International Journal of Computer Vision, 4, 185–210 (1990)

The Theory and Practice of Bayesian Image Labeling

PAUL B. CHOU AND CHRISTOPHER M. BROWN
IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598; and Computer Science Department, University of Rochester, Rochester, NY 14627

Abstract

Image analysis that produces an image-like array of symbolic or numerical elements (such as edge finding or depth map reconstruction) can be formulated as a labeling problem in which each element is to be assigned a label from a discrete or continuous label set. This formulation lends itself to algorithms, based on Bayesian probability theory, that support the combination of disparate sources of information, including prior knowledge.

In the approach described here, local visual observations for each entity to be labeled (e.g., edge sites, pixels, elements in a depth array) yield label likelihoods. Likelihoods from several sources are combined consistently in abstraction-hierarchical label structures using a new, computationally simple procedure. The resulting label likelihoods are combined with a priori spatial knowledge encoded in a Markov random field (MRF). From the prior probabilities and the evidence-dependent combined likelihoods, the a posteriori distribution of the labelings is derived using Bayes' theorem.

A new inference method, Highest Confidence First (HCF) estimation, is used to infer a unique labeling from the a posteriori distribution that is consistent with both prior knowledge and evidence. HCF compares favorably to previous techniques, all equivalent to some form of energy minimization or optimization, in finding a good MRF labeling. HCF is computationally efficient and predictable and produces better answers (lower energies) while exhibiting graceful degradation under noise and least commitment under inaccurate models. The technique generalizes to higher-level vision problems and other domains, and is demonstrated on problems of intensity edge detection and surface depth reconstruction.

1 Introduction

We propose a mathematical framework, data structures, and algorithms for combining disparate sources of visual information (observational and prior) consistently into a raster representation of the scene. The representation may contain both symbolic (segmentation) and quantitative (reconstruction) information. The computation is formulated as a labeling problem: for each image location (site) find the appropriate attribute or label (such as depth, or whether it is an object boundary) describing the corresponding portion of the scene. The work involves the representation of knowledge, reasoning procedures for combining distinct bodies of knowledge, and methods for deriving scene properties from available knowledge. We have successfully applied the framework to two instances of the labeling problem—boundary detection from irradiance data and surface reconstruction from irradiance and range data. The paper deals with the following questions. (1) How to integrate multimodal visual data in a hierarchically structured hypothesis space. (2) How to incorporate a priori spatial knowledge with visual observations. (3) How to compute the optimal solution given a nonconvex a posteriori probability distribution of the possible solutions. (4) How simultaneously to reconstruct and segment the surfaces in a three-dimensional scene using sparse depth measurements and intensity observations.

The answers we propose to the four questions are, respectively, the following:

1. A computationally simple, probabilistically justified procedure that provides consistent and coherent integration of early visual observations using hierarchically structured label trees.

2. A priori spatial knowledge is encoded as potential energy functions that determine the distribution of a Markov random field (MRF). The a posteriori probability distribution is thus derived by combining the pooled external observations and the a priori distribution.
3. A new method, called Highest Confidence First (HCF), provides a robust and efficient technique for solving the labeling problem given the a posteriori probability distribution. HCF embodies the principles of graceful degradation and least commitment. It is a sequential deterministic calculation whose running time is predictable and small.
4. HCF is extended to handle both symbolic and numerical labels simultaneously using coupled MRFs. The unified treatment is applied to the problem of three-dimensional surface segmentation and reconstruction from intensity discontinuity and sparse depth data.

2 Probabilistic Information Fusion and the Labeling Problem

Most research on evidence combination has focused on updating the "belief" in a given hypothesis about an individual entity when bodies of new evidence become available [1, 2]. When dealing with problems involving large number of entities with local interactions (such as the labeling problem), this approach of maintaining "marginal belief" involves propagating the effects of updating local "belief" to other sites. For instance, the classic probabilistic relaxation techniques [3, 4, 5] use rather ad hoc and heuristic rules to update a set of weights, reminiscent of a probability distribution of the labels, associated with each site. It is difficult, if not impossible, to interpret their weights after a couple of iterations from a Bayesian decision point of view [6, 7].

Instead of updating marginal beliefs, our approach maintains joint probability distributions of the sites by decoupling the notion of external evidence and a priori knowledge. Since bodies of evidence based on local image observations bear directly upon individual sites, they can be combined locally without having to interfere with other sites. On the other hand, a priori knowledge is mostly about the interactions among the sites. It is best described by a joint probability distribution of all variables. When a priori knowledge and external evidence are combined using Bayes' rule, the resultant distribution reflects the a posteriori belief in the global configurations. Inference methods can thus be applied to find the true labeling based on the a posteriori beliefs. Figure 1 gives a schematic view of this approach.

