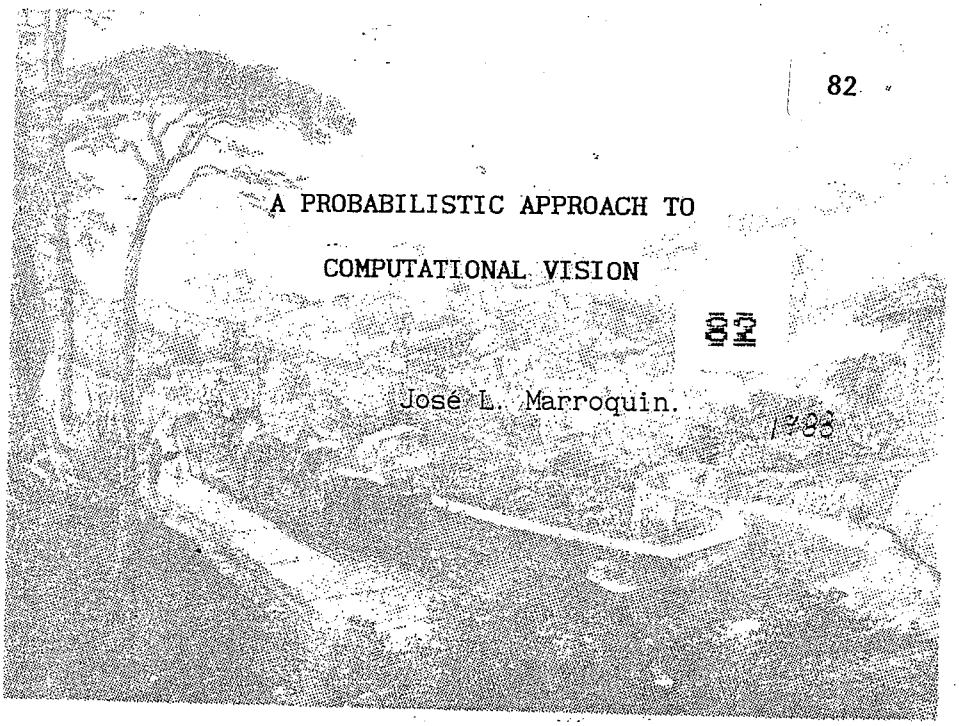
# COMUNICACIONES DEL CIMAT



82

A PROBABILISTIC APPROACH TO

COMPUTATIONAL VISION

82

José L. Marroquin.

1983

## CENTRO DE INVESTIGACION EN MATEMATICAS

Apartado Postal 402
Guanajuato, Gto.
México
Tels. (473) 2-25-50
2-02-58

# A PROBABILISTIC APPROACH TO COMPUTATIONAL VISION

Jose L. Marroquin.
Centro de Investigacion en Matematicas.
Apdo. Postal 402 ; Guanajuato, Gto. ; MEXICO.

## 1. Introduction.

The study of what has been called "Early Vision" (Brown, 1984; Brady, 1981; Barrow and Tannenbaum, 1981; Poggio,1984) has, as one of its goals, the elucidation of the mechanisms by which a living being constructs internal representations of the physical structures of the "outside world" - namely, the surfaces of objects and a picture of the way they move.

A good way of improving our understanding of these mysterious biological processes, is by studying the problem they are supposed to solve, from a purely mathematical (computational) viewpoint (see Marr, 1982). From this perspective, the problem becomes one of designing a distributed algorithm (see note [1]) for the reconstruction of a function on the sites of a two-dimensional, finite lattice or "Retina", given some observations that constrain its value. This function can be real-valued (i.e., a "surface" or an "intrinsic image"), as in the case of the recovery of: depth from stereoscopic pairs of images; lightness; shape from shading; the restoration and segmentation of images and the formation of perceptual clusters. It can also be vector-valued, as in the case of the recovery of the velocity field from succesive frames of the same scene, etc.

An important, common characteristic of this class of problems, is that the observations are so noisy or incomplete, that it is not possible, based only on them, to determine the solution in a unique way ( in mathematical terms, we say that the problem is ill--posed (Poggio and Torre, 1984), or that the mapping, from the original function to the observations is many-to-one). This means that, in order to find a unique

solution, it is necessary to use, in some appropriate way, prior generic knowledge about the behavior of the solution (for example, we may use the assumption that the solution should be smooth, or that it should be piecewise constant, etc.).

One possible way in which this prior knowledge can be included, is to define a functional (a variational principle) that expresses the desired global condition (such as smoothness), and then to define the solution as the global maximizer of this functional, subject to the constraints imposed by the observations (see, for example, Horn, 1974 and 1981; Hildreth, 1984; Blake, 1985). When these constraints are linear, Poggio et al (1984, 1985) have shown that this functional can be constructed in a systematic way, using the so called "Standard Regularization" method (Tikhonov, 1977). They have applied this construction to the formulation of a variety of Early Vision problems, and have also proposed some extensions to cover some non-linear cases as well.

There is, however, another powerful mathematical tool that deals with the use of prior knowledge for the reconstruction of unknown fuctions, namely, statistical - in particular, Bayesian - Estimation Theory. To use it, we must formulate the original reconstruction task as an estimation problem, which means that we have to propose a probabilistic model for the observations (that is, we must assume that they are corrupted by noise whose statistical properties are, at least partially, known). Our prior knowledge about the behavior of the solution must also be expressed in probabilistic terms, i.e., we must construct a probability distribution , on the space of all possible solutions, that reflects the fact that some functions (for example, the smooth ones) are more likely to occur than others.

The increased effort involved in formulating the problem in this terms, has a definite payoff: we can use the machinery of Estimation Theory to construct efficient distributed algorithms that perform reconstructions that are optimal, with respect to very natural cost functionals (such as the expected value of the reconstruction error).

This approach has other advantages as well:

Firstly, it provides a general framework, not only for the separate formulation of a wide variety of perceptual problems, but also for the incorporation of qualitatively different measurements in a single cooperative estimation process; thus, it can be used to construct models for the interaction of "perceptual modules" that so far have been trated in an independent way (see Marroquin, 1985; Poggio, 1985).

Secondly, the parameters that appear in the reconstruction algorithms that are derived using this approach, have a precise interpretation (for example, the relative weight of the evidence provided by each set of observations, is determined by the variance of the associated noise process), and its optimal value can, in principle, be determined by statistical methods.

Finally, the plausibility of the prior assumptions about the behavior of the solution, can be explicitely verified by generating sample functions of the corresponding random field, by means of a Monte Carlo procedure.

## 2. Probabilistic Models.

### 2.1. General Definitions.

The basic problem in which we are interested is the reconstruction of the values of a function on the N sites of a finite, regular lattice $\mathcal{L}$. To formulate this problem in probabilistic terms, we need the concept of a random field defined on $\mathcal{L}$, that is, a collection $\mathcal{F}$ of random variables indexed by the sites of $\mathcal{L}$:

$$\mathcal{F} = \{ F_i \ , \ i \in \mathcal{L} \}$$

Suppose that each one of these random variables $F_i$, can take values on some set $Q_i$. We will call any possible sample realization $f = \{ f_i \ , i \in \mathcal{L} \}$, with $f_i \in Q_i$, for all $i \in \mathcal{L}$, a "configuration" of the field (the set of all valid configurations is called the "sample space" $\Omega$).

If we assume that any valid configuration can occur with positive probability, we can always write the probability distribution of the configurations (i.e., the joint probability density of the random variables $\{ F_i , i \in L \}$) in the "Standard Gibbs form" (see note [2]):

$$P_F ( f) = ( 1 / Z ) \exp [ - U( f) / T] \qquad (1)$$

where Z is a normalizing constant; U is called the "energy function", and T is a parameter.

As an example, consider a field of N independent, identically distributed, zero mean, Gaussian random variables. In this case,

$$Z = (2 \pi \sigma^2)^{N/2} ; \qquad U( f) = \sum_{i \in L} f_i^2 \qquad \text{and } T = 2\sigma^2$$

Note that for this simple model, it is straightforward to generate a sample function: we only have to generate N independent, identically distributed (i.i.d.) Gaussian pseudo random numbers.

## 2.2. Models for the Observations.

Suppose that a sample configuration $f$ of the field is given. We will assume that the observations $g$ are obtained from $f$ by a degrading operation (such as sampling or blurring), and by combining the resulting degraded field with a noise field formed by i.i.d. random variables. This combining operation can take several forms; it can be, for example, addition or multiplication, or it can correspond to an error being commited in the transmission of each value of $f$, with certain probability. In any case, we require that it be reversible, in the sense that it should be possible to obtain the value of the noise from the values of $f$ and $g$. If this is the case, the conditional distribution of $g$ given $f$, $P_{g|f}$, can be obtained directly from the noise distribution; assuming that the latter can be written in the standard Gibbs form, we

get:

$$P_{g|f} = (1 / Z_C) \exp [ - \alpha \Phi( g ; f) ] \qquad (2)$$

where $\alpha$ is called the noise parameter, and $\Phi$, the noise statistic.

Two examples will clarify this representation:

a) Suppose that the observations are obtained by adding to the value of $f$, at the sites of a subset S of $\mathcal{L}$, zero mean, i.i.d. Gaussian noise of variance $\sigma^2$. In this case, $Z_C = (2\pi\sigma^2)^{m/2}$ ; $\alpha = 1/(2\sigma^2)$ , and the noise statistic corresponds to the squared "distance" between the fields $f$ and $g$ , taken over the sites where an observation is present:

$$\Phi( f ; g) = \sum_{i \in S} (f_i - g_i)^2 \qquad (3)$$

b) Suppose that $f$ is a binary field, and that the observations are the output of a Binary Symmetric Channel (BSC) with error rate $\rho$ (see Gallager, 1975), so that

$$\Pr (g_i | f_i) = \begin{cases} (1-\rho) & , \text{ if } f_i = g_i \\ \rho & , \text{ if } f_i \neq g_i \end{cases} \quad , i \in S$$

In this case, $Z_C = 1$ ;

$$\alpha = \ln [ (1-\rho)/\rho ] \quad , \qquad (4)$$

and the noise statistic corresponds to the number of sites where an error has been committed, i.e.,

$$\Phi = \sum_{i \in S} (1-\delta(f_i - g_i)) \qquad (5)$$

$$\text{where } \delta(x) = \begin{cases} 1, & \text{if } x=0 \\ 0, & \text{otherwise} \end{cases} \qquad (6)$$

## 2.3. Models for the Solutions: Markov Random Fields.

Let us now consider the problem of constructing a probability distribution on the space of all possible solutions (i.e., of random fields defined on $\mathcal{L}$) that correspond to a given qualitative behavior. The simplest way to accomplish this, is to specify the probabilistic dependencies between different elements of the field. Since we want to be able to process images that consist of separate objects (i.e., to reconstruct surfaces that are, for example, piecewise smooth), we are mostly interested in fields where these dependencies are local , in the sense that the value of the field at each location depends, probabilistically, only on a small set of neighboring sites (the "neighborhood" of a given location). These fields are called Markovian, and the probability distribution of their configurations has the property that, when written in the standard Gibbs form (1), the energy function U can always be written as the sum of a set of "Potential Functions" which are supported on the "cliques" of the neighborhood system of the field ( a "clique" is either a single site, or a set of sites such that any two sites belonging to it are neighbors of each other) :

$$U = \sum_c V_c(f) \qquad (7)$$

where c ranges over the cliques of the system, and each function $V_c$ has c as its support.

In this way, the desired probability distribution can be constructed simply by specifying the neighborhood system and the local potential functions. As an example, the behavior of piecewise constant surfaces on a square lattice can be modeled in the following way: the neighborhood of a site i of the lattice is defined as its nearest neighbors, i.e.,

$N_i = \{j : \|i-j\| = 1\}$. Note that, for an interior site, the size of this neighborhood will be 4; if the site lies at the boundary, but not at a corner, it will be 3, and for the corners , it will be 2 (this field is called a "first order MRF with free boundaries"). For the potential functions, we use the generalized Ising potentials (Ising, 1925; Geman and Geman, 1984):

$$V_C(f_i,f_j) = \begin{cases} -1 \text{ , if } \|i-j\| = 1 \text{ and } f_i = f_j \\ +1 \text{ , if } \|i-j\| = 1 \text{ and } f_i \neq f_j \\ 0 \text{ , otherwise} \end{cases} \qquad (8)$$

where we assume that $f_i \in \{q_1, q_2, \ldots, q_m\}$, for all i. The Gibbs distribution (1), with U given by (7) and (8), defines a one parameter family of models (indexed by T), describing piecewise constant patterns with varying degrees of granularity.

