation where the line-process field $l(x, \alpha)$ controls the anisotropy of the diffusion process. It is equivalent to a method defined by Perona and Malik [1987] that produced good experimental results for image segmentation. An advantage of (18) is that, since it gives a solution of the MF equations, it is easier to interpret. We can also vary the dependence of γ on α to obtain different edge thresholds during the evolution of the process. The Markov field formalism has the additional advantage that it can be used as a general formulation of early vision with an elegant probabilistic interpretation.

It should be emphasized that tracking a solution in parameter space by repeatedly minimizing the energy function for different parameter values may be more reliable than using the explicit equation for $\partial f / \partial \alpha$.

6.1.1 An Alternative Analogy to the Diffusion Equation. There is an alternative analogy to the diffusion equation (also noticed by Nordstrom [1990] and a related method has been found by Lumesdaine—personal communication). The dynamic equation (12) in the continuous case gives

$$\frac{\partial f(x, t)}{\partial t} = f(x, t) - g(x) + \nabla \cdot \{\alpha \nabla [f(x, t)(1 - l(x, t))]\} \qquad (20)$$

where

$$l(x, t) = \frac{1}{1 + \exp\{\beta[\gamma - \alpha \nabla^2 f(x, t)]\}}$$

The first term on the r.h.s. of (20) is given by $f(x, t) - g(x)$. If we start with initial conditions $f(x, 0) = g(x)$, then this term is initially zero. In a number of situations, for example when the noise is small for small t, this term can be neglected. In this case, (20) becomes

$$\frac{\partial f(x, t)}{\partial t} = \alpha \nabla \cdot \{\nabla [f(x, t)(1 - l(x, t))]\} \qquad (21)$$

This is an anisotropic nonlinear diffusion equation and is basically the same (after substituting for $l(x, t)$ as (19). However, as we discussed before the MF solution is only obtained from (21) as $t \mapsto \infty$. Therefore at any step the value of f can not be interpreted as the mean field value. Although, formally (21) is like (19) their difference resides on the different roles of the parameters α and the time t. More precisely, (21) is a gradient descent method and the only meaningful solution is obtained at the minimum, while (19) produces mean field values throughout the whole path.

6.2 Minimal Length Encoding

We can also relate the line process method to an alternative approach to image segmentation suggested by Leclerc [1988] based on Rissanen's work on minimal length encoding [Rissanen 1978]. In its simplest version it corresponds to fitting the image to a set of piecewise constant regions while minimizing the length of the boundaries between regions. The intuitive idea of the approach is that for a given problem a solution for it can be seen as the one that requires minimal encoding length. There is a direct correspondence between minimal length encoding and Bayesian probability. Given a cost function for encoding we can define the corresponding probability theory using the Gibbs distribution. Then the maximum a posteriori (MAP) estimate corresponds to the minimal length encoding.

Leclerc starts with the minimal length idea and arrives at a functional of the following form (in one dimension)

$$\sum_i \left\{ \frac{(f_i - g_i)^2}{\sigma} + b[1 - \delta(f_i - f_{i-1})] \right\} \qquad (22)$$

and in order to solve it he suggests a continuation method starting from the functions.

$$\sum_i \left\{ \frac{(f_i - g_i)^2}{\sigma^2} + b\left[1 - \exp\left(-\frac{f_i - f_{i-1}}{s}\right)^2\right] \right\} \qquad (23)$$

where as the parameter s decreases to zero we obtain the original energy function. He noticed that as $s \rightarrow \infty$ the solution is $f_i = g_i$. He uses this as initial conditions and minimizes (23) with respect to f_i while decreasing s.

By comparing (23) with (8) we can see that both potentials have essentially the same form, shown in figure 3. The parameter $1/\sigma^2$ is equivalent to our β (it is easier to see the whole comparison by multiplying (8) with β). In this way Leclerc works with a finite temperature σ^2. The parameter b relates directly to γ since in the limit of large gradients of f, both potentials become "flat" with constants b and γ respectively. The parameter $1/s$ is equivalent to α. We can see that a similar result will be achieved by our model as we increase α, since the membrane term will now be strictly enforced (leading to approximately piecewise constant regions). We can find a more precise relation between s and α by finding the inflection point of both effective potentials (without the data term) with respect to the gradient, that is, the zero of the second derivative of the potential (without the data term) with respect to the gradient, $(f_i - f_{i-1})$. For (23) the inflection point is at $(f_i - f_{i-1})^2 = s/2$ and for (8) it occurs at

$$\frac{2\alpha}{1 + \exp(-\beta \bar{H}_i)} \left[1 - \frac{2\alpha\beta(f_i - f_{i-1})^2}{1 + \exp(-\beta \bar{H}_{ik})} \right] = 0$$