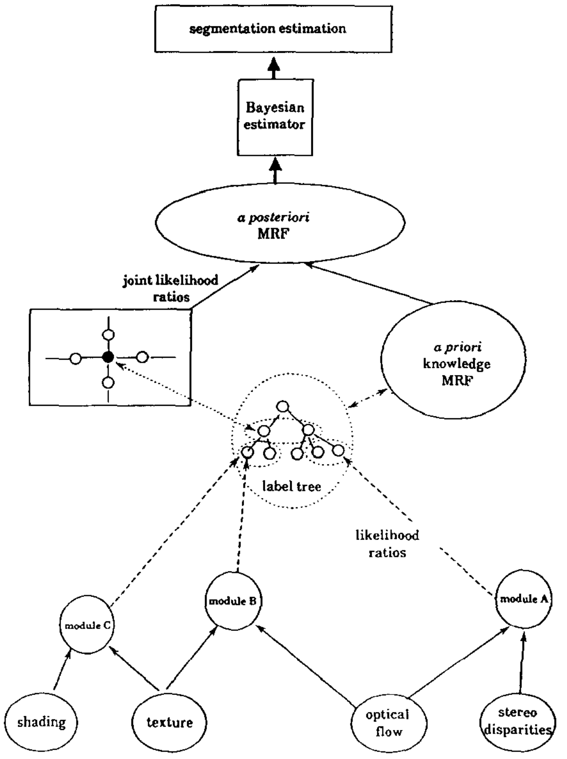


Fig. 1. System overview.

2.1 Hierarchical Integration of Early Visual Observations

Unordered symbolic labels are commonly used for the high-level representations of a scene. Each label corresponds to some event directly or indirectly observable in the scene. Labeling a site has the semantics of hypothesizing the occurrence of the corresponding event ("what?") at the corresponding location ("where?"). Since the existence and uniqueness of the label assignments are assumed, the corresponding set of events must be mutually exhaustive and exclusive. Frequently, certain subsets of the events have a semantic interpretation, and can be organized as an abstraction or causal hierarchy or tree. Each internal node in the tree represents the disjunct of its son events.

Figure 2 shows one organization of knowledge about various types of image edges. An abrupt change of the

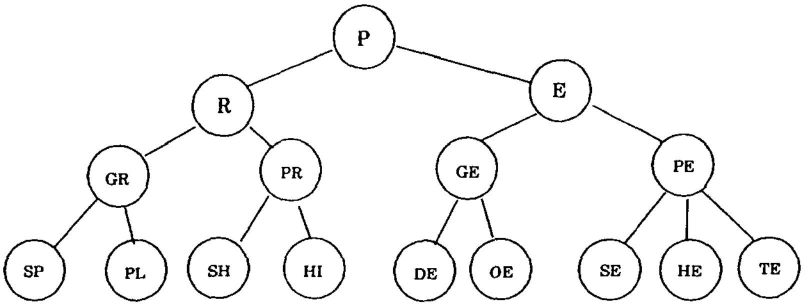


Fig. 2. A label tree. E—edge; R—region; PE—photometrical edge; GE—geometrical edge; PR—photometrical region; GR—geometrical region; TE—texture edge; HE—highlight edge; SE—shadow edge; OE—orientation edge; DE—depth edge; HI—highlight region; SH—shadow region; PL—planar region; SP—spherical region.

image irradiance (edge) may be caused by the variations of underlying scene structure (geometrical edge) or surface reflectance (photometric edge), and the changes of surface reflectance may be due to the variation of surface albedo (e.g., texture edge) or the amount of incident light (e.g., shadow edge). It is desirable to organize various types of edges into such trees. It is often vital to have the fincst-level descriptions of the edges for object recognition tasks (when obstacle avoidance is the primary concern, only depth discontinuities are of interest.) However, a visual module such as an intensity edge detector might not be able to tell geometrical edges from photometric edges.

One advantage of hierarchically structured trees, pointed out by Gordon and Shortliffe [8] and later by Pearl [1], is the ability to represent a particular piece of knowledge at the appropriate level of abstraction. Also, in practice, it is easy independently to design and maintain modules that are experts at detecting particular events. Their opinions about the label events can then be pooled according to the known semantic relations. In addition, many visual tasks can simultaneously share the knowledge accumulated in a label tree. Since events on different branches of the tree are mutually exclusive, every cross-section corresponds to a mutually exclusive and exhaustive set of events. Instances of the labeling problem can thus be defined with respect to appropriate cross-sections in a label tree dependent on the particular goals of the tasks.