## 2.4. Generation of Sample Configurations of MRF's.

In this section we present some algorithms for generating configurations that are random, but whose values satisfy the probabilistic dependencies of the given MRF, i.e. , sample functions from the Gibbs distribution (1), with U given by (7). To do this, we identify the value of the field at each site with the state of a particle of a hypothetical physical system whose energy is given by U. If this system is allowed to reach thermal equilibrium at temperature T, its configurations will be distributed according to (1) (Reif, 1965); the equilibrium behavior of such a system can be modeled by the steady state of a Markov chain (whose states correspond to instantaneous configurations of the field), provided that:

a) The chain is regular (which means that any two allowed configurations are mutually accessible with positive probability).

b) It has (1) as its invariant measure (Kemeny and Snell, 1960).

After the work of Metropolis et al (1956), several algorithms have been

proposed for generating this chain (specifically, the Heat Bath algorithm (Hastings, 1970) and the Gibbs Sampler (Geman and Geman, 1984)). The basic idea in all of them, is to visit every site of the lattice in some prescribed (deterministic or pseudorandom) order, and, when a site is visited, to generate, randomly, a new value for its state, using a transition probability that depends on the current state of its neighbors, and that guarantees the invariance of the distribution (1). The Metropolis and "Heat Bath" algorithms generate this transition probability by selecting a new "candidate state" at random from the set of allowable states (using a uniform distribution), and then accepting or rejecting this transition, with a probability that depends on the value of the energy increment associated with it. The Gibbs Sampler, on the other hand, equates the transition probability to the conditional distribution for the state of the visited site, given the state of its neighbors (we will discuss this algorithm in detail in section 4).

The Metropolis algorithm itself is efficient and simple to implement in a serial computer; in a parallel machine, however, one would like to update the state of all the non-neighboring sites at the same time, and in this case, the regularity of the Metropolis chain is lost, so that it is no longer possible to guarantee its convergence to the desired Gibbs distribution (see note [3]). Therefore, if a parallel implementation is desired, either the Heat Bath or the Gibbs Sampler schemes must be used (a more detailed description of all these algorithms, and a discussion of their serial and parallel complexity may be found in Marroquin, 1985).

These algorithms, not only provide a practical way of checking the appropriatness of the proposed MRF for modeling the desired qualitative behavior of the field (and for the calibration of the parameters of the model), but more importantly, they can be used for the computation of the optimal solutions to the reconstruction problem, as we will explain in the next section. Their use is illustrated in figure 1, where we show three typical configurations of a first order MRF, obtained after 200 iterations of the HBA, for different values of the interaction parameters.

---

Figure 1 around here

---

## 3. Bayesian Estimation.

Once we have formulated the reconstruction problem in probabilistic terms, its solution can be found as the optimal Bayesian estimator for the Markovian field $f$, given the observations $g$. This estimator is defined as the minimizer of the expected value (taken with respect to the posterior distribution $P_{f|g}$) of a given cost functional. The posterior distribution is found using Bayes rule:

$$P_{f|g} = \frac{P_{g|f} \cdot P_f}{P_g}$$

Since for a given set of observations $P_g$ is just a normalizing constant, and $P_f$ has the form:

$$P_f = (1 / Z) \exp [ - U_0/T_0 ]$$

with $U_0$ of the form (7), we can use (2) to write $P_{f|g}$ in the standard Gibbs form :

$$P_{f|g} = (1 / Z_p(\alpha,T_0)) \exp [ - U_0/T_0 - \alpha \Phi ] \quad (10)$$

This Gibbs distribution corresponds to the equilibrium behavior of a physical system in which the original field is coupled with a fixed, external field whose intensity is given by the observations. The interaction term is given by $\Phi(f,g)$, and the coupling strength by $\alpha$.

## 3.1. Cost Functionals.

A cost functional $C(f, \hat{f})$ measures how close is an estimated configuration $\hat{f}$ from the true one $f$. The one that has been most widely used (although never in an explicit way) is one that is equal to zero only if $f_i = \hat{f}_i$, for all i, and is equal to some positive number M otherwise. Since minimizing the expected value of such cost functional, with respect to $P_{f|g}$, is equivalent to maximizing $P_{f|g}$ itself, the corresponding maximizer is called the Maximum a Posteriori (or MAP) estimator.

This estimator works reasonably well if the signal to noise ratio is relatively high. In the high noise situation, however, it tends to be too conservative (in the sense that it practically disregards the observations, and relies almost exclusively on the prior generic knowledge), since from its viewpoint it is equally costly to make one, or one thousand mistakes.

A better approach is to define, for each particular problem, a cost functional that is in closer agreement with one's intuitive assessment of the performance of the estimator. As an example, we will now propose cost functionals (and derive the corresponding estimators) for two particular problems: image segmentation and surface reconstruction.

Consider a field $f$ with N elements, each of which can belong to one of a finite set $Q_i$ of classes. Let $f_i$ denote the class to which the $i^{th}$ element belongs. The segmentation problem is to estimate $f$ from a set of observations $\{g_1,..., g_p\}$. Note that $f_i$ does not necessarily correspond to the image intensity. It may represent, for example, the texture class for a region in an image (as in Elliot et. al., 1983), etc.

A reasonable criterion for the performance of an estimate $\hat{f}$ is the number of elements that are not classified correctly. Therefore, we can define the segmentation error as:

$$e_S(f, \hat{f}) = \sum_{i=1}^{N} (1 - \delta(f_i - \hat{f}_i)) \quad , \quad f_i , \hat{f}_i \in Q_i \qquad (11)$$

where $\delta(\cdot)$ was defined by (6).

In the case of the reconstruction problem, an estimate $\hat{f}$ should be considered "good" if it is close to $f$ in the ordinary sense, so that the total squared error:

$$e_r(f, \hat{f}) = \sum_{i=1}^{N} (f_i - \hat{f}_i)^2 \qquad (12)$$

is a reasonable measure for its performance.

To derive the optimal estimators with respect to the criteria stated above, we first present the general result (which can be found, for example, in Abend, 1968) which states that if the posterior marginal distributions for every element of the field are known, the optimal Bayesian estimator with respect to any additive, positive definite cost functional C may be found by minimizing independently the marginal expected cost for each element.

In more precise terms, we consider cost functionals of the form:

$$C(f, \hat{f}) = \sum_{i \in L} C_i(f_i, \hat{f}_i) \qquad (13)$$

with

$$C_i(a,b) \begin{cases} = 0 , & \text{if } a = b \\ > 0 , & \text{if } a \neq b , \text{ for all } i. \end{cases}$$

We will assume the value of each element $f_i$ of the field is constrained to belong to some finite set $Q_i$ (the generalization to the case of compact sets is straightforward). The optimal Bayesian estimator $f^*$

with respect to the cost functional C, and given some observations $g$, is defined as the global minimizer of the expected value of C over all possible $f$ and $g$. One can prove that this estimate can be found by minimizing independently the marginal expected cost for each element, i.e.,

$$f^*_i = q \;:\; \sum_{r \in Qi} C_i(r,q)\, P_i(r) < \sum_{r \in Qi} C_i(r,s)\, P_i(r)$$
$$\text{for all } s \neq q \text{ and all } i \in L$$

where $P_i(r)$ is the marginal posterior distribution of the element i:

$$P_i(r) = \sum_{f:\, f_i = r} P_{f|g}(f\,;g) \qquad (14)$$

The optimal estimators for the error criteria defined above, can be easily derived from this result (see Marroquin, 1985):

In the case of the segmentation problem, we get that:

$$f^*_i = q \in Q_i \;:\; P_i(q) > P_i(r) \qquad (15)$$
$$\text{for all } r \neq q$$

We will call this estimate the "Maximizer of the Posterior Marginals" ($f_{MPM}$).

For the reconstruction problem, the optimal estimator is:

$$f^*_i = q \in Q_i \;:\; (\overline{f_i} - q)^2 < (\overline{f_i} - s)^2 \qquad (16)$$
$$\text{for all } s \neq q$$

We will call this estimate the "Thresholded Posterior Mean" ($f_{TPM}$).

To illustrate the difference between the MAP and these estimators, let us consider the following simple example:

Suppose we have a one-dimensional, binary field $f$ with Ising potentials

(see equation (6)), on a lattice that has only two sites ( so that the set of all possible configurations of the field is $\Omega = \{00,10,01,11\}$ ), and suppose that the observations $g$ are the output of a BSC with a given error rate. Using equations (2), (4), (5) and (6), we can compute the posterior distribution:

$$P_{f|g}(f;g) = (1/Z_p) \exp\left[ -(1/T_0) V_C(f_1,f_2) - \alpha \Phi(f,g) \right]$$

Suppose, in particular, that $g_1 = 1$ and $g_2 = 0$. If the SNR is relatively low (especifically, if $\alpha < 2/T_0$ ), we will have that:

$$P_{f|g}(11) = P_{f|g}(00) > P_{f|g}(10) > P_{f|g}(01)$$

so that the MAP estimator in this case is not unique, and is either 11 or 00.

On the other hand, the posterior marginals are:

$$P_1(1) = P_{f|g}(11) + P_{f|g}(10) \ ;$$
$$P_1(0) = P_{f|g}(01) + P_{f|g}(00) \ ;$$
$$P_2(1) = P_{f|g}(11) + P_{f|g}(01) \ ;$$
$$P_2(0) = P_{f|g}(00) + P_{f|g}(10) \ .$$

Clearly, $P_1(1) > P_1(0)$ and $P_2(0) > P_2(1)$ , so that the MPM estimator is 10 .

For larger lattices, the situation is similar, and the differences in performance, more dramatic (particularly for low SNR). This is illustrated in the example portrayed in figure 2 : panel (a) represents a typical realization of a 64 by 64 Ising net with free boundaries, using a value of $T_0 = 1.74$ ; panel (b), the output of a BSC with error rate equal to 0.4 ; panel (c), the MAP estimate, and panel (d), an approximation to the MPM estimate, which is clearly better than the MAP from almost any viewpoint.

Figure 2 around here

## 3.2. General Monte Carlo Algorithms.

All the estimators that are optimal with respect to cost functionals of the form (13) are easy to compute, once the marginal posterior probabilities are known. These marginal distributions, in turn, are ensemble averages (with respect to $P_{f|g}$) of $\delta$ functions (see equation (6)):

$$P_i(r) = \sum_f \delta(f_i - r)\, P_{f|g}(f;g)$$

(in some particular cases, such as the TPM estimator, it may not be necessary to compute the marginals, since the estimator is easily computed directly from the ensemble average of $f$). In any case, the computation of optimal estimators can be reduced to the approximation of ensemble averages, with respect to $P_{f|g}$, of specific functions. To perform this approximation, we recall that, using the Heat Bath or Gibbs Sampler algorithms, it is possible to generate a regular Markov chain that has $P_{f|g}$ as its equilibrium distribution. The law of large numbers for regular chains (see, for example, Kemeny and Snell, 1960) establishes that the fraction of time that the chain spends on a given state $f$, will tend to $P_{f|g}(f;g)$ as the number of time steps gets large, independently of the initial state (i.e., we can approximate ensemble averages by time averages), so that the posterior marginals may be approximated by:

$$P_i(q) \approx (1/(n-k)) \sum_{t=k}^{n} \delta(f_i^{(t)} - q) \qquad (17)$$

and the posterior mean by:

$$\bar{f}_i \approx (1/(n-k)) \sum_{t=k}^{n} f_i^{(t)} \qquad (18)$$

where $f^{(t)}$ is the configuration generated by the Monte Carlo algorithm at time t, and k is the time required for the system to reach equilibrium.

A similar Monte Carlo procedure can be used to approximate the MAP estimator (Geman and Geman, 1983). In this case, the associated cost functional is not of the form (9), and thus, $f_{MAP}$ cannot be obtained directly in terms of ensemble averages. It is possible, however, to introduce a new "temperature" parameter and form a family of distributions:

$$P_T = (1/Z_p) \exp [ -(1/T) (U_0/T_0 + \alpha \Phi) ]$$

(note that $P_T$ coincides with $P_{f|g}$ when T = 1). It can be shown the global maximizer of $P_{f|g}$ will correspond to the equilibrium configuration of a chain that has $P_T$ as its invariant distribution, as T→0. The method for finding this equilibrium configuration (the "ground state" of the system) is called "Simulated Annealing" (Kirkpatrick at. al., 1983), and it consists in slowly decreasing the "temperature" parameter T while the system (the Heat Bath or Gibbs Sampler chain) is mantained in equilibrium. It should be noted that this process will be, in general, computationally more expensive than the approximation of the ensemble average of a function at a fixed temperature T = 1.0. Besides, since in the latter case we are using a Monte Carlo prcedure for approximating the value of some integrals, we should expect a nice convergence behavior, in the sense that coarse approximations can be computed very fast, and then refined to an arbitrary precision (in fact, it can be proved (see Feller, 1950) that the expected value of the squared error of the

estimates (17) and (18) is inversely proportional to n-k).

Finally, we point out that, since maximizing $P_{f|g}$ is equivalent to the minimization of the posterior energy:

$$U_p = U_0/T_0 + \alpha \Phi \ ,$$

the MAP estimator can be considered a generalization of the regularization method for solving this class of problems (see Poggio et. al. , 1984). $\Phi$ represents the constraints generated by the observations, and $U_0$, the smoothness assumption. If these terms are quadratic (which corresponds to a Gaussian assumption, both for the noise and for the field), the MAP estimate coincides with the "Standard Regularization" solution, and its value can be found by efficient, deterministic methods (see Terzopoulos, 1985).