Where $H_i = [\gamma - \alpha(f_i - f_{i-1})^2]$ is the one dimensional version of H_{ij}. For the special case of zero temperature (corresponding to $\sigma = 0$) the inflection occurs at $(f_i - f_{i-1})^2 = \gamma/\alpha$ and the inverse relation between s and α becomes transparent, for a constant value of γ. In the general case, for finite β, we can see that as $\alpha \mapsto \infty$ the inflection point moves to $(f_i - f_{i-1})^2 = 0$, which corresponds to $s = 0$. In this case the solution is piecewise constant.

Therefore Leclerc's algorithm can be interpreted within our framework as a continuation method, corresponding to increasing α in a parameter space, for the mean-field effective energy and finite $\beta = 1/\sigma^2$. This also suggests that a mean length encoding may be more appropriate than a minimal length encoding.

Effectively this approach will lead to results similar to the nonlinear diffusion one, discussed above, but it does not fully exploit the temperature parameter (as, for example, the annealing methods do). More sophisticated versions of the minimal length encoding approach using polynomial patches will similarly correspond to higher-order smoothness terms.

Leclerc's penalty for the creation of lines is proportional to the total length of the line, hence it is similar to our model. More sophisticated methods of minimal length encoding [Keeler 1990] put penalties on lines that are proportional to the amount of information needed to describe the line. For example a straight edge requires four numbers to specify it, the coordinates of the initial and final points, and hence its cost is independent of its length.

7 Results

We have implemented a combination of the gradient-descent method with the continuation method on the smoothing parameter (α) and the temperature (β). We used equations (13) and (10) starting with α, $\beta = 0$ and varying these parameters as the gradient descent was applied. This initial value of the parameters gives the exact initial condition that $f_{ij} = g_{ij}$. The final value of β is ∞ (in our case, $\infty = 2$, which is a high enough value that guarantees $l_{ij} = 1, 0$). We used a linear schedule for updating β and α and we demonstrate in two image examples to a final $\alpha = 1$ and 4 (see figures 5, 6, and 7).

8 Extensions

This section describes two simple modifications that can be done to the energy function.

8.1 Robust Statistics

The energy function model we have been considering is inappropriate for a certain class of images. In particular, for images corrupted with salt and pepper noise it seems inferior to classical techniques such as median filtering. We briefly show how the energy function can be easily modified to deal with this situation. The method presented here is obtained from the work of Geiger and Pereira [1990] for the problem of minimal visual encoding and from Woodward Yang (private communication), however an alternative derivation is presented by Girosi, Poggio, and Caprile [1989]. By use of mean field techniques we can show [Yuille et al. 1990] that this reformulation is closely related to robust statistics [Huber 1981] and can be thought of as