We have developed a novel evidence-combination technique for a hierarchy of hypotheses [9, 10]. This scheme, besides having all the desirable characteristics listed in [1], has the following advantages:

1. The computations involved are extremely simple. Simpler and fewer messages must be passed, compared with Pearl's procedure [1]. Normalizations are never needed since relative degrees of confirmation/disconfirmation are maintained instead of probabilities.
2. This scheme decouples the notion of evidence and belief. That is, the evidence can be collected and combined consistently and coherently without having to maintain probability distributions. In the next section we show this characteristic is very helpful when the prior knowledge is represented as a Markov random field.

Let us introduce some notation. Represent an image as a set of sites indexed by the set $S = \{s_1, s_2, \ldots, s_N\}$. Without loss of generality, assume all sites have the same set of interesting labels organized as a hierarchically structured tree H (e.g., figure 2). Node H_l denotes the hypothesis that s can be labeled l. Each internal node stands for the disjunct hypothesis of its sons. For convenience purposes, let γ_l denote the set of sons of H_l. Let $L = \{l_1, l_2, \ldots, l_Q\}$ be a mutually exclusive and exhaustive set of labels of H. Let $X_s \in L$ be a random variable associated with $s \in S$. A labeling ω of the image with respect to L is a realization of the set of random variables $X = \{X_s, s \in S\}$. Let $\omega_s = \omega(X_s) \in L$ represent the label attached to s_i according to the labeling ω, and Ω be the set of all labelings—the admissible solution space of the labeling problem with respect to S and L.

We consider early visual computations as the computations performed by a set of independent modules. A module encodes a piece of knowledge that relates image observations to some label events. The input for these modules is noise-corrupted visual data, such as image irradiance, texture, stereo disparities, etc. When forming an opinion about site s, a module may restrict its consideration of input to some spatial region dependent on s. Typically, the region will include s and its spatially adjacent sites. Also, a module might not use all the input available for a given site, that is, it might use only irradiance when other data are also available. The subset of input used to form an opinion about s is observation O_s.

In our treatment, the *opinions* of the modules are presented in terms of likelihood ratios. For example, module A is an expert on a label l. After observing O_s, the module reports its opinion on l as a likelihood ratio:

$$\lambda_l = \frac{P(O_s|l)}{P(O_s|\neg l)}$$

The semantics of likelihood ratios are well known [11, 12, 13]. Confirmation of l based on O_s is encoded by $\lambda_l > 1$, and disconfirmation by $\lambda_l < 1$.

More generally, it is possible that a module can suggest several labels that would explain the data. In such cases, the module distributes the support of the evidence to the set of labels and its complement. More precisely, if module A is an expert on the subset $L_A \subseteq L$, it reports one likelihood ratio for each label in L_A. In this more general case, a *likelihood ratio* is the probability of the observation given that one label truly applies divided by the probability of the observation should none of the labels in L_A apply. For example, the likelihood ratio reported for label $l_i \in L_A$ is

$$\lambda_i = \frac{P(O_s|l_i)}{P(O_s|\neg(\bigcup_{l \in L_A} l))}$$

Note that we define $P(O|\neg(\bigcup_{l \in L_A} l))$ to be 1 when L_A is exhaustive.

The values of likelihoods and likelihood ratios can be derived from stochastic models of relationships between the label events and the observations [14, 15, 16], or estimated from statistical data [13]. Sometimes they are subjectively assigned by experienced human experts [12].

We assume conditional independence between spatially distinct observations:

$$P(O^A|\omega) = \prod_{s \in S} P(O_s^A|\omega_s) \tag{1}$$

where the superscript A indicates the observations of the module A. This assumption has been used implicitly in numerous applications [17, 18, 19, 15, 16]. Conditional independence may not always hold. For example, the noise of an ultrasound image may be spatially dependent given the true scene due to the change in conductance. There are also cases in which O_s may contain information not only from site s but also from its adjacent sites. In such cases, the spatial independence assumption is still valid, but equation (1) needs to be modified so that labels for (more structured and complex) neighborhoods are used rather than labels for individual sites. Conditional independence is an interesting and sometimes vexing question that we do not propose to consider further here.

In combining evidence, numbers (α values) are maintained for each node of the label tree H, except for the root. Unlike other approaches, the numbers here do not indicate our states of belief in the hypotheses but rather the degrees of hypothesis confirmation or disconfirmation provided by the collected evidence. Technically, the semantics is that of a conditional probability (see Theorem 1 below). Initially, all α's are set to unity indicating "neither confirmed nor disconfirmed" before observing any evidence.