## 4. Deterministic Cooperative Networks.

We will now discuss a general procedure for constructing deterministic cooperative networks that can be used for obtaining good approximations to the optimal estimators for Markovian fields with finite state space (i.e., when the random variables of the field take values only on a finite set $Q = \{q_1,..., q_M\}$ ).

As we mentioned above, the optimal estimators for these fields can be easily obtained once the marginal probabilities $P_i$ (q) are known; these marginals correspond, in the Monte Carlo scheme, to the fraction of time during which the variable associated with each site i takes a particular value $q \epsilon Q$, as the Markov chain generated by the algorithm evolves. We will now use a "mean field" approximation to construct a deterministic network that simulates the evolution of these marginal frequencies.

To understand how this is done , let us consider the Gibbs Sampler algorithm in detail: it can be represented by an M-layer binary stochastic network F, with each layer corresponding to an allowable value for the

random variables of the field $f$ (thus, if $f_i^{(t)} = q$, we have $F_{i,q}^{(t)} = 1$, and $F_{i,r}^{(t)} = 0$, for $r \neq q$). Each site of the lattice (each column of F) is visited in turn. When the $i^{th}$ site is visited (say, at time t), the conditional probabilities $w_i^{(t)}$ are computed using the expression:

$$w_i^{(t)}(q) = \Pr\left(f_i^{(t)} = q \mid f_j^{(t)},\ j \neq i\right) = \frac{\exp\left[-u_i^{(t)}(q)\right]}{\sum_{r \in Q} \exp\left[-u_i^{(t)}(r)\right]}$$

where $u_i^{(t)}(q)$ is the local energy or "excitation" received by the (i,q) cell of the network F at time t. For example, for a first order field, we have:

$$u_i^{(t)}(q) = \sum_{j \in Ni} V_C\left(q, f_j^{(t)}\right) + \alpha\, \Phi_i(q) \qquad , \quad q \in Q$$

where Ni is the neighborhood of site i; $V_C$ are the potentials for the MRF model for $f$, and $\Phi_i$ is the local noise term (for example, in the case of additive, white Gaussian noise, if there is an observation present at site i, we have that $\Phi_i(q) = (q - g_i)^2$).

Next, a random number n, distributed according to $w_i^{(t)}$, is generated. Suppose $n = q$. Then, we put $f_i^{(t+1)} = q$ (i.e., $F_{i,q}^{(t+1)} = 1$, and $F_{i,r}^{(t+1)} = 0$, for $r \neq q$). Finally, the estimates for the marginal probabilities are updated using the expression :

$$P_i^{(t+1)}(r) = \lambda\, P_i^{(t)}(r) + (1 - \lambda)\, F_{i,r}^{(t)} \qquad , r \in Q$$

where $\lambda \in [0,1]$ is a parameter related with the size of the time window used for the update.

In the deterministic scheme, we model the average behavior of the stochastic network, so that the estimate for $P_i(q)$ is replaced by its (conditional) expected value. The update expression becomes:

$$P_i^{(t+1)}(r) = E\,[P_i^{(t+1)} \mid P^{(t)}] = \lambda\, P_i^{(t)}(r) + (1-\lambda)\, E\,[F_{i,r} \mid P^{(t)}]\ ,\ r \in Q$$

since $\quad E\,[F_{i,r} \mid P^{(t)}] = E\,[\,E\,[F_{i,r} \mid w_i^{(t)}(r)]\,\mid P^{(t)}]$

and $\quad E\,[\,F_{i,r} \mid w_i^{(t)}(r)\,] = 1 \cdot w_i^{(t)}(r) + 0 \cdot (1-w_i^{(t)}(r)) = w_i^{(t)}(r)$

so that $\quad E\,[\,F_{i,r} \mid P^{(t)}] = E\,[w_i^{(t)}(r) \mid P^{(t)}] = \overline{w}_i(r)\ ,$

we finally get :

$$P_i^{(t+1)}(r) = \lambda\, P_i^{(t)}(r) + (1-\lambda)\, \overline{w}_i(r) \qquad (19)$$

$\overline{w}_i(r)$ can be approximated using a "mean field" assumption, which in this case takes the form:

$$\overline{w}_i^{(t)}(q) = \frac{\exp\,[-\overline{u}_i^{(t)}(q)]}{\sum_{r \in Q} \exp\,[-\overline{u}_i^{(t)}(r)\,]} \qquad (20)$$

where $\overline{u}_i^{(t)}(q)$ is the average excitation of cell (i,q) at time t :

$$\overline{u}_i^{(t)}(q) = E\,[u_i(q) \mid P^{(t)}] = \sum_{j \in Ni} \sum_{r \in Q} V_c(q,r)\, P_j^{(t)}(r) + \propto \Phi_i(q)$$

It is also possible to use a weaker assumption to approximate $\overline{w}$, namely, the linear independence of non-neighboring sites of the field $f$. Using this assumption, one gets the formula:

$$\overline{w}_i^{(t)}(q) = \frac{\exp[-\propto \Phi_i(q)]\ \beta_i(q)}{\sum_{r \in Q} \exp[-\propto \Phi_i(r)]\ \beta_i(r)} \qquad (21)$$

where $\beta_i(q) = \Pi_{j \in Ni}\,(\sum_{s \in Q} P_j^{(t)}(s)\, \exp[-V_c(q,s)]\,)$.

Note that in either case, in this deterministic scheme, the value of the

field is never computed explicitly. The algorithm corresponds to an M-layer cooperative network P where the state of each cell (i,q) corresponds to the current estimate of the marginal probability $P_i(q)$. At the fixed points, the system will satisfy the mean field equation:

$$P_i(q) = \overline{w}_i(q) \qquad (22)$$

## 4.1. Stability.

Heuristically, the stability of the system (19) (using either (20) or (21) to approximate $\overline{w}$), follows from the regularity of the underlying Markov chain. This consideration is reinforced by the experimental behavior of the algorithm, which, in fact, is found to converge to a solution of the form (22). More formally, one can guarantee the stability of the system by introducing a time-varying "temperature" parameter T in the approximation (20) to $\overline{w}_i$ , which now becomes:

$$\overline{w}_i^{(t)}(q) = \frac{\exp\left[-\overline{u}_i^{(t)}(q) / T(t)\right]}{\sum_{r \in Q} \exp\left[-\overline{u}_i^{(t)}(r) / T(t)\right]} \qquad (23)$$

Where the function T(t) (the "annealing schedule") is a non-negative, decreasing function satisfying T(0) = 1. (we have used the schedule:

$$T(t) = \begin{cases} 1, & \text{for } t < t_0 \\ \exp\left[-(t-t_0)^2/\tau^2\right], & \text{for } t > t_0 \end{cases}$$

with good experimental results. Typical values for the parameters are : $t_0 = 0$ and $\tau = 29$ ).

In the appendix, we show that if $T_f = 0$ , the deterministic system given by (19) and (20), for a first order, M-ary field will always be stable, and converge to a (local) maximum of the posterior probability $P_{f|g}$ .

We have found, experimentally, that the convergence of the system described by (19) and (23) (with the annealing schedule given above) is very fast: one can get reasonably good results in less than 10 iterations. The convergence of the system given by (19) and (21) is slightly slower: it might take about 20 iterations instead of 10 (one cannot use "annealing" in this case), but the final results are better than in the former case. The results of a typical experiment are illustrated in figure 3, where we estimate a non-isotropic, ternary field corrupted by additive Gaussian noise (the values of the original gray levels are assumed to be known). The estimate obtained after 500 iterations of the (asymptotically optimal) Monte Carlo scheme described in section 3.2 is also presented, for comparison.

---

**Figure 3 around here**

---

### 4.2. Analog Networks.

Let $h$ denote the time increment for system (19). Set $\lambda = 1 - h/\tau$, where $\tau$ is a parameter. Taking the limit as $h \to 0$, one can construct a continuous-time system equivalent ( in the sense of having the same set of fixed points) to the discrete-time system (19) :

$$\frac{d\, p_i(q)}{d\, t} = -(1/\tau)\,(p_i(q) - \overline{w}_i(q)) \quad , \quad i \in L \quad , \quad q \in Q \qquad (24)$$

This system can be implemented using an analog network with non-linear amplifiers whose gains are given by the function $\overline{w}$. For example, for binary fields, and using equation (23) to approximate $\overline{w}$, we get a single-layer network, with

$$\overline{w}_i(t) = 1/(1 + \exp[\Delta u/T])$$

and $$\Delta u = \bar{u}_i(1) - \bar{u}_i(0) \qquad (25)$$

This network is related to the ones proposed by Hopfield (1985) (see also Koch et.al., 1986) for the solution of combinatorial optimization problems (see note [4] ). It is interesting to note that, although in the zero temperature limit both types of network will have fixed points at the local minima of the posterior energy $U_p$ , their dynamic behavior will be very different, so that even for the same initial state (for example, $P_i = 0.5$ , for all i), they will, in general, converge to different solutions. We have performed simulations of both systems, and have found a much better performance for the scheme described by (24).

### 4.3. Higher Order Fields.

It is possible to construct deterministic cooperative networks to compute estimates of Markovian fields that have non-zero potentials for cliques with more than 2 elements, provided that the appropriate formulae for the average excitations are used. As an example, consider a binary, second order MRF whose only non-zero potentials are given by:

$$V_3(f_i,f_j,f_k) = \begin{cases} \varphi(f_i + f_j + f_k) & \text{if} \quad j,k \in N^3_i \\ 0 \ , & \text{otherwise} \end{cases}$$

where $N^3_i = \{ j, k : \|i-j\| = \|i-k\| = 1 ; \|j-k\| = \sqrt{2} \}$ (see figure 4); $f_i \in \{0,1\}$, for all i, and $\varphi$ is some real-valued function.

---

**Figure 4 around here**

---

If one wishes to use the approximation given by equation (23), the average excitations should be computed using:

$$\overline{u}_i(1) = \sum_{j,k \in N^3_i} [\, \varphi(1)\, P_j(0)P_k(0) + \varphi(2)\,(\, P_j(1)P_k(0) + P_j(0)P_k(1)) + \varphi(3)\, P_j(1)P_k(1)\, ] + \alpha\, \Phi_i(1)$$

$$\overline{u}_i(0) = \sum_{j,k \in N^3_i} [\, \varphi(0)\, P_j(0)P_k(0) + \varphi(1)\,(\, P_j(0)P_k(1) + P_j(1)P_k(0)\,) + \varphi(2)\, P_j(1)P_k(1)\, ] + \alpha\, \Phi_i(0)$$

We will present a practical application for these higher order fields when we discuss the reconstruction of piecewise continuous functions in section 7.

### 4.4. "Winner-Takes-All" (WTA) Networks.

In the zero temperature limit, and for the particular case where $\lambda = 0$, the deterministic system defined by (19) and (23) works, in fact, as a WTA network, in the sense that only the maximally stimulated elements in each column will have a non-zero updated value.

This scheme can be implemented in a binary network , if a "restoring" mechanism is introduced, so that the updating scheme becomes:

$$P_i^{(t+i)}(q) = \begin{cases} 1\,, \text{ if } \overline{w}_i^{(t)}(q) > 1/2M \text{ (or equivalently, if } \overline{u}_i(q) > \overline{u}_i(r)\,,\ r \neq q\,) \\ 0\,, \text{ otherwise.} \end{cases} \qquad (26)$$

Networks of this kind are very efficient in the use of memory, and can be useful in the solution of certain classes of problems (such as the computation of depth from stereoscopic pairs of images, a problem that we will discuss in more detail below), where their convergence to "good" solutions can be guaranteed.

## 5. Parameter Estimation.

The algorithms we have described give excellent results when the parameters that characterize the system (the "field" and "noise" parameters) are known. In general, however, this is not the case, and these parameters have to be estimated directly from the (noisy) observations. This problem, whose solution remains very much open, is not exclusive of this formulation. In all other approaches to the reconstruction problem (such as the "regularization" one), there are also "free" parameters that have to be adjusted (usually, using a trial and error procedure).

In our case, since we are using a probabilistic formulation, we can, at least in principle, define the "best" value for these parameters in a precise way: it is the value that maximizes the likelihood function:

$$L(\theta \mid g) = \log P(g \mid \theta) = \log \frac{\sum_f \exp\left[ - U_p(f;g \mid \theta) \right]}{\sum_{f,h} \exp\left[ - U_p(f;h \mid \theta) \right]}$$

where $\theta$ is the parameter vector, and $U_p$ is the posterior energy.

Unfortunately, the extraordinary complexity of this function makes its direct maximization unfeasible. There are indirect methods, such as the "EM algorithm" (Dempster et. al., 1977; see note [5]), but they are practical only if the number of unknown parameters is small (one or two).