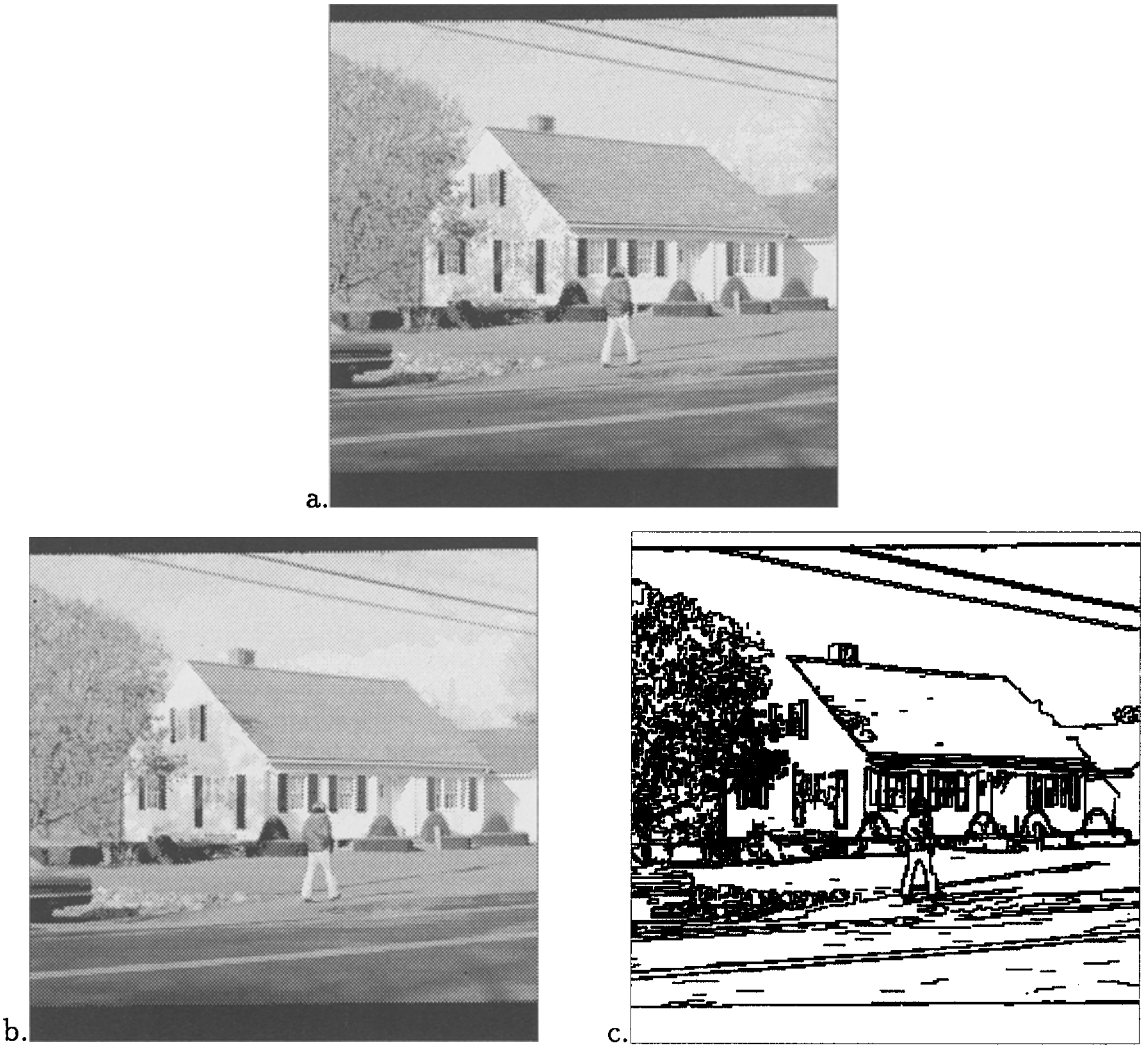

Fig. 5. a. An 8-bit image of 256 × 256 pixels. b. The smoothed image for final $\mu = 0.7$, $\gamma_{ij} = 60$, and 500 iterations. c. The corresponding line process.

robust visual reconstruction. Robust techniques are designed to be reliable in the presence of noise and to small errors in the assumptions of the model [Huber 1981].

The basic idea is to allow the algorithm to ignore certain data points by paying a penalty. We introduce, for the 1-D case, a binary valued field V_i such that $V_i = 0$ if the ith data point is included in the surface reconstruction and $V_i = 1$ if it is discarded. Discarded points must pay a penalty ν. This modifies the [Geman & Geman 1984] cost function to give

$$E[f, l, V] = \sum_i (1 - V_i)(f_i - g_i)^2$$

$$+ \lambda \sum_i (f_{i+1} - f_i)^2(1 - l_i)$$

$$+ \mu \sum_i l_i + \nu \sum_i V_i$$

We can apply the same mean field theory techniques as above to average out the V field as well as the l field. This gives

$$E_{\text{eff}}[f] = \frac{-1}{\beta} \sum_i \log \{e^{-\beta(f_i - d_i)^2} + e^{-\beta\nu}\}$$