The evidence-combination algorithm can be best described in terms of local updating and message passing between the nodes of H. It plays a central role in the general labeling system. It generates the likelihood ratios of labels given the observations that are used in the labeling algorithm. Although its details are unnecessary to understand the rest of this work, it has the following important property.

THEOREM 1. Let α_l^t denote the α value associated with H_l at the consistent state t, and $P(O_t|l)$ be the probability of O_t given $X_s = l$, where O_t denotes the union of those observations that form the set of opinions that yields the state t. If $O_t \neq \emptyset$, then

$$\alpha_l^t = c_t P(O_t|l) \qquad \forall H_l \in h \tag{2}$$

where c_t is a constant depending only on t, given conditional independence assumptions between observations and between the probabilities of labelings at a node and the labels of its descendents (for details and proofs see [9]).

3 Spatial Priors and Markov Random Fields

Markov random fields have been used for image modeling in many applications for the past several years

[20, 21, 22, 23, 24, 18, 25, 26, 27, 17, 28, 29, 30]. In this section, we review the properties of MRFs and discuss how to encode prior knowledge in this formalism. We refer the reader to [31] for an extensive treatment of MRFs.

3.1 Noncausal Markovian Dependency

Let E be a set of unordered pairs (s_i, s_j)'s representing the "connections" between the elements in S. The semantics of the connections will become clear shortly. E defines a symmetric and nonreflexive neighborhood system $\Gamma = \{N_s | s \in S\}$, where N_s is the neighborhood of s in the sense that

1. $s \notin N_s$, and
2. $r \in N_s$ if and only if $(s, r) \in E$.

X is a *Markov random field* with respect to Γ and P, where P is a probability function, if and only if

$$(\textit{positivity}) \qquad P(X = \omega) > 0 \quad \text{for all} \quad \omega \in \Omega \tag{3}$$

$$\begin{aligned}(\textit{Markovianity}) \qquad P(X_s = \omega_s | X_r = \omega_r,\ r \in S,\ r \neq s) \\ = P(X_s = \omega_s | X_r = \omega_r,\ r \in N_s)\end{aligned} \tag{4}$$

The set of conditional probabilities on the left-hand side of equation (4) is called the *local characteristics* that characterizes the random field. It can be shown that the joint probability distribution $P(X = \omega)$ of any random field satisfying (3) is uniquely determined by these conditional probabilities [20]. An intuitive interpretation of (4) is that the contextual information provided by $S - s$ to s is the same as the information provided by the neighbors of s. Thus the effects of members of the field upon each other is limited to local interaction as defined by the neighborhood. Notice that any random field satisfying (3) is an MRF if the neighborhoods are large enough to encompass all the dependencies.

3.2 Encoding Prior Knowledge and Gibbs Distributions

The utility of the MRF concept for image labeling problems is that the prior knowledge about spatial dependencies among the image entities can be adequately modeled with neighborhoods that are small enough for practical purposes. Very often, the image entities are regularly structured and prior distributions on the image are homogeneous and isotropic. The following theorem further simplifies the specification of MRFs.

HAMMERSLEY-CLIFFORD THEOREM: A random field X is an MRF with respect to a neighborhood system Γ if and only if there exists a function V such that

$$P(\omega) = \frac{e^{-\frac{1}{T}U(\omega)}}{Z} \ \forall\ \omega \in \Omega \tag{5}$$

where T and Z are constants and

$$U(\omega) = \sum_{c \in C} V_c(\omega) \tag{6}$$

C denotes the set of totally connected subgraphs (cliques) with respect to Γ. Z is a normalizing constant and is called the *partition function.*

The probability distribution defined by (5) and (6) is called a *Gibbs distribution* with respect to Γ. The class of Gibbs distributions has been extensively applied to model physical systems, such as ferromagnets, ideal gases, and binary alloys. When such systems are in a state of *thermal equilibrium*, the fluctuations of their configurations follow a Gibbs distribution. In statistical mechanics terminology, U is the *energy* function of a system. The V_c functions represent the *potentials* contributed to the total energy from the local interactions of the elements of clique c. T, the *temperature* of the system, controls the "flatness" of the distribution of the configurations.

The MRF-Gibbs equivalence not only relates the local conditional probabilities to the global joint probabilities, but also provides a conceptually simpler way of specifying MRFs—specifying potentials. The importance of the joint probabilities will soon become evident. Based on (4) and the Hammersley-Clifford theorem, the local characteristics can be computed from the potential function through the following relation:

$$P(X_s = \omega_s | X_r = \omega_r, r \neq s) = \frac{e^{-\frac{1}{T}\sum_{c \in C_s} V_c(\omega)}}{\sum_{\omega'} e^{-\frac{1}{T}\sum_{c \in C_s} V_c(\omega')}} \tag{7}$$

where C_s is the set of cliques that contain s, and ω' is any configuration of the field that agrees with ω everywhere except possibly s.