We have developed a different approach, which is computationally more efficient, and may,in principle, be used for more that two unknown parameters. The basic idea is to use some statistics computed from the data to constrain the space of plausible values for the estimates to a smooth curve . In this way, we can perform an exhaustive search for the global minimum of an appropriate merit function by varying continuously the values of the parameters, so that the equilibrium of the Gibbs chain is mantained.

To illustrate this idea, we consider, for example, the case of a binary

Ising field where the noise corruption corresponds to a BSC (the idea can be easily extended to M-ary Ising fields and other noise models - see note [6]).

We define the statistic $U_g$ as:

$$U_g = \sum_{i,j} V(g_i , g_j)$$

where V is an Ising potential (see section 2). If the error rate of the channel is $\epsilon$, we have that

$$E [ U_g \mid \alpha, T_0 ] = E [ U_0 \mid \alpha, T_0 ] (4\epsilon^2 - 4\epsilon + 1)$$

Note that the function

$$E [ U_0 \mid \alpha, T_0 ] = E [ U_0 \mid T_0 ] = \Psi (T_0)$$

is independent of the data, and thus, it can be pre-computed for any given lattice size, using, for example, the Monte Carlo procedure of section 3, but this time with the prior energy $U_0$ instead of $U_p$ (in figure 5 we show this function for a 30 × 30 binary Ising lattice with free boundaries).

---

**Figure 5 around here**

---

Therefore, we get:

$$E [U_g \mid \alpha, T_0 ] = \Psi (T_0) (4\epsilon^2 - 4\epsilon + 1)$$

If we make the assumption that

$$E [ U_g \mid \alpha, T_0 ] = \overline{U}_g$$

where $\overline{U}_g$ is the observed statistic (see note [7]), we can constrain the search for the estimates $\hat{\alpha}$, $\hat{T}_0$ to the curve given by the equations:

$$\hat{\epsilon} = {}^1/_2 [ 1 - ( \overline{U}_g / \Psi(\hat{T}_0) )^{1/2} ]$$

$$\hat{\alpha} = \ln [ (1 - \hat{\epsilon}) / \hat{\epsilon} ] \qquad (27)$$

As a merit function, we have used the squared difference between the conditional and unconditional expected values of the sufficient statistics (see notes [2] and [5]):

$$\begin{aligned} \mathcal{L}(\hat{\alpha},\hat{T}_0) &= (E[\alpha \mid \mathcal{J},\hat{\alpha},\hat{T}_0] - E[\alpha \mid \hat{\alpha},\hat{T}_0])^2 + \\ &+ (E[U_0 \mid \mathcal{J},\hat{\alpha},\hat{T}_0] - E[U_0 \mid \hat{\alpha},\hat{T}_0])^2 = \\ &= (\overline{\alpha_0} - \hat{\alpha})^2 + (\overline{U_0} - \Psi(\hat{T}_0))^2 \qquad (28) \end{aligned}$$

with $\overline{\alpha_0} = \ln [ (1 - \overline{\epsilon_0}) / \overline{\epsilon_0} ]$ .

$\overline{\epsilon_0}$ and $\overline{U_0}$ are the conditional expected values of the noise density and the field potentials, respectively, and can be approximated either as time averages of the corresponding Gibbs chain (using the posterior distribution given by equation (10), and with $\hat{\alpha}$ and $\hat{T}_0$ as parameters), or from the stationary (posterior) marginal probabilities, if the deterministic system is used (see below). The optimal estimate for ($\alpha$, $T_0$) can now be obtained as the global minimizer of $\mathcal{L}$ over the curve (27).

Other merit functions may also be used (see section 6.3) ; this one, however, has the advantage of having a precise interpretation: it corresponds to the norm of the derivative of the true likelihood function (see note [5]), so that it will be zero at its local maxima. This also suggests the use of higher order cumulants (for example, the covariance matrix) to approximate the corresponding higher order derivatives, and improve the optimization strategy, but we will not pursue this point here. Note that if $T_0$ (and hence $\alpha$) are varied slowly enough, so that the associated Gibbs chain is mantained approximately in equilibrium, the computational cost of this search will be equivalent to that of a single "simulated annealing" experiment.

This estimation algorithm allows us to reconstruct a pattern $\mathcal{I}$ from the noisy observations $\mathcal{J}$ without having to adjust any free parameters. The only prior assumptions correspond to the qualitative structure of the

field $f$ (first order, isotropic MRF) and to the nature of the noise process. In practice, this means that we can apply it to restore any piecewise uniform image with uniform granularity, even if it has not been generated by a Markov process. We have used this algorithm to reconstruct a variety of binary images with excellent results. In figure 6 we show such a restoration. The observations (b) were generated from the synthetic image (a) with an actual error rate of 0.30 (assumed unknown). The optimal estimate for $f$ (using the deterministic system given by equations (19) and (21) ) is shown in (c). The behavior of the function $L$ along the curve (27) is shown in (d).

---

**Figure 6 around here**

---

It should be noted that all the average quantities required for this parameter estimation procedure , can be approximated using formulae of the form:

$$\overline{a_i} = \sum_{q \epsilon Q} P_i(q)\, a_i(q)$$

where $\overline{a_i}$ is the desired average; $a_i(q)$ is the value of the variable $a_i$ obtained assuming that $f_i = q$ , and $P_i(q)$ is the marginal probability, estimated using either the Monte Carlo or the deterministic procedures described in section 3 and 4. Thus, for example, the expected value of the noise density over the lattice, $\overline{\epsilon_0}$ , can be estimated as:

$$\overline{\epsilon_0} = 1/N \sum_{i \epsilon L} \sum_{q \epsilon Q} P_i\,(q)\,(1 - \delta(q-g_i))$$

where N is the number of sites of the lattice L.

## 6. Examples.

In this section, we present some examples of the application of the methods that we have presented, to some problems which are relevant in computer vision, In particular, we will discuss: the reconstruction of piecewise smooth surfaces from sparse observations; the formation of

perceptual clusters, and the reconstruction of depth from stereoscopic pairs of images.

## 6.1. Reconstruction of Piecewise Smooth Surfaces.

In several problems relevant to computer vision - for example, in the reconstruction of depth from stereoscopic pairs of images- one can frequently estimate the desired property (say, depth), only at a set of sparse locations in the image. With these data, one then wishes to reconstruct the surfaces of the corresponding objects (which one assumes to be piecewise smooth), but preserving the discontinuities that correspond to the boundaries between them.

To apply the general reconstruction algorithms developed above to this problem, the main issue is the representation of the concept of "piecewise continuity" in the form of a prior Gibbs distribution in a meaningful way.

A flexible construction involves the use of two coupled MRF models: one to represent the function (the surface) itself, and another to model the curves where the field is discontinuous. This last field lives in a lattice, whose sites correspond to links between pairs of adjacent sites of the "surface" field (a coupled model of this kind was first used by Geman ans Geman (1984) in the context of the restoration of piecewise constant images).

In our case, the potentials that model the coupling between the two fields take the form:

$$V(f_i,f_j,b_{ij}) = \begin{cases} (f_i-f_j)^2\,(1-b_{ij}) & , \quad \text{for } \|i-j\| = 1 \\ 0 & , \quad \text{otherwise.} \end{cases}$$

Where $b_{ij}$ denotes the "line element" that lies between sites i and j, and is equal to one, if that line element is "on", and equal to zero, otherwise.

The shape of the lines is controlled using the value of potential functions , that are associated with the configurations of different "line cliques". We have used, for example, the cliques (a) and (b) of figure 7.

---

**Figure 7 around here**

---

The corresponding potentials, $V_a$ , $V_b$ ,encourage the formation of thin lines with smooth curvature; their values depend only on the number of active lines in the clique, according with the table:

| Number of active lines: | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $V_a$ : | 0 | 0.4 | .25 | 1.2 | 2.0 |
| $V_b$ : | 0 | 0.0 | 10.0 | - | - |

Assuming that the observations are corrupted by i.i.d. Gaussian noise, we get the following expression for the posterior energy:

$$U_p(f, b; g) = 1/T_0 \sum_{i,j} (f_i - f_j)^2 (1 - b_{ij}) + 1/(2\sigma^2) \sum_{i \in S} (f_i - g_i)^2 + \\ + \sum_{Ca} V_a(b) + \sum_{Cb} V_b(b)$$

where S is the set of sites where an observation is present. As a performance criterion, we use a mixed cost functional of the form:

$$C(f, b, \hat{f}, \hat{b}) = \sum_i (f_i - \hat{f}_i)^2 + \sum_{ij} (1 - \delta(b_{ij} - \hat{b}_{ij}))$$

where the sums range over the sites of the "surface" and "line" lattices, respectively, and $\delta(\cdot)$ is defined in (6). This error criterion means that the reconstructed surface should be as close as possible to the true (unknown) surface, and that we should commit as few errors as possible in the assertions about the presence and absence of discontinuities. Applying the results of section 3, we find that the optimal estimators will be: the posterior mean for $f$, and the maximizer of the posterior marginals for $b$.

The computation of these estimates poses some special problems, due to the fact that the variables of the $f$ field are continuous-valued (or can take a large number of discrete values). The details of the algorithms used to approximate them are given in (Marroquin et.al., 1986); here, we only present, in figure 8, an illustration of their performance.

---

**Figure 8 around here**

---

## 6.2. Reconstruction of Depth from Stereoscopic Pairs.

The reconstruction of depth from stereoscopic pairs of images of natural scenes is a difficult problem, whose solution (i.e., the construction of an algorithm whose performace matches that of human beings) is still open. One of its main parts (although not the only one) is the problem of matching "tokens" that occur in both images along epipolar lines (see for example, Poggio, 1984; Marr and Poggio, 1976). To illustrate the potential usefulness of the techniques that we have presented here for the solution of this problem, we consider now a simple version of it: the matching of "Random Dot Stereograms" (Julesz, 1960).

We will consider binary images, and assume that each row of the right image is obtained as a sample function of a Bernoulli process of density $p$. The left image is formed from the right one by shifting it along the x direction, by a variable amount given by the disparity function $d$, except at some points, where an error is commited with probability $\epsilon$. Note that some regions that appear in the right image will be occluded in the left one. The "occlusion indicator", $\varphi_d$, can be computed deterministically from $d$ in the following way:

$$\varphi_d(i) = \begin{cases} 1, & \text{if } d_{i-k} > d_i + k, \text{ for some integer } k \in (0,m] \\ 0, & \text{otherwise} \end{cases} \qquad (29)$$

The occluded areas are assumed to be "filled in" by an independent

Bernoulli process B. The final model is then:

$$g_L(i) = \begin{cases} g_R(i+d_i) & \text{, with prob. } 1-\epsilon, \text{ if } \psi_d(i) = 0 \\ 1 - g_R(i+d_i) & \text{, with prob. } \epsilon \text{ , if } \psi_d(i) = 0 \\ B(i) & \text{, with prob. } 1 \text{ , if } \psi_d(i) = 1. \end{cases} \qquad (30)$$

Note that in the two-dimensional case, the index i denotes a site of the lattice, and therefore, it can be represented as a two-vector $(i_1,i_2)$ whose components denote the column and row of the site, respectively. To simplify the notation, we will adopt the following convention throughout this section: when a scalar is added to this vector index (as in $g_R(i+d_i)$ and $d_{i+k}$), it will be implicitly assumed that it is multiplied by the vector (1,0) (so that the above expressions should be understood as $g_R(i+(d_i,0))$ and $d_{i+(k,0)}$ , respectively). Using this convention, the observation model of equation (30) can be applied either to the one or to the two-dimensional cases.

Notice that even if the observations are noise-free ($\epsilon$=0), the solution of the problem remains ambiguous, and it cannot be uniquely determined, unless some prior knowledge about $d$ (for example, in the form of a MRF model) is introduced. The use of a MRF model in this case, corresponds to a quantification of the assumption of the existence of "dense solutions" (this term was introduced by Julesz (1960), and essentially corresponds to the assumption that the disparity varies smoothly in most parts of the image; see also Marr and Poggio (1976)). The use of the occlusion indicator, corresponds to the "ordering constraint" (i.e., the requirement that, if i > j, then, $i + d_i > j + d_j$ ; see Baker (1981). We put $\psi_d(i) = 1$ whenever this constraint is violated).

To formulate the estimation problem, we consider the sequence $g_L$ as "observations", while $g_R$ will play the role of a set of parameters. Thus, from (30), we have (assuming for simplicity that $p = 0.5$):

$$P(g_L(i)=k \mid \mathcal{d}, g_R) = P_{g|d}(k) = \begin{cases} 1-\epsilon, & \text{if } \psi_d(i)=0 \text{ and } g_R(i+d_i) = k \\ \epsilon, & \text{if } \psi_d(i)=0 \text{ and } g_R(i+d_i) \neq k \\ 0.5, & \text{if } \psi_d(i)=1 \end{cases}$$

As a prior model for the disparity field, we may use a first order MRF with generalized Ising potentials, such as the one presented in section 2. Other models may also be used, including the coupled depth and line fields that we discussed in the previous section. For the present, let us assume that the simple Ising model is adequate. Note that even when the matching problem is one-dimensional, (we are assuming that there is no vertical disparity between the images, so that the matching can be done on a row-by-row basis), the two-dimensional nature of the prior MRF model for the disparity introduces a coupling between matches at adjacent rows. The posterior energy is:

$$U_p(\mathcal{d};g) = (1/T_0)\sum_{i,j} V(d_i,d_j) + {}^1/_2\sum_i \psi_d(i) \ln 2 + \\ + (\alpha/2)\sum_i (1-\psi_d(i))(\delta(g_L(i)-g_R(i+d_i))$$

where $\alpha = \ln [ (1-\epsilon)/\epsilon ]$.