$$\frac{-1}{\beta} \sum_i \log \{e^{-\beta(f_{i+1} - f_i)^2} + e^{-\beta\mu}\}$$

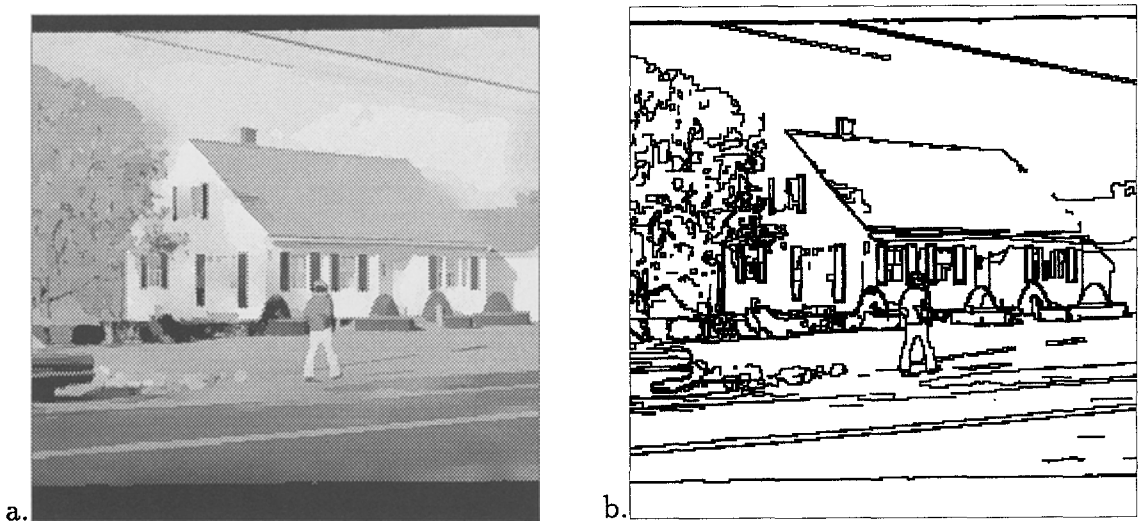

Fig. 6. a. For figure 5a. the smoothed image for final $\mu = 16$, $\gamma_{ij} = 256$, and 500 iterations. b. The corresponding line process.

The first part of the energy function has the same functional form shown in figure 3, though the argument of the function is now $(f_i - d_i)^2$. This shows that the interaction between f_i and d_i is quadratic for small $(f_i - d_i)^2$ but becomes constant as $(f_i - d_i)^2$ becomes large. Hence data points will have little influence if their values are too far from the surface and so salt-and-pepper noise will be ignored.

The property of interactions being locally quadratic and becoming constant for large distances is characteristic of *redescending M-estimators* [Huber 1981] used in robust statistics. Huber describes three estimators of this type derived using different criteria. The forms of these estimators are very close to figure 3. Huber usually graphs the derivatives of these functions (which fall to zero for large distances, hence the name *redescending*).

Techniques from robust statistics have been applied to different aspects of vision (e.g., [Pavlidis 1986; McKendall & Mintz 1989; Hallinan & Mumford 1990]) and are probably desirable for visual reconstruction. The connection described here between robust statistics and Markov models is extended by Yuille et al. [1990] where we also show a relation to the *α-trimmed mean estimator* [Huber 1981] and a method used by Girosi et al. [1989] for a neural network learning algorithm.

8.2 Hysteresis and Nonmaximum Suppression

There are two important directions at a boundary, parallel to the boundary and perpendicular to it. We want to encourage edges in one direction, *hysteresis*, and suppress them in the other, *nonmaximum suppression*. Hysteresis and nonmaximum suppression [Canny 1986] are two important heuristics used for segmenting and detecting discontinuities in an image.

Hysteresis acts by extending strong edges to places where the intensity discontinuities are weak. This is a desirable property for an edge detector and one can enforce it by an additional term E^{hys} in the energy function, with

$$E^{\text{hys}} = \xi \int (\nabla^{\perp} f \cdot \nabla l)^2 \, dx \tag{24}$$

where f is the intensity field, l is the discontinuity field, and $\nabla^{\perp} f$ is the vector perpendicular to ∇f and with the same norm. This energy, when minimized, will enforce the discontinuity field to be smooth (∇l small) along the line perpendicular to ∇f (hence along the edge) and therefore it will be associated to the hysteresis term for edge detection. Since the fields f and l are coupled, this additional term will reduce the smoothing of the field f across that line.