There has been some work that applies statistical methods to estimate parameters used for specifying MRFs. The standard approach is maximum likelihood estimation [28, 32]. Variations of this approach are also used [33, 22]. Elliott and Derin [34] use a least-square-fit method to estimate parameters of their texture models. These methods are good when many uncorrupted realizations are available, such as in the case of natural texture modeling. When such data are difficult to acquire, choosing the clique potentials on an ad hoc basis has been reported to produce promising results [24, 23]. Our experiments (see sections 5 and 6) have also shown good results. These results are not surprising since the notion of clique potentials provides a simple mapping from "qualitative" spatial knowledge to numeric values of the parameters specifying the MRFs.

3.3 A Posteriori Markov Random Fields

In this section, we move our attention to the relationships of segments of the image S. Let β_s denote the set of α's associated with $s \in S$, and $\beta_s(l)$ be the α value for label l in β_s. Define a *global consistent state* to be a state of the β's at which each β_s is in a consistent state.

Assume that the prior knowledge about the image is represented as an MRF X over S, $Z_s \in L$—a mutually exclusive and exhaustive label set in H, with respect to a neighborhood system N. The Gibbs measure that characterizes the prior MRF is

$$P_0(\omega) = \frac{e^{-\frac{1}{T} U_0(\omega)}}{Z_0} \tag{8}$$

where

$$U_0(\omega) = \sum_{c \in C} V_c(\omega)$$

and Z_0 is a normalizing constant.

Given equation (1), theorem 1, and Bayes' rule $P(AB) = P(A|B)P(B)$, the a posteriori Gibbs measure of a configuration ω at a global consistent state t can be computed as

$$P_t(\omega) = \frac{e^{-\frac{1}{T} U_t(\omega)}}{Z_t} \tag{9}$$

where the a posteriori energy is

$$U_t(\omega) = \sum_{c \in C} V_c(\omega) - T \sum_{s \in S} \ln[\beta_s^t(\omega_s)] \tag{10}$$

It is easy to see that the a posteriori Gibbs measure characterizes a MRF over S with respect to the neighborhood system N with the local characteristics

$$P_t(X_s = \omega_s | X_r = \omega_r, r \neq s) = \frac{e^{-\frac{1}{T} \sum_{c \in C_s} V_c(\omega) + \ln[\beta_s^t(\omega_s)]}}{\sum_{\omega'} e^{-\frac{1}{T} \sum_{c \in C_s} V_c(\omega') + \ln[\beta_s^t(\omega'_s)]}} \tag{11}$$

Equations (10) and (11) imply that only simple local operations are involved in updating the energy measure and local characteristics as new opinions from the early visual modules become available. Therefore, inference methods depending only upon these measures, such as stochastic MAP [24] and MPM estimations [23], can easily be implemented in this framework.

4 Highest Confidence First Estimation

At various levels of a visual hierarchy, available data and knowledge are used to estimate (infer) more condensed, symbolic information and to reduce the amount of data passed between levels of visual tasks. In this section, we describe how to make label inferences using a Bayesian decision rationale with imperfect knowledge represented as the a posteriori Gibbs distributions described in the previous section.

The goodness of a labeling $\hat{\omega}$, following the Bayesian formalism, is evaluated in terms of its expected loss,

$$\text{loss}(\hat{\omega}) = \sum_{\omega \in Q} \text{loss}(\hat{\omega}, \omega) P(\omega) \tag{12}$$

where loss $(\hat{\omega}, \omega)$ is a penalty associated with the estimate $\hat{\omega}$ while the "truth" is ω, and $P(\omega)$ is the (a posteriori) probability of ω.

One question concerning the applicability of (12) is which loss function should be used for a given task. Except for a few simple cases, the answer to this question usually depends on subjective judgements. One popular choice is assigning the same penalty to incorrect estimates: loss $(\hat{\omega}, \omega)$ equals to a constant (positive) value whenever $\hat{\omega} \neq \omega$, and 0 otherwise. Using this loss function, the configuration minimizing (12) maximizes the a posteriori probability $P(\omega|O)$, and therefore minimizes the a posteriori energy (10) in the MRF formalism. This maximum a posteriori (MAP) criterion has been widely applied to the labeling problem [35, 24, 18, 25, 26]. Marroquin et al. [23], also Besag [17], suggest that the number of mislabeled image entities of an estimation is a better loss measure for the labeling problem. They derive the maximizer of the a posteriori marginals (MPM) estimation—choosing the configuration $\hat{\omega} = (\hat{\omega}_{s_1}, \ldots, \hat{\omega}_{s_N})$ such that

$$\hat{\omega}_s = \max_{l \in L} P_s(l|O) \; \forall \; s \in S$$

where $P_s(l|O)$ denotes the a posteriori marginal probability of l on s. Their experiments show the MPM estimator to be superior to the MAP criterion when the signal-to-noise ratio is low.