It is possible to apply the general Monte Carlo or deterministic algorithms presented above to approximate the optimal estimator for $\mathcal{d}$, with respect to a given performance measure (such as the mean squared error). Their use in this case, however, is complicated by the introduction of the occlusion function $\psi$ in the posterior energy: the size of the support for this function equals the total number of allowed values for the disparity (see equation 27). If this number is large, the computation of the increment in energy, or of the expected value of the conditional distributions (if the deterministic scheme is used) may be quite expensive. In many cases, however, the size of the regions of constant disparity is relatively large compared with the size of the occluded areas. In these cases, one can approximate the posterior energy by:

$$U_p(\mathcal{d};g) = (1/T_0)\sum_{i,j} V(d_i,d_j) + (\alpha/2)\sum_i (1-\psi_d(i))(\delta(g_L(i)-g_R(i+d_i)) \quad (31)$$

and increase significantly the computational efficiency.

The form of this expression (in particular, the non-monotonicity of the noise statistic) causes the solution to this problem to remain ambiguous, even if the signal to noise ratio is arbitrarily high. In this case, however, it is possible to use the efficient deterministic scheme discussed in section 4.4 - namely, a WTA network. Moreover, for perfect observations (zero noise), it is possible to guarantee the convergence of this scheme to the correct solution in a small number of iterations. The informal argument is as follows (the technical details may be found in Marroquin, 1985):

Given, as initial state of the network P, $P_i^{(0)}(q) = 0$, for all i and q, at the first iteration of the algorithm (26), the network will turn "on" the cells (i,d) for which $\delta(g_L(i)-g_R(i+d_i)) = 0$, that is, all the cells in the correct places (since there is no noise), and some cells in the wrong layers as well, due to accidental correlations in the texture of both images.

After the first iteration, the cells that are "on" in the correct places, will have at least as many neighbors that are "on" as the corresponding cells in the wrong layers, so that the algorithm will only turn "off" some of the latter. This will cause the cells that lie at the boundaries of clusters in the wrong layers to lose, in the subsequent iterations, against the corresponding strongly stimulated cells that lie in the interior of the "correct" regions. This will result in a progressive shrinking of the wrong clusters, and will end up with their dissapearance.

The only situation in which this behavior will not take place, is when there is a significant overlap between wrong clusters and the boundaries of correct regions. In this case, the algorithm will not be able to solve correctly this ambiguity based only on smoothness considerations (i.e., on the prior MRF model), and it will locate the boundary at a position, within the region of overlap, which will depend on the detailed shape of this region. Also, the solution may not be so clean in this case; a few cells, corresponding to different disparities at the same spatial location, may be left "on" in the final state (limit cycles involving some of these few cells are also possible). It should be noted that the human visual

system may exhibit a similar behavior in these cases.

This type of ambiguity (accidental overlap) is relatively frequent in sparse stereograms. However, the regions of overlap are typically "blank" regions (i.e., without tokens), and the algorithm will give the correct disparity at all token locations.

Figure 9 illustrates the performance of algorithm (26) with sparse and dense random dot stereograms portraying a "wedding cake". As predicted by the theory, the convergence to the correct solution is fast (less than four iterations) in both cases. In the case of the sparse stereogram, the boundaries are slightly misplaced, but, as can be verified by direct inspection, all the dots are correctly located.

---

**Figure 9 around here**

---

To apply this algorithm to the processing of real images, there are some modifications and extensions that should be made. They fall in two categories:

**Neighborhood size:** It is possible to increase the robustness of the algorithm, with respect to the presence of noise in the images, by increasing the size of the excitatory neighborhood (i.e., by postulating a more global MRF prior model) and decreasing the value of the parameter $\alpha$. This increased robustness is traded off by a decrease in resolution: small, correct regions may be treated as "noise", and therefore disappear from the solution. Also, the shape of the piecewise constant regions may be altered (corners may be rounded off, and small concavities "filled in").

**Token Selection:** In the case of continuous-toned images of natural objects, the distribution of the reflected light may vary as the viewpoint is changed (particularly the specular component); also, the two retinas (cameras) may have different point spread functions, and be affected by independent sources of noise. This means that the simple model for the observation process given above should be replaced by another that reflects the formation of natural images in a more realistic way. The use of a better model will cause the term $\delta(g_L(i)-g_R(i+d_i))$ in equation (31) to be replaced by a different compatibility measure $\eta_{i,d}$, which

may be obtained by first preprocessing the right and left images with an operator T whose output should be, ideally, invariant under the changes in viewpoint, optics, etc., and then computing a suitably defined distance D between the two images:

$$\eta_{i,d} = D( Tg_L(i) , Tg_R(i+d_i) )$$

The WTA algorithm (26) can still be used, if we now compute $\overline{u}_i(d)$ using:

$$u_i(d) = \sum_{j \in Ni} c( \|i-j\| ) P_i(d) + \alpha \eta_{i,d}$$

where Ni is the extended neighborhood of i, and c(·) denotes a set of parameters that depend only on the distance $\|i-j\|$ .

We have performed some experiments using this kind of mechanism, and the preliminary results are encouraging. Other researchers (Prazdny, 1985; Pollard et. al., 1985; Drumheller and Poggio, 1986) have also reported good results with algorithms of a similar form (although not derived from probabilistic considerations).

## 6.3. Formation of Perceptual Clusters.

At the heart of a general purpose perceptual system, one must have a mechanism for deciding which parts of an image should be considered to "belong" together. A simple instance of this problem is the grouping of dots in an image into perceptual clusters. Some heuristic schemes have been proposed to model this phenomenon (see, for example, O'Callahan, 1974). We will now show how this problem can be formulated in an elegant way, that is also biologically motivated, as a particular case of the reconstruction of binary patterns from noisy observations.

The conceptual model is as follows:

let us consider the dots that form the original pattern as patches belonging to some objects of uniform color that are partially hidden, say, by some foliage. In this way, the formation of clusters is equivalent to

the problem of reconstructing these objects (whose cohesive nature is modeled by a first order MRF with Ising potentials) from observations that are formed according with the following model:

Suppose that $f_i = 1$ only if the image of an object overlaps the $i^{th}$ site of the lattice. We assume that the "foliage" will hide this point (i.e., make $g_i = 0$) with probability $\epsilon$, and that spurious values of $g_i = 1$ can appear in sites where $f_i = 0$ with a very small probability $\rho$:

$$g_i = \begin{cases} 1, \text{ with prob. } (1-\epsilon), \text{ if } f_i = 1 \\ 0, \text{ with prob. } \epsilon \quad , \text{ if } f_i = 1 \\ 0, \text{ with prob. } (1-\rho), \text{ if } f_i = 0 \\ 1, \text{ with prob. } \rho \quad , \text{ if } f_i = 0. \end{cases}$$

with $\rho << 1$. The posterior energy is:

$$U_p(f;g) = (1/T_0)\sum_{i,j} V_c(f_i,f_j) + \alpha\sum_{i\epsilon S} (1-\delta(1-g_i)) + M\sum_{i\epsilon S} (1-\delta(g_i))$$

where: $V_c$ is given by equation (8); $\alpha = \ln[\,(1-\epsilon)/\epsilon\,]$; S is the set of sites i where $f_i = 1$; $\delta(\cdot)$ is defined in (6), and $M = \ln[\,(1-\rho)/\rho\,]$ is a very large number.

The clustering task is now equivalent to the problem of estimating $f$ and the parameters $\alpha$ and $T_0$ from the noisy observations $g$. To accomplish this, we minimize, over the appropriate region of the parameter space, a merit function, which is related to the degree of uniformity in the spatial distribution of the corresponding residuals. We have defined, for example, a likelihood function L, by covering the lattice with a set of m non-overlapping squares (say, 8 pixels wide); computing the relative variance of the noise parameter, estimated over each square, and adding all these terms together:

$$L(\hat{\alpha},\hat{T}_0) = -\sum_{j=1}^{m} \left[ \frac{\overline{\epsilon_j} - \overline{\epsilon_0}}{\overline{\epsilon_0}} \right]^2$$

where $\overline{\epsilon_0}$ and $\overline{\epsilon_j}$ are the (conditional) expected values of the noise density over the intersection of the set $A = \{i : \hat{f}_i = 1\}$, with the whole lattice, and with the $j^{th}$ square, respectively (thus, for example, $\overline{\epsilon_0} = (1/|A|)\sum_{i\epsilon A} \delta(g_i)$, where $|A|$ is the size of set A, for a particular value of $\hat{\alpha}$ and $\hat{T}_0$). The performance of this procedure is illustrated in figure 10, where we show: the original dot pattern (upper left) and the reconstructed objects for decreasing values of $\hat{\sigma}=\hat{\alpha}\hat{T}_0$ (we have found that for this task, a fast, deterministic approximation to the optimal estimator, which depends on the parameters $\alpha$ and $T_0$ only through their product, gives good enough results; the technical details may be found in Marroquin, 1985). The maximizer of the likelihood L is marked with an arrow. We beleive that these preliminary results are encouraging, although, clearly, more numerical and psychophysical experiments are needed to assess the plausibility of this scheme to model human perceptual processes.

---

**Figure 10 around here**

---

## 7. Discussion.

In this chapter we have presented a probabilistic approach to the solution of perceptual problems. We showed that a large class of these problems can be reduced to the reconstruction of a function on the sites of a finite lattice, from a set of degraded observations; we pointed out that, in order to solve them, one has to include prior, generic knowledge about the behavior of the desired solution; then, we presented a general class of probabilistic models that permit the inclusion of this knowledge, and derived the Bayesian estimators that provide an optimal solution. The distributed (deterministic and stochastic) algorithms that we have presented for approximating these estimators, can be efficiently implemented, either in a general purpose serial computer, or in special hardware: analog and hybrid computers, and massively-parallel machines

(see Marroquin et. al., 1986; Koch et.al., 1986; Drumheller and Poggio, 1986). The use of these special purpose architectures will make these algorithms practical, even for real-time applications.

We illustrated the practical value of this approach with several examples: the reconstruction of piecewise smooth surfaces from sparse data; the reconstruction of depth from stereoscopic measurements, and the formation of perceptual clusters.

There are a number of perceptual problems which are related in such a way, that the solutions that can be obtained, should improve if the mutual constraints between them were taken into account. Thus, the presence of a brightness edge should increase the likelihood of a depth edge, and viceversa; the depth estimated from stereo should be compatible with the shape derived from shading, etc. We beleive that the probabilistic approach that we have presented here, can provide a framework for the integration of the solutions to these problems (via the use of coupled potential functions for the corresponding MRF models) into a unified cooperative process with enhanced performance.

## Appendix: Asymptotic Convergence of the "Deterministic Annealing" Scheme for M-ary Ising Fields.

In this appendix, we analyze the convergence of the deterministic system defined by equations (19) and (22), in the limit when the "annealing temperature" $T \to 0$, for the particular case when the field $f$ is an M-ary field with generalized Ising potentials, and the noise is continuous-valued. The posterior energy is given by:

$$U_p = \sum_i \{ a_x \sum_{j \in Nx(i)} V(f_i, f_j) + a_y \sum_{j \in Ny(i)} V(f_i, f_j) + \alpha \Phi_i(f_i) \} \qquad (A1)$$

where V is a generalized Ising potential, defined by equation (8); $a_x$ and $a_y$ are the interaction strengths, and Nx(i), Ny(i) are the set of nearest neighbors to i in the x and y direction, respectively. The functions $\Phi_i$, and the parameter $\alpha$ depend on the noise distribution.

$\overline{w}_i$ will be approximated by equation (22), with the average local excitations given by:

$$\overline{u}_i^{(t)}(q) = a_x \sum_{j \in Nx(i)} (1-2P_j^{(t)}(q)) + a_y \sum_{j \in Ny(i)} (1-2P_j^{(t)}(q)) + \alpha\Phi_i(q) =$$

$$= -2a_x \sum_{j \in Nx(i)} P_j^{(t)}(q) - 2a_y \sum_{j \in Ny(i)} P_j^{(t)}(q) + \eta_i(q)$$

$$\text{with } \eta_i(q) = a_x |Nx(i)| + a_y |Ny(i)| + \alpha\Phi_i(q) \qquad \text{(A2)}$$

At T=0, we will have that:

$$\overline{w}_i^{(t)}(q) = \begin{cases} 1 \text{ , if } \overline{u}_i^{(t)}(q) < \overline{u}_i^{(t)}(r) \text{ , for } r \neq q \\ 0 \text{ , otherwise} \end{cases} \qquad \text{(A3)}$$

(Note that for continuous-valued noise, such as in the i.i.d. Gaussian case, we do not have to consider the case $u_i^{(t)}(q) = \overline{u}_i^{(t)}(r)$ , for $q \neq r$, since it will occur only with probability 0).