Nonmaximimum suppression thins edges selectively on the basis of their strengths. To account for it we propose the energy term

$$E^{\text{nms}} = -\xi \int (\nabla f \cdot \nabla l)^2 \, dx \tag{25}$$

where f is the intensity field, l is the discontinuity field, and ∇f is the gradient of the field f. This energy, when minimized, will enforce the discontinuity field to be discontinuous (∇l) along the gradient of f (∇f). This

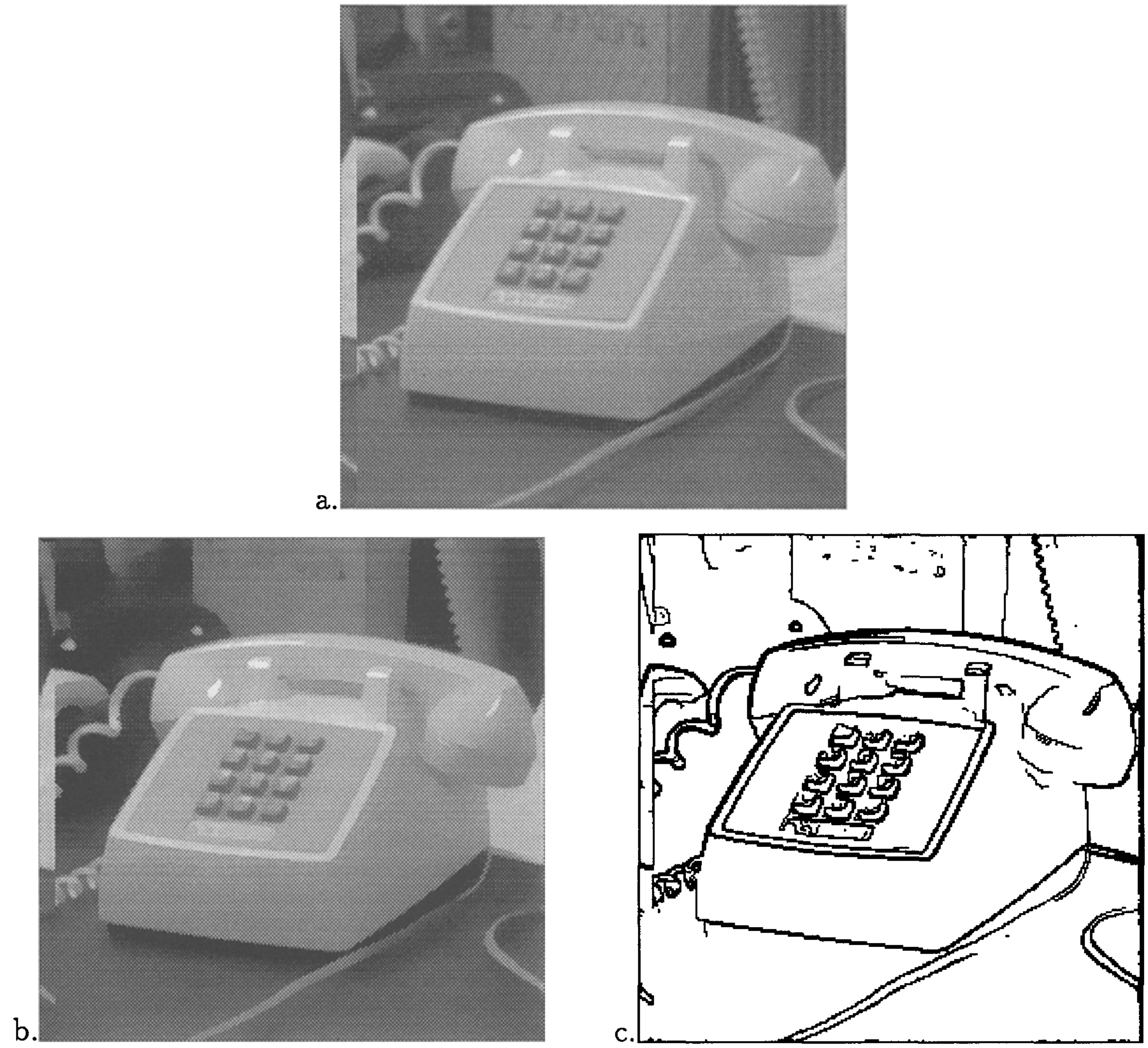

Fig. 7. a. An 8-bit image of 256 × 256 pixels. b. The smoothed image for $\mu = 9.0$ $\gamma_{ij} = 256$ and 1000 iterations. c. The corresponding line process.