There are three problems with the MAP and MPM estimations for the labeling problem. The first problem is the cost of calculation of the estimates. Notice that the rationale of minimizing the loss function in (12) does not take the cost of computation into account, despite the fact that computational cost is usually a primary consideration in image understanding MRF applications because of their immense configuration spaces. A suboptimal but useful estimator with an effective computation procedure is more valuable than an intractable optimal estimator. It is currently believed that the exact evaluation of MRF statistical moments, and therefore (12) is generally impossible since no analytic solutions exist [21, 24]. For the same reason, MAP and MPM cannot be exactly determined, except for some simple energy functions.

The second problem has to do with the large-scale characteristics induced by using MRFs to model local dependencies among the neighboring sites. Besag [17] points out that even relatively simple MRFs, such as the binary Ising model, exhibit positive correlations over arbitrarily large distances when adjacent sites have high probability to be the same label. Thus there is a strong tendency to form infinitely large patches of a single label. This phenomenon has also been observed repeatedly in our experiments with Monte Carlo simulations of MRFs.

The third problem concerns the match between the image of interest and the prior model. It is possible for the true labeling of an instance of the problem to be very unlikely according the a priori distribution. Therefore, the a posteriori probability of the true labeling and the corresponding marginal probabilities are relatively low in such cases, even though at some image sites there may be external evidence that strongly supports or refutes certain labels.

In this section we describe a new inference method—highest confidence first estimation, based on the a posteriori Gibbs distribution described in the previous section. Compared to previous methods, it is computationally inexpensive since it is a deterministic algorithm with local computations and global scheduling that makes decisions in the order of their importance. It is less affected by the large-scale characteristics of the MRFs because it tends to create small clusters of labels that are firmly based on the relative strength of the external evidence. It is more robust against inaccurate prior models because it favors prior expectations only when image evidence is weak. Two sets of applications discussed in section 5 and section 6 illustrate these points.

4.1 Previous Stochastic and Deterministic Relaxation Methods

MAP and MPM estimations can be approximated using procedures that generate configurations according to Gibbs distributions in the form of (9) such as the Metropolis algorithm [36] and the Gibbs sampler [24]. The necessary iterative stochastic relaxation algorithms are often implemented by simulated annealing, but the cost is intolerable for many applications.

This cost motivated work on deterministic techniques to approximate MAP. For vision systems that require predictable results in reasonable time periods, possible alternatives to the stochastic relaxation scheme include using suboptimal estimation criteria and/or heuristics in searching for solutions [18, 37]. One approach is the Metropolis algorithm without randomness: Start with an initial configuration. At each iteration through the image sites, the state of each site is either changed to the state that yields maximal decrease of the energy, or is left unchanged if no energy reduction is possible. The process always stops at a local minimum when no more changes can be made. Since every step of the energy reduction attempts to maximize a local conditional probability, Besag [17] names this method the Iterative Conditional Modes (ICM) estimation. In a parallel implementation, convergence is assured if the neighboring entities are not updated simultaneously.

The local minimum obtained by ICM may be far from optimal. Two enhancements are apparently helpful:

1. Start with a better initialization of the MRF. One possibility is to use the *maximum-likelihood estimates* (MLE)—$X_s(0) = \omega_s$ if $\max_{l \in L} P(O_s|l) = P(O_s|\omega_s)$ [17].
2. Escape from shallow valleys by changing the states of more than one entity at once [26].

Using (1) is not adequate to achieve robust estimations. The error rate of local MLEs can be low only when the likelihood function correctly models the relation between the label hypotheses and the observations, and, more importantly, there must be significant differences among the likelihoods of the hypotheses. Frequently these conditions cannot be met, resulting in initial configurations far away from the true labeling in the energy space. Moreover, since the energy space is usually characterized by many local minima, significantly different estimates may result under different visiting orders to the sites, even with the same initial estimate. Cohen and Cooper's procedure [26] with reasonable initializations has shown good results. It uses extra computations in exchange for better performance, and can be used as a postprocessing step following the method presented here.