We will consider an asynchronous version of the algorithm, so that only one site (one column of the P network) is updated at a time. Suppose that we update site i at time t, and let m be such that $\overline{u}_i^{(t)}(m) < \overline{u}_i^{(t)}(q)$, for all $q \neq m$. Then, at T=0 we will have, using (19):

$$P_i^{(t+1)}(m) = \lambda P_i^{(t)}(m) + (1-\lambda)$$
$$P_i^{(t+1)}(q) = \lambda P_i^{(t)}(q) \quad \text{, for } q \neq m.$$

$$\text{Let} \quad L(t) = \sum_q \{ -a_x \sum_{i,j \in Nx(i)} P_i^{(t)}(q) P_j^{(t)}(q) -$$
$$- a_y \sum_{i,j \in Ny(i)} P_i^{(t)}(q) P_j^{(t)}(q) + \sum_i P_i^{(t)}(q) \eta_i(q) \}$$

which is bounded, since $P_i^{(t)}(q) \in [0,1]$ , for all i,q. We will now show that L decreases at every iteration of the system at T=0, except at a fixed point:

$$L(t+1) - L(t) = \Delta L = \sum_q (P_i^{(t+1)}(q) - P_i^{(t)}(q)) \, \overline{u}_i^{(t)}(q)$$

but $$\sum_q P_i^{(t+1)}(q) \; \overline{u}_i^{(t)}(q) = \lambda \, P_i^{(t)}(m) \, \overline{u}_i^{(t)}(m) + (1-\lambda) \, \overline{u}_i^{(t)}(m) + + \sum_{q \neq m} \lambda \, P_i^{(t)}(q) \, \overline{u}_i^{(t)}(q) =$$

$$= (1-\lambda) \, \overline{u}_i^{(t)}(m) + \lambda \sum_q P_i^{(t)}(q) \, \overline{u}_i^{(t)}(q)$$

so that $\Delta L = (1-\lambda) \left( \overline{u}_i^{(t)}(m) - \sum_q P_i^{(t)}(q) \, \overline{u}_i^{(t)}(q) \right) < 0$

with equality only if $P_i^{(t+1)}(q) = P_i^{(t)}(q)$, for all q, which implies that

$$P_i^{(t)}(m) = 1 \text{ , and } P_i^{(t)}(q) = 0 \text{ , for } q \neq m.$$

It is not difficult to see that if $L(t+N) = L(t) = L^*$, where N is the number of asynchronous iterations needed to update all the sites of the lattice, the system will be at a stable fixed point $P^*$. If we now make $f_i^* = m$, if $P_i^*(m) = 1$, for all i, we will have that $L^* = U_p(f^*)$, so that, at $T = 0$, the deterministic system will always converge to a fixed point, which will correspond to a local minimum of the posterior energy.

Notes:

[1] The requirement that the algorithms that perform the reconstruction should be distributed, i.e., implementable in some kind of "cooperative network", is justified, both from a theoretical viewpoint (so that it represents a plausible biological mechanism, according with our current knowledge of neurophysiology and psychophysics), and form a practical one: I beleive that artificial systems with real-time perceptual abilities will only be possible with the use of algorithms that are implemented in fine grain, distributed multiprocessing architectures.

[2] In many practical cases, the Standard Gibbs form belongs to a more general class of distributions, which is called the "Regular Exponential

Family", whose form is:

$$P(x) = (b(x)/a(\varphi)) \exp [ -t^T \varphi ]$$

where t is the vector of sufficient statistics ; T denotes transpose, and $\varphi$ is the parameter vector (in equation (10), for example, $t = (U_0, \bar{\varepsilon})^T$ , and $\varphi = (1/T_0, \alpha)^T$ . This definition will be important when we talk about parameter estimation.

[3] As a simple example for which the regularity of the Metropolis chain is destroyed, consider a 3×3 binary Ising lattice with periodic boundary conditions. It is easy to see that for the initial state:

1 0 1
0 1 0
1 0 1

the Metropolis algorithm, either with lexicographic updating order, or with simultaneous updating of all non-neighboring sites, will produce, deterministically, the sequence:

$$\begin{matrix} 1\,0\,1 \\ 0\,1\,0 \\ 1\,0\,1 \end{matrix} \rightarrow \begin{matrix} 0\,1\,0 \\ 1\,0\,1 \\ 0\,1\,0 \end{matrix} \rightarrow \begin{matrix} 1\,0\,1 \\ 0\,1\,0 \\ 1\,0\,1 \end{matrix} \rightarrow \ldots$$

for any finite temperature.

[4] The "Hopfield network" in this case would be described by:

$$\frac{d z_i}{d t} = - \Delta u - z_i / \tau \quad ; \quad p_i = 1 / (1 + \exp [-z_i/T])$$

[5] Although it is computationally unfeasible to perform the maximization of the likelihood function L directly, due to the extraordinary complexity of
$P( g \mid \theta )$, the form of the "complete data" distribution $P( f, g \mid \theta)$ (the so called "regular exponential family form" ; see note [2]) is such that the derivatives of the likelihood function L will be given by:

$$\frac{\partial L}{\partial \theta_1} = E [ U_0 \mid g, \theta ] - E [ U_0 \mid \theta ]$$

$$\frac{\partial L}{\partial \theta_2} = E [ \Phi \mid g, \theta] - E [ \overline{\epsilon} \mid \theta]$$

where $\Phi$ is the noise statistic (see equation (2)); $U_0$ is the prior energy (equation (7)); $\theta_1$ is the field parameter ($1/T_0$), and $\theta_2$ is the noise parameter ($\alpha$) (see Dempster et.al., 1976). This means that at a local maximum of L we will have that:

$$E [ U_0 \mid g, \theta ] = E [ U_0 \mid \theta ]$$

$$E [ \Phi \mid g, \theta] = E [ \overline{\epsilon} \mid \theta]$$

Note that both the left and the right hand sides of the above equations can be approximated using the Monte Carlo procedure described in section 4 (using the posterior and prior energy, respectively), and that the right hand side is independent of the observations.

These relations form the basis of the EM algorithm: for example, for a noise model that corresponds to a BSC with error rate $\epsilon$, the EM algorithm takes the following form:

We start with some estimates $\alpha_0^{(0)}$, $T_0^{(0)}$ for the parameters. The $p^{th}$ iteration (for $p = 1, 2, \ldots$) consists of 2 steps:

**Expectation (E-step):** Find the conditional estimates for $U_0$ and $\overline{\epsilon}$ :

$$U_0^{(p)} = E\left[ U_0 \mid g, \alpha^{(p)}, T_0^{(p)} \right]$$

$$\bar{\epsilon}^{(p)} = E\left[ \bar{\epsilon} \mid g, \alpha^{(p)}, T_0^{(p)} \right]$$

These estimates are ensemble averages taken with respect to the posterior distribution $P_{f|g}$.

**Maximization (M-step):** Find $T_0^{(p+1)}$, $\alpha^{(p+1)}$ such that:

$$E\left[ U_0 \mid \alpha^{(p+1)}, T_0^{(p+1)} \right] = U_0^{(p)}$$

$$E\left[ \bar{\epsilon} \mid \alpha^{(p+1)}, T_0^{(p+1)} \right] = \bar{\epsilon}^{(p)}$$

Note that, since the left hand side of the above expressions is independent of the data, it can be precomputed, so that this step may be implemented using a table lookup procedure.

[6] Consider an M-ary field f with Ising potentials, corrupted with 0-mean, additive white Gaussian noise with variance $\sigma^2 < \sigma^2_{max}$.

Suppose that

$$f_i \in Q = \{ q : q = q_0 + 2k\delta \ , \ k = 1, 2, \ldots, M \} \quad \text{for all } i .$$

We define the statistic $W_g$ as:

$$W_g = \frac{1}{N_C} \sum_{i,j} W( g_i , g_j ) \ ,$$

where g is the observation process; $N_C$ is the number of nearest-neighbor

pairs in the lattice;

$$W(g_i, g_j) = \begin{cases} -1, \text{ if } \hat{g}_i = \hat{g}_j \text{ and } |i-j| = 1 \\ 1, \text{ if } \hat{g}_i \neq \hat{g}_j \text{ and } |i-j| = 1 \\ 0, \text{ if } |i-j| \neq 1 \end{cases}$$

and $\hat{g}_i = q_0 + 2n\delta$, with n an integer such that

$$q_0 + (2n-1)\delta < g_i < q_0 + (2n+1)\delta$$

(note that it is possible that $\hat{g}_i \notin Q$).

Define

$$\Psi(r, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \int_r^\infty \exp[-x^2/2\sigma^2]\, dx$$

It is not difficult to see that

$$E[W_g \mid \sigma, T_0] = 1 - (A+B) + E[U_0 \mid T_0](A-B)$$

where $\overline{U}_0 = U_0 / N_C$;

$$A = Pr(W(g_i, g_j) = -1 \mid V(f_i, f_j) = -1)$$
$$B = Pr(W(g_i, g_j) = -1 \mid V(f_i, f_j) = 1)$$

(note that $E[U_0 \mid T_0] = \Psi(T_0)$ is data independent, and therefore, it can be computed off-line).

Assuming that

$$Pr(|f_i - f_j| = q \mid f_i \neq f_j) = \frac{1}{M-1}, \quad \text{for } q = 1, 2, \ldots, M-1$$

We can approximate A and B by:

$$A(\sigma) = 6a^2 + 4b^2 - 4ab - 4a + 1$$

$$B(\sigma) = \frac{1}{M - 1} ( - 3a^2 - 3b^2 + 2ab + 2a )$$

where $a = \psi(\delta,\sigma)$ and $b = \psi(3\delta,\sigma)$ . (The above approximation has been computed assuming that $\psi(5\delta,\sigma_{max}) \approx 0$ . If this is not true, more terms can be easily included).

Assuming, as before, that

$$E [ W_g \mid \sigma, T_0 ] = \overline{W}_g \text{ (computed from the data),}$$

we can find the optimal estimate for $(\sigma,T_0)$ as the global maximizer of the merit function (28) along the curve:

$$T_0 = \Psi^{-1} \left[ \frac{\overline{W}_g + A(\sigma) + B(\sigma) - 1}{A(\sigma) - B(\sigma)} \right]$$

using a "composite annealing" strategy.

**[7]** Since both the random field $f$ and the noise process are stationary, we have that

$$E [ ( \overline{U}_g - E [ U_g \mid \alpha, T_0 ] )^2 ] \sim \frac{1}{\text{\# of cliques of the lattice}}$$

so that this assumption becomes asymptotically correct for large lattices.

Reading List.

Abend, K. "Compound Decision Procedures for Unknown Distributions and for Dependent States of Nature". in Pattern Recognition pp. 207-249 L. Kanal, ed. Thompson Book Co. Washington, D.C. (1968).

Baker, H.H. and Binford, T.O. "Depth from edge and intensity based stereo" Proc. Seventh International Joint Conference on Artificial Intelligence, August, 1981, 631-636.

Barrow, H.G. and Tennenbaum, J.M. "Interpreting line drawings as three dimensional surfaces" Artificial Intelligence, 17, 75-117, 1981.

Blake, A. "Parallel computation in low level vision" Ph.D. Thesis. Univ. of Edinburg (1985).

Brady, J.M. Computing Surveys, 14, 3-71, (1982).

Besag, J. "Spatial interaction and the statistical analysis of lattice systems". J. Royal Stat. Soc. B 34 75-83 (1972).

Cross G.C. and Jain A.K. "Markov Random Field Texture Models". IEEE Trans. PAMI 5 25-39 (1983).

Dempster, A.P., Laird N.M. and Rubin D.B. "Maximum likelihood from incomplete data via the EM algorithm". J. Royal Stat Soc. B 39 1-38, (1977).

Drumheller, M. and Poggio, T. "On parallel stereo" IEEE trans. on PAMI (in press) (1986).

Elliot H.,Derin R.,Christi R. and Geman D. "Application of the Gibbs distribution to image segmentation" Univ. of Massachusetts Technical Report (1983).

Feller W. "An introduction to probability theory and its applications" Vol. I. Wiley, New York (1950).

Gallager R.G. "Information theory and reliable communication" J. Wiley (1968).

Geman S. and Geman D. "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images" IEEE trans. on PAMI 6, 721-741 (1984).

Grimson W.E.L. "From images to surfaces" MIT press, Cambridge, Mass. (1981).

Grimson W.E.L. "A computational theory of visual surface interpolation" Phil. Trans. Royal Soc. London, B 298 395-427 (1982).