implies that if a line is created it will inhibit a creation of a line along the gradient field and therefore this gives rise to nonmaximum suppression. It should be emphasized that the discretized (lattice) total energy function can be made bounded below, despite the negative sign in (25), by choosing ξ sufficiently small.

9 Conclusion

One of the chief goals of this article was an attempt to unify different methods of image segmentation. We used MF theory [Geiger & Girosi 1991] to show that several deterministic approaches were essentially equivalent and were closely related to the statistical approaches (Markov random fields). By introducing the concept of parameter space we could relate these energy function methods to alternative approaches using nonlinear diffusion equations or minimal description length. An overview is provided by figure 8.

An important advantage of the mean field approach is that it enables to integrate out the line process fields and get an effective energy that depends only on the smoothed image *f*. This can be used to show that some, apparently different, energy functions are closely related. For example, it was shown [Geiger & Girosi 1991] that the graduated non-convexity algorithm [Blake & Zisserman 1987] can be directly related to the Geman

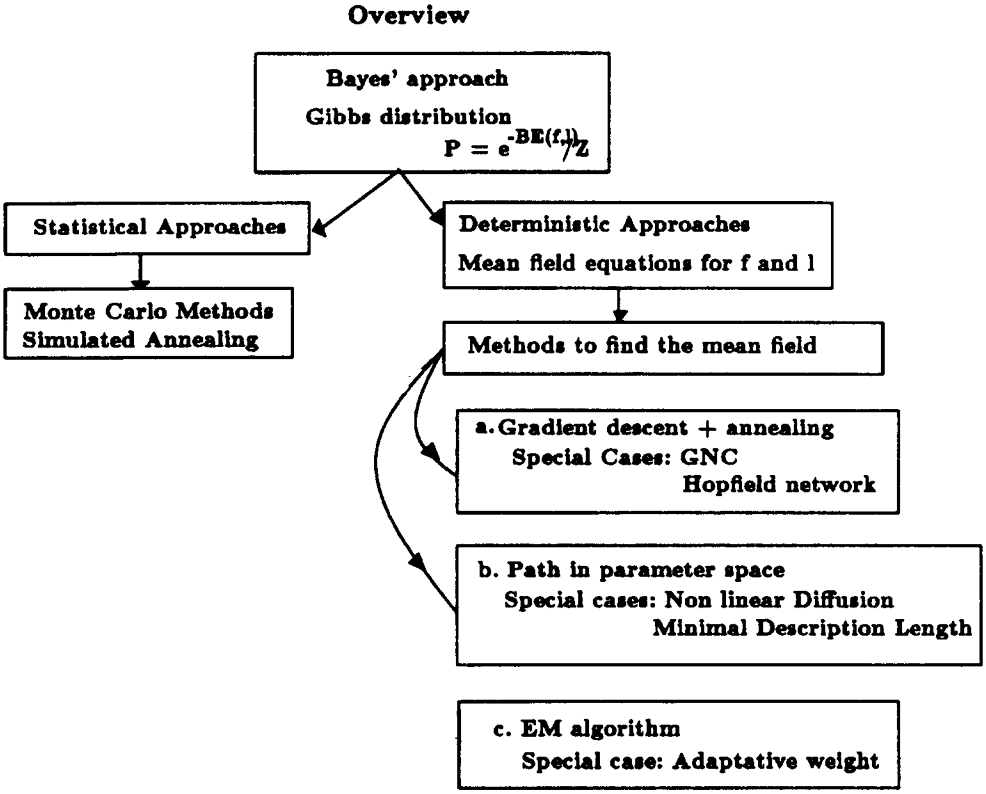
Overview
Bayes' approach
Gibbs distribution
P = e^{-BE(f,l)}/Z
Statistical Approaches
Monte Carlo Methods
Simulated Annealing
Deterministic Approaches
Mean field equations for f and l
Methods to find the mean field
a. Gradient descent + annealing
Special Cases: GNC
Hopfield network
b. Path in parameter space
Special cases: Non linear Diffusion
Minimal Description Length
c. EM algorithm
Special case: Adaptative weight

Fig. 8. An overview of the relations between different methods of image segmentation.

and Geman [1984] approach. Similarly it can also be shown that the effective energy terms used by Geman and McClure [1987] can be related to Geman and Geman [1984] by the mean field approach (although the data terms are rather different).