4.2 Estimation with Highest Confidence First

We have seen that an estimation method should possess the following properties.

Efficiency: The cost of computation meets the demands of the visual tasks. Ideally, an iterative procedure should not only be deterministic, but at each step should make the maximum improvement, thus at any stage it should embody the best current estimate.

Predictability: The final estimate depends only on the inputs and the chosen a priori distribution. Thus the user is not responsible for choosing important performance-affecting parameters such as initial estimate, annealing schedule, and visiting order.

Robustness: The estimates degrade gracefully with the increase of noise and modeling error, and they are unaffected by the large-scale characteristics of the chosen MRFs.

The highest confidence first (HCF) method introduced in this section satisfies the above requirements, while existing methods generally do not. HCF blends the initialization into the estimation process. Instead of working only in the configuration space, HCF constructs a configuration with a local minimal energy measure by following a path, suggested by the observations, in an augmented space. Observable evidence and spatial prior knowledge are combined in the process of the construction, resulting in robust estimates and better efficiency. The details of HCF are described next.

4.2.1 Augmented Search Space. Recall that $L = \{l_1, \ldots, l_Q\}$ is a set of mutually exclusive and exhaustive labels with respect to which the labeling problem is defined, and Ω denotes the corresponding configuration space. The a posteriori knowledge about the labelings is represented by a Gibbs distribution; the a posteriori probability of a labeling ω, according to (9) is

$$P(\omega|O) = \frac{e^{-\frac{1}{T}\{\sum_{c \in C_s} V_c(\omega) - T \sum_{s \in S} \ln[P_s(\omega_s|O_s)]\}}}{Z} \qquad (13)$$

Let $\bar{L} = L \cup \{l_0\}$ denote the *augmented label set*, where l_0 is the null label corresponding to the "uncommitted" state in the construction. Let $\bar{\Omega} = \{\omega = (\omega_1, \ldots, \omega_N) | \omega_s \in \bar{L}, \forall\, s \in S\}$ denote the *augmented configuration space*. The basic idea of HCF is to construct a sequence of configurations $\omega^0, \omega^1, \ldots$ of $\bar{\Omega}$ with the starting configuration $\omega^0 = (l_0, \ldots, l_0)$, and a terminal configuration—the final estimate $\omega^f \in \Omega$, where the energy measure $U_O(\omega^f)$ is a local minimum with respect to Ω.

We say a site s has *committed* to a label $l \in L$ at step t of the construction if $\omega_s^t = l$, and it is uncommitted if $\omega_s^t = l_0$. We impose a rule that states once a site has committed to a label, it cannot nullify its commitment, but it is allowed to *change* its commitment to other labels of L. The rationale behind this rule will soon become clear.

Define the *augmented a posteriori local energy* of $l \in L$ with respect to $s \in S$ and a configuration $\omega' \in \bar{\Omega}$ as

$$E_s(l) = \sum_{c: s \in c} V'_c(\omega') - T \ln P(O_s|l) \qquad (14)$$

where $\omega' \in \bar{\Omega}$ is the configuration that agrees with ω everywhere except $\omega_s' = l$, and V_c' is 0 if $\omega_r = l_0$ for any r in c, otherwise it is equal to V_c—the potential function. This measure quantifies the goodness of a label with respect to the current configuration of the neighbors. It is related to the conditional probabilities (local characteristic) with respect to an MRF that has the same potential functions as the prior MRF, but consists of only the committed sites. Notice that only committed neighbors contribute to this measure, thus uncommitted sites do not actively influence others' commitments based on this measure. However, an uncommitted site always takes into account the states of the active neighbors when making a commitment.

*4.2.2 **Local Stability Measures***. To ensure the quality of the resulting estimate ω^f, we impose the following rule that decides the "updating" order: At each stop of the construction, only the least "stable" site is allowed to change its label or make its commitment. The *stability* of s with respect to the current configuration ω is defined as follows:

$$G_s(\omega) = \min_{k\in L, k\neq \omega_s} \Delta E_s(k, \omega_s) \qquad \text{if } \omega_s \in L \qquad (15a)$$

$$G_s(\omega) = -\min_{k\in L, k\neq j} \Delta E_s(k, j) \qquad \text{if } \omega_s = l_0 \qquad (15b)$$

where j in (15b) satisfies: $j \in L$ s.t. $E_s(j) = \min_{k\in L} E_s(k)$, and $\Delta E_s(j, k) = E_s(j) - E_s(k)$ with respect to ω.