Hansen A.R. and Elliot H. "Image segmentation using simple Markov field models". Comp. Graph. and Image Proc. 20 101-132 (1982).

Hastings W.K. "Monte Carlo sampling methods using Markov chains and their applications" Biometrika 57 97-109 (1970).

Hassner M. and Sklansky J. "The use of markov random fields as models of texture" Comput. Graph. and Image Proc. 12 357-370 (1980).

Hillis D. "The Connection Machine" Ph. D. Thesis MIT (1985).

Hildreth E.C. "Computation of the velocity field" Proc. Roy. Soc. London, B, 221, 189-220 (1984).

Hopfield J. and Tank D.W. "Neural computation of decisions in optimization problems" , Biol Cybern., 52, 141-152 (1985).

Horn B.K.P. "Determining lightness from an image" Computer Graphics and Image Processing, 3, 211-299 (1974).

Horn B.K.P. and Schunck B.G. "Determining optical flow". Artificial Intelligence, 17, 185-203 (1981).

Ising E. "Beitrag sur theorie des ferromagnetismus" Zeit. fir Physik 31, 253-258 (1925).

Julesz B. "Binocular depth perception of computer generated patterns" Bell Syst. Tech. J. 39, 1125-1162 (1960).

Kemeny J.G. and Snell J.L. "Finite Markov Chains" Van Nostrand, New York (1960).

Kirkpatrick S., Gelatt C.D. and Vecchi M.P. "Optimization by simulated annealing" Science 220, 671-660 (1983).

Koch, C., Marroquin J.L. and Yuille A. "Analog neuronal networks in early vision". Proc. Nat. Ac. of Science (in press) (1986).

Marr D. "Vision, A computational investigation into the human representation and processing of visual information" W.H.Freeman, San francisco (1982).

Marr D. and Poggio T. "Cooperative computation of stereo disparity" Science , 194, 283-287 (1976).

Marr D. and Poggio T. "From understanding computation to understanding neural circuitry" Neur. Res. Bull., 15, 470-488 (1977).

Marroquin J.L. "Surface reconstruction preserving discontinuities" , Artificial Intelligence Lab. Memo 792, MIT (1984).

Marroquin J.L. "Probabilistic solution of inverse problems" Ph.D. Thesis. MIT (1985).

Marroquin J.L., Mitter S.K. and Poggio T. "Probabilistic solution of ill-posed problems in computational vision" J. Am. Stat. Asoc. (in press) (1986).

Metropolis N. et. al. "Equation of state calculations by fast computing machines" J. Phys. Chem. , 21, 6, 1087 (1953).

O'Callahan J. "Human perception of homogeneous dot patterns". Perception, 3, 33 (1974).

Poggio T. "Vision by man and machine" Artificial Intelligence Lab. Memo 776. MIT (1984).

Poggio T. and Torre V. "Ill-posed problems and regularization analysis in early vision" A.I. Lab. Memo 773 , MIT (1984).

Poggio T., Torre V. and Koch C. "Computational vision and regularization theory", Nature, 317, 6035 (1985).

Poggio T. "Integrating vision modules with coupled MRF's" AI Working Paper 285. MIT (1985).

Reif F. "Fundamentals of Statistical and Thermal Physics". McGraw-Hill (1965).

Terzopoulos D. "Multiresolution computation of visible-surface representations". Ph.D. Thesis. MIT (1984).

Terzopoulos D. "Integrating visual information from multiple sources for the cooperative computation of surface shape". in "From pixels to predicates: Recent advances in computational and robotic vision" ed. A. Pentland, Ablex (1985).

Tikhonov A.N. and Arsenin V.Y. "Solutions of ill-posed Problems". Winston & Sons (1977).

Figure Captions:

Fig 1: Three typical configurations of a first order, ternary MRF with Ising potentials. The interaction strengths in the x and y directions ($a_x$, $a_y$) are: (a) $a_x = 1.0$, $a_y = .2$ (b) $a_x = a_y = 0.5$ (c) $a_x = a_y = 0.4$.

Fig 2: (a) Sample function of a first order, binary MRF with Ising potentials. (b) The previous pattern sent through a binary symmetric channel (error rate: 0.4). (c) MAP estimator. (d) Monte Carlo approximation to the MPM estimate.

Fig 3: (a) Sample configuraton of a ternary (with values 1, 2, or 3), first order MRF with Ising potentials. (b) Pattern (a) corrupted with additive, "white" Gaussian noise (only the integer part of the result is represented). (c) Approximation to the TPM estimate obtained after 20 iterations of the deterministic system given by equations (19) and (21). (d) Approximation to the TPM estimate obtained after 5 iterations of the deterministic system given by equations (19) and (20). (e) Approximation to the optimal TPM estimator, obtained after 500 iterations of the Monte Carlo (Metropolis) algorithm.

Fig 4: The neighborhood $N^3_i$ of site i is formed by the pairs: {1,2}, {2,3}, {3,4} and {4,1} (see text).

Fig 5: Average value of $U_0$ versus interaction strength fo a 30 × 30 binary Ising field with free boundaries. Solid line: Monte Carlo approximation; dashed line: deterministic system (19)/(21).

Fig 6: (a) Original binary image. (b) Output of a BSC with error rate 0.3 (assumed unknown). (c) Maximum likelihood estimator (deterministic approximation). (d) Behavior of the likelihood function (28) along the curve (27) (the minimum corresponds to the correct estimates for both the error rate and the interaction strength).

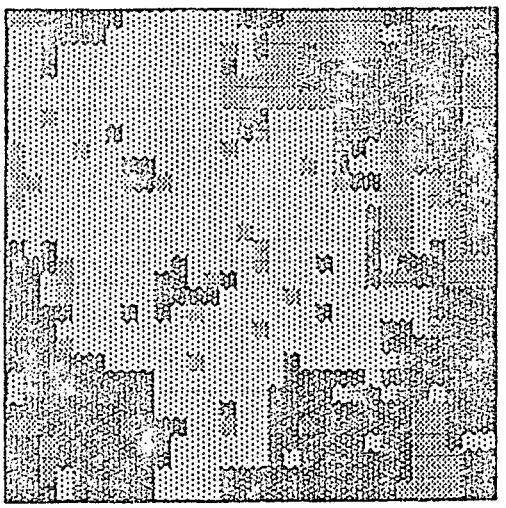
Fig 7: Cliques for the line field (a cross denotes a line element, and a circle, a "surface" site).

Fig 8: Observations of three rectangles at heights 2.0, 3.0 and 2.0, over a background at height 1.0 (height coded by gray level; a white pixel means the observation is absent at that point). (b) "Membrane" interpolation obtained with all lines turned "off". (c) Optimal estimate.

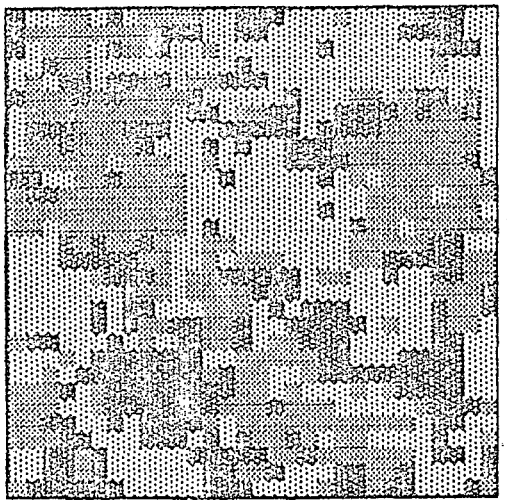
Fig 9: (a) Dense stereogram (density: 0.4) portraying a pyramid. (b) Fixed point for algorithm (26) (each panel represents the final state of a disparity layer, with a black pixel representing an "on" cell: from left to right, the disparity is: -3, -2, -1, 0, 1, 2 and 3). (c) Sparse stereogram (density: 0.1) portraying the same pyramid. (d) Fixed point for algorithm (26).

Fig 10: Formation of perceptual clusters. We show: the original dot patter (upper left) and the reconstructed objects for decreasing values of $\mathcal{C} = \alpha T_0$. The maximum likelihood estimate (i.e., the optimal clustering) is marked with an arrow.

(a)

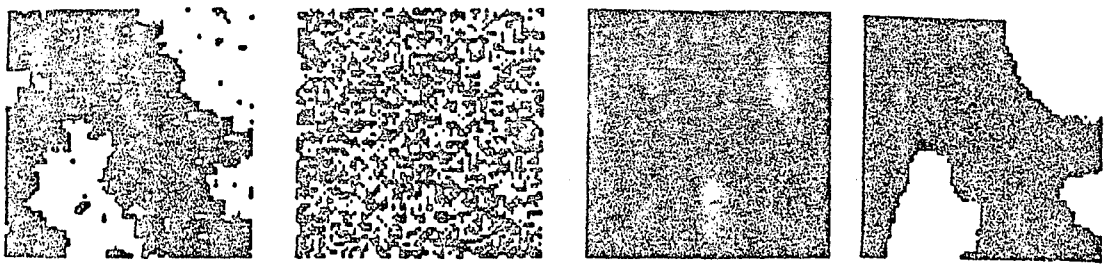


(b)



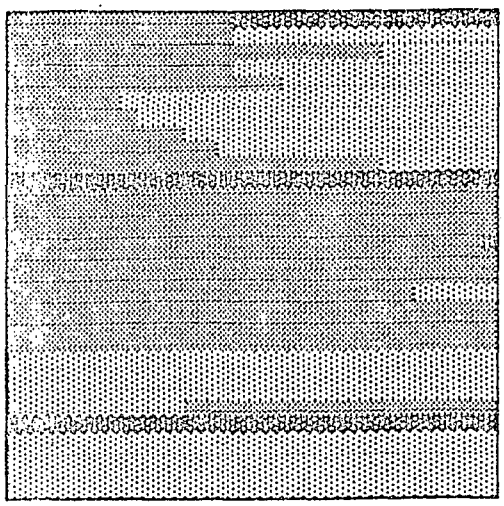
(c)

Figure 1.

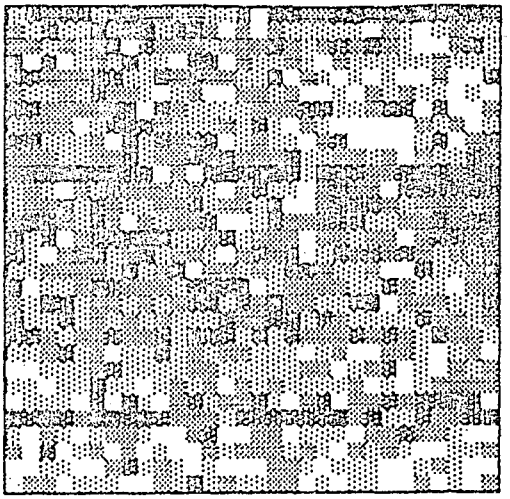


(a) (b) (c) (d)

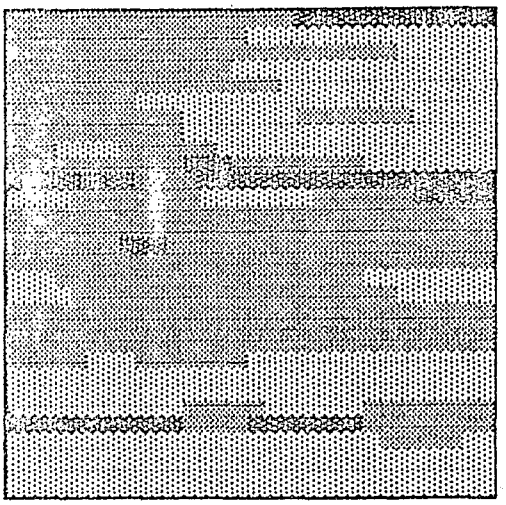
Figure 2



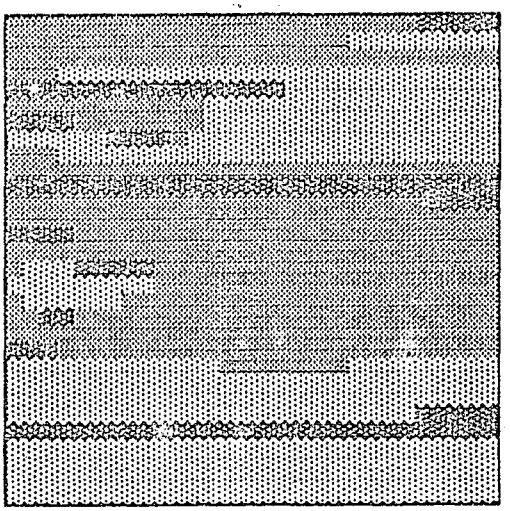
(a)