The relation between energy functions and nonlinear diffusion approaches may lead to new efficient ways to obtain approximate solutions. This is supported by recent work by Nitzberg and Shiota [1990] on a nonlinear filter which preserves edges. They relate their filter both to that of Perona and Malik [1987] and to the adaptive filtering for noise cleaning in television pictures [Graham 1962].

We also introduced the idea of a single line process field on a lattice, instead of the conventional pair of horizontal and vertical line process fields. This helped us relate to the continuous formulations of Mumford and Shah [1985] and Ambrosio [1988]. By encouraging the creation of diagonal lines it also decreased the bias toward horizontal and vertical lines found in most lattice formulations of the problem. We point out that we had to redefine discrete derivatives in order to preserve the symmetry of both diagonals' interactions.

We also showed that the energy function, or Markov random field, approach can be easily extended to give robust methods [Huber 1981] for smoothing with discontinuities. It could also incorporate many of the basic elements used in most edge detectors, such as nonmaximal suppression and hysteresis.

Finally we believe that these results support the view that the Bayesian approach, including Markov random fields as a special case, is sufficiently rich to give an overall theoretical framework for early vision. We stress that most of the work described here assumed a specific Markov model corresponding to Gaussian noise in the data and a specific smoothness assumption. If this assumption is invalid then other Markov models should be considered, for example our robust model in section 7.1.

Acknowledgments

We would like to thank J. Clark, A. Elfadel, N. Grzywacz, B. Horn, C. Marcus, D. Mumford, M. Nitzberg, T. Poggio, and J. Wyatt for helpful conversations. We also thank the reviewers for their careful comments and suggestions. A.L.Y. would like to thank the Brown, Harvard, and M.I.T. Center for Intelligent Control Systems for a United States Army Research Office grant number DAAL03-86-C-0171 and an exceptionally stimulating working environment. D.G. would like to thank the exceptionally stimulating working environment of the MIT AI lab.

References

Ambrosio, L. 1988. Variational problems in SBV. Technical Report Lectures Notes, M.I.T., Cambridge, MA

Berger, J.O. 1985. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag: New York.

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

Canny, J.F. 1986. A computational approach to edge detection. *IEEE Trans. Patt. Anal. Mach. Intell.*, PAMI-8(6): 679–698.

Chou, P.B., and Brown, C.M. 1988. Multimodal reconstruction and segmentation with Markov random fields and HCF optimization. *Proc. Image Underst. Work.* Morgan Kaufmann, San Mateo, CA, Cambridge, MA, February, pp. 214–221.

Durbin, R., and Willshaw, D.J. 1987. An analogue approach to the travelling salesman problem using an elastic net method. *Nature*, 326: 689–691.

Gamble, E.B., and Poggio, T. 1987. Visual integration and detection of discontinuities: The key role of intensity edges. A.I. Memo No. 970, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, October.

Geiger, D., and Girosi, F. 1991. Parallel and deterministic algorithms for MRFs: Surface reconstruction and integration. PAMI-13(5), May. Also A.I. Memo No. 1114, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge MA.

Geiger, D., and Pereira, R.M. 1990. On intelligent sparse data. *IEEE 7th Israeli Symp. Artif. Intell. Comput. Vision*, Haifa, Israel.

Geiger, D., and Poggio, T. 1987. An optimal scale for edge detection. *10th Intern. Joint Conf. Artif Intell.*, Milan, August.

Geman, S., and Geman, D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Patt. Anal. Mach. Intell.* PAMI-6: 721–741.

Geman, S., and McClure, D.E. 1987. Statistical methods for tomographic image reconstruction. *Proc. 46th Session of the ISI*, Bulletin of ISI.

Girosi, F., Poggio, T., and Caprile, B. 1989. Extensions of a theory of networks for approximation and learning: Outliers and negative examples. Technical Report, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

Graham, R.E. 1962. Snow removal: A noise-stripping process for TV signals. *IRE Trans. Inform. Theory* 9: 129–144.

Grimson, W.E.L., and Pavlidis, T. 1985. Discontinuity detection for visual surface reconstruction. *Comput. Vision, Graph. Image Process.* 30: 316–330.

Hallinan, P.W., and Mumford, D. 1990. In preparation.

Hopfield, J.J. 1984. Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Nat. Acad. Sci.* 81: 3088–3092.

Hopfield, J.J., and Tank, D. 1985. Neural computation of decision in optimization problems. *Biological Cybernetics* 52: 141–152.

Huber, P.J. 1981. *Robust Statistics.* John Wiley: New York.

Keeler, K.C. 1990. Map representation and optimal encoding for image segmentation. Ph.D. thesis.

Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P. 1983. Optimization by simulated annealing. *Science* 220: 219–227.

Koch, C., Marroquin, C., and Yuille, A. 1985. Analog 'neuronal' networks in early vision. *Proc. Natl. Acad. Sci.* 83: 4263–4267.

Leclerc, Y.G. 1988. Constructing simple stable descriptions for image partitioning. *Proc. Image Underst. Work.* 1: 365–382.

Lee, D., and Pavlidis, T. 1987. One-dimensional regularization with discontinuities. *Proc 1st Intern. Conf. Comput. Vision*, London, June.

Marroquin, J.L. 1985. Probabilistic solution of inverse problems. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Marroquin, J.L. 1987. Deterministic Bayesian estimation of Markovian random fields with applications to computational vision. *Proc. 1st Intern. Conf. Comput. Vision*, London, England, June.

Marroquin, J.L., Mitter, S., and Poggio, T. 1985. Probabilistic solution of ill-posed problems in computational vision. In L. Baumann, ed. *Proc. Image Underst. Work.*, Scientific Applications International Corporation, McLean, VA, August, pp. 293–309.

McKendall, R., and Mintz, M. 1989. *Robust Fusion of Location Information.* Preprint. Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. 1953. Equation of state calculations by fast computing machines. *J. Phys. Chem.* 21: 1087.

Mumford, D., and Shah, J. 1985. Boundary detection by minimizing functionals. *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, San Francisco, CA.

Nitzberg, M., and Shiota, T. 1990. Nonlinear image smoothing with edge and corner enhancement. Technical Report 90-2, Harvard Robotics Laboratory.

Nordstrom, N.K. 1990. Variational edge detection. Ph.D. thesis, University of California at Berkeley.

Parisi, G. 1988. *Statistical Field Theory.* Addison-Wesley: Reading MA.

Pavlidis, T. 1986. A critical survey of image analysis methods. *Proc. 8th Intern. Conf. Patt. Recog.*, Paris.

Perona, P., and Malik, J. 1987. Scale space and edge detection using anisotropic diffusion. *Proc. 5th IEEE Work. Comput. vision*, Miami.

Richardson, T. 1990. Department of electrical engineering and computer science. Ph.D thesis, Massachusetts Institute of Technology, Cambridge MA.

Rissanen, J. 1978. Modeling by shortest data description. *Automatica* 14: 465–471.

Terzopoulos, d. 1986. Regularization of inverse visual problems involving discontinuities. *IEEE Trans. Patt. Anal. Mach. Intell.* PAMI-8: 413–424.

Wasserstrom, E. 1973. Numeral solutions by the continuation method. *SIAM Review* 15: 89–119.

Witkin, A.P., 1983. Scale-space filtering. *Proc. 8th Intern. Joint Conf. Artif. Intell., Karlsruhe,* pp. 1019–1022.

Yuille, A.L. 1987. Energy functions for early vision and analog networks. A.I. Memo No. 987, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, October.

Yuille, A.L., and Poggio, T. 1983. Fingerprint theorems for zero crossings. A.I. Memo No. 730, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

Yuille, A.L., Yang, T., and Geiger, D. 1990. Towards a theory of transparency. Harvard Robot. Lab. Tech. Report (in preparation), Harvard University, Cambridge MA.