Thus stability is a combined measure of the observable evidence and the a priori knowledge about the preferences of the current state over the other alternatives. A negative value of G indicates that a more stable (lower energy) configuration will result from an alternative commitment. By equation (15b), all uncommitted sites have negative G. The magnitude of G corresponds to how much energy would be lost or gained by changing the current state: the larger the negative value of G, the more confidence we have in a decision to change state. Since a site has no effect on its neighbors unless it has committed, the sites with large likelihood ratios of one label over the others—strong external evidence in favor of a label—are visited early in the construction sequence. Observe that when no neighbor is active, the augmented local energy measure reduces to the local likelihood of the label. Therefore, commitments under such circumstances are equivalent to the local maximum-likelihood estimates. The sites without strong evidence from the observations will take the neighbors' configuration into account when making their commitments. As mentioned previously, such commitments are equivalent to the conditional modes. An early commitment will be altered if the neighbors' later commitments are strongly against it—thus an error estimate based on local evidence can be corrected when more contextual information becomes available.

*4.2.3 **Serial Implementation***. The highest confidence first estimation method can be implemented serially with a heap (priority queue) maintaining the visiting order of the construction according to the values of Gs in such a way that the *top* of the heap is the site with the smallest G value. Updating the *top*'s decision will cause the changes of its neighbors' G-values, and therefore the structure of the heap. The following is pseudo code for the HCF algorithm.

```
  $\omega = (l_0, \ldots, l_0)$;
  $top$ = Create_Heap($\omega$);
  while ($G_{top} < 0$) {
     $s = top$;
     Change_State($\omega_s$);
     Update_G($G_s$);
     Adjust_Heap($s$);
     foreach ($r \in N_s$) {
        Update_G($G_r$);
        Adjust_Heap($r$);
     }
}
return($\omega$);
```

Change_State(ω_s) changes the current state ω_s of s to the state l such that $\Delta E_s(l, \omega_s) = \min_{k\in L, k\neq\omega_s} \Delta E_s(k, \omega_s)$ if if $\omega_s \in L$, or $E_s(l) = \min_{k\in L} E_s(k)$ if $\omega_s = l_0$. Upon this change taking place, the stability of s changes to positive. Update_G is invoked for every site that is affected by this change, namely the neighbors of s according to (14), to update their stability measures with respect to the new configuration. Adjust_Heap(r) maintains the heap property by moving r up or down according to its updated G-value.

The HCF algorithm always returns in finite time, with a feasible solution with locally minimal energy. This implementation takes $O(N)$ comparisons to create the heap and $O(\log(N))$ to maintain the heap invariance for every visit to a site, provided the neighborhood size is small relative to N—the number of sites. The overheads of heap maintenance are well repaid since the procedure makes progress for every visit, in contrast to the iterative relaxation procedure that may make only few changes per iteration (N visits). Our edge detection experiments (section 5) show that on the average, less than one percent of the sites are visited more than once using the proposed algorithm while the deterministic relaxation procedure takes around 10 iterations to reach a local minimum. This advantage becomes more evident as the number of sites gets larger.

4.3 Discussion and Possible Extensions

The HCF estimation method meets two important design principles for visual procedures suggested by Marr [38], namely, the principle of graceful degradation and the principle of least commitment. A common effect of lowering signal-to-noise ratio is the decrease of the feature saliency. For instance, a sharp edge in a badly degraded image may appear rather weak. As suggested previously, the search is guided by salient features; the utilization of contextual information increases as the degree of saliency decreases. Therefore, the results degrade gracefully with the increase of the noise level. Also, spatial priors are used less when external evidence is strong, thus model inaccuracy and large-scale characteristics are less likely to affect the estimates at the sites with apparent answers. There are two more advantages of delaying the commitments of the sites with weak observations: (1) To minimize the possibility of *undoing* previous commitments. It is likely that those sites would make incorrect commitments without enough contextual information, therefore they should commit late, and the principle of least commitment follows. (2) To reduce the chance of misinforming other sites. A site can do better without the incorrect information of a neighbor.

The concept of highest confidence first can be used as a heuristic search strategy for large state-space optimization or a rule for the nodes of a cooperative network to reach mutual agreement. It can be extended in many directions to achieve better results. Let us look more closely at the construction process of the HCF estimate. At each stage, S_D consists of a set of isolated clusters. A cluster is a set of spatially connected (with respect to Γ) sites. We say two clusters are isolated from each other if none of the sites of a cluster is a neighbor of any site of the other cluster. Each cluster corresponds to an MRF with free boundaries in our formalism. When a site makes a commitment, a cluster is created or expanded, or clusters are merged. When a site changes a commitment, the energy of the corresponding MRF is reduced. Eventually, all the clusters are merged and the final estimate corresponds to a local minimum configuration of the corresponding MRF.