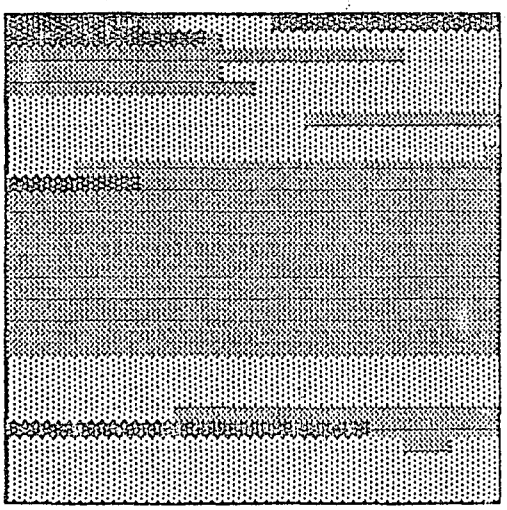
(b)



(c)



(d)



(e)

Figure 3

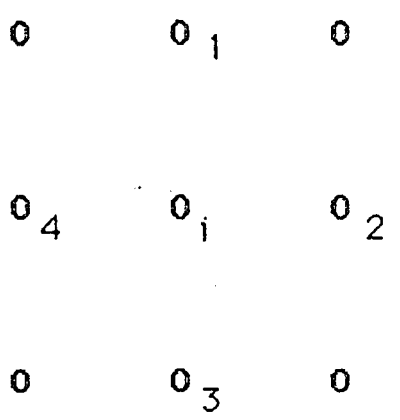



Figure 4

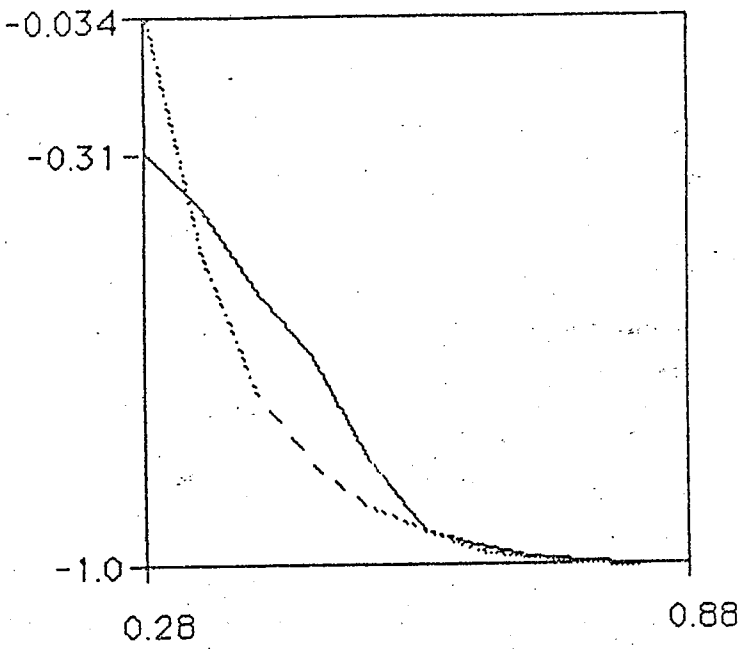



Figure 5.

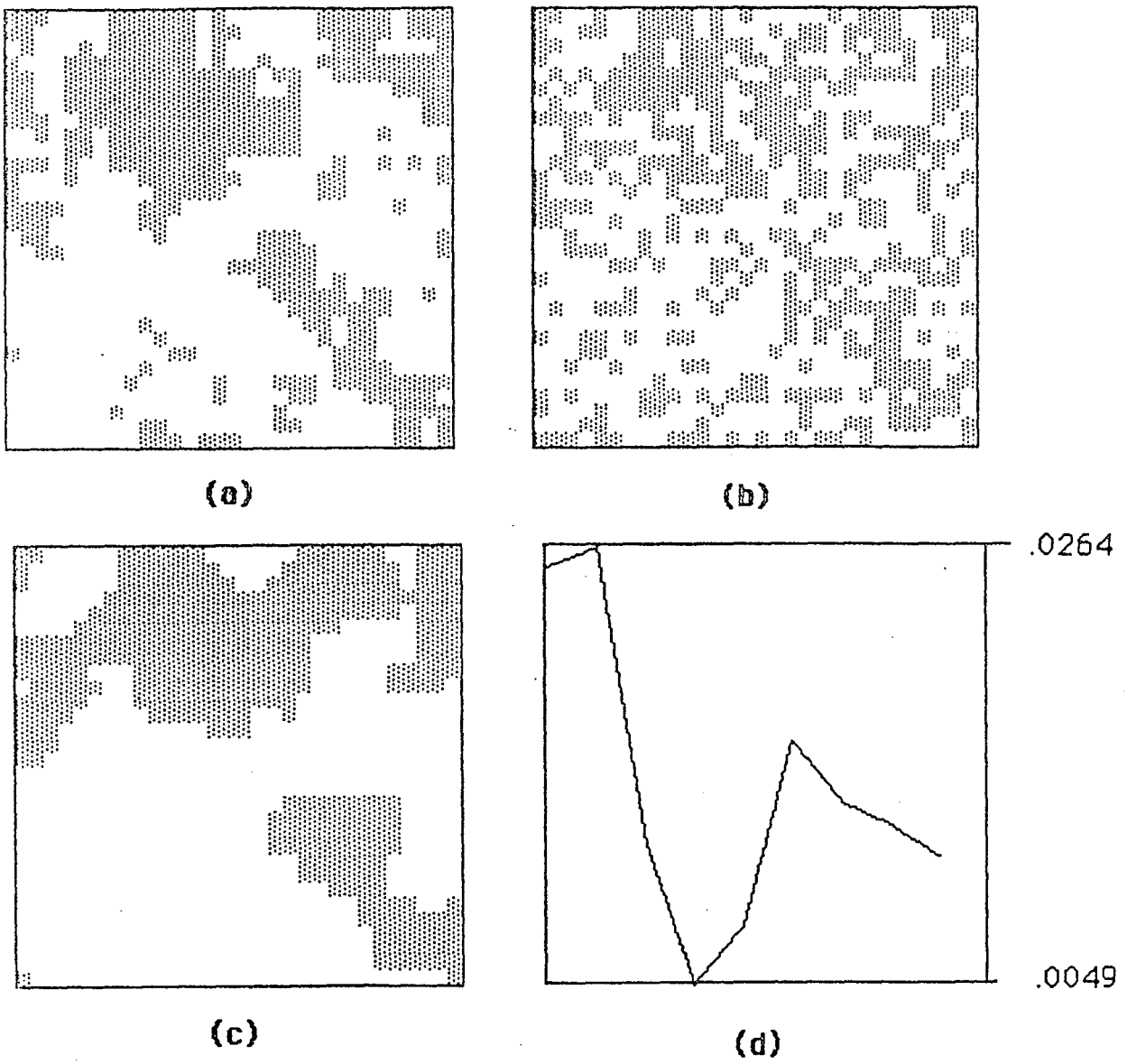



Figure 6.

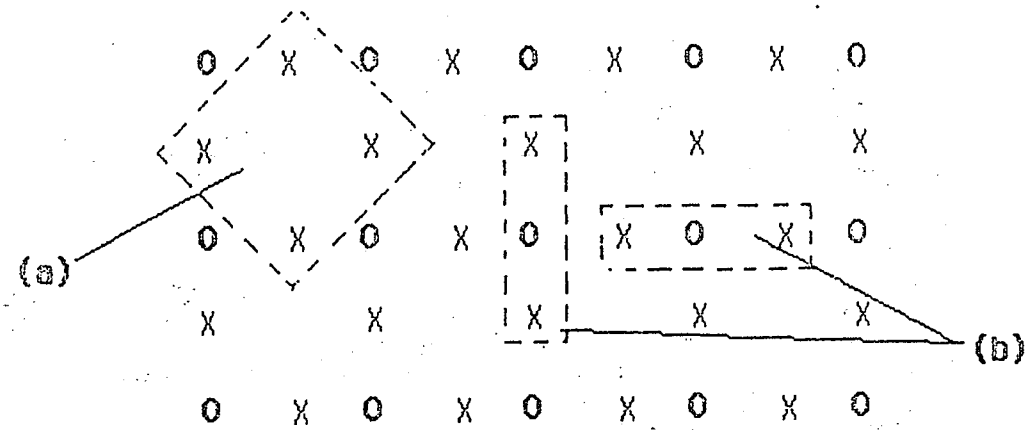
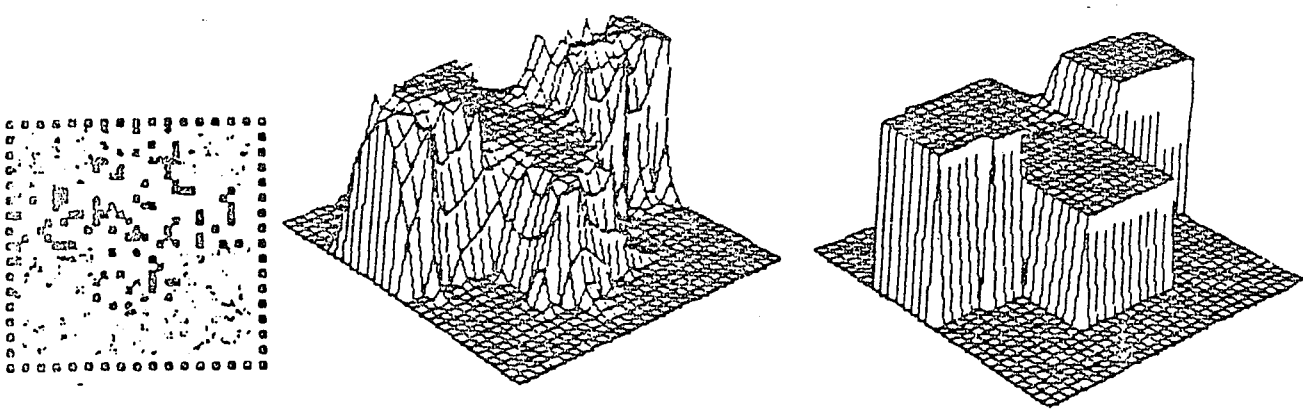
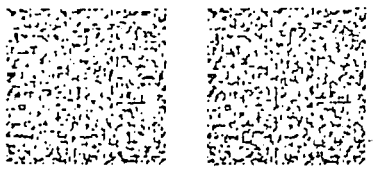



Figure 7

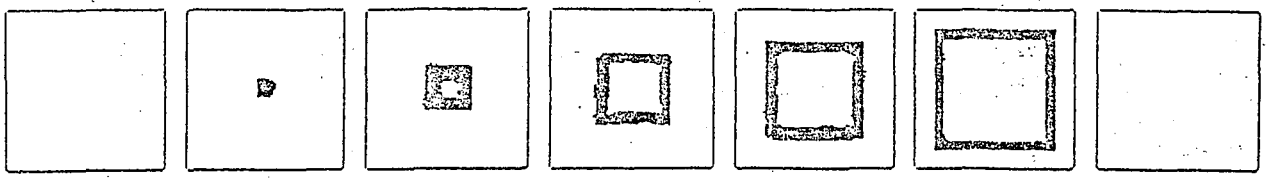


(a) (b) (c)

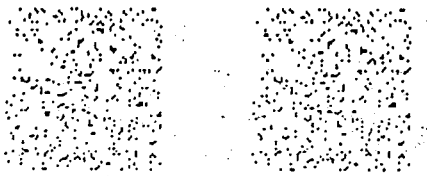
Figure 8



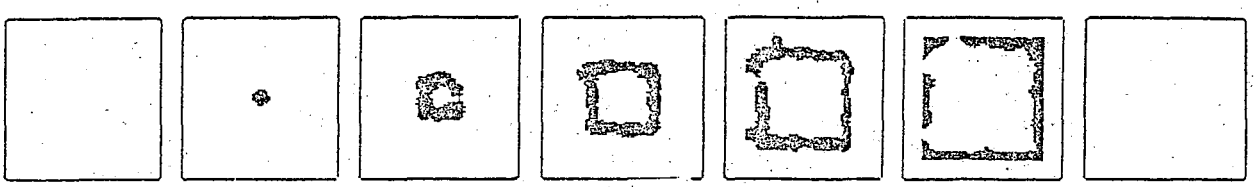
(a)



(b)



(c)



(d)

Figure 9

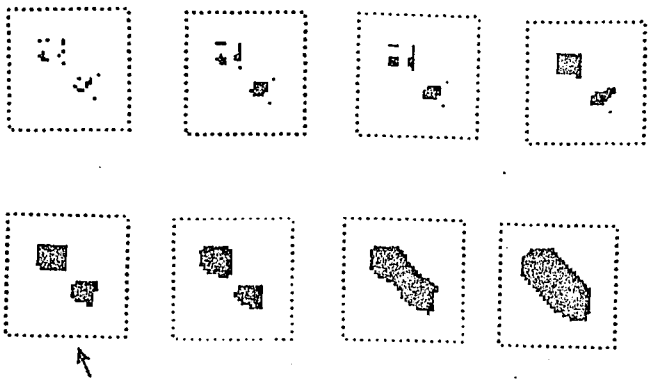


Figure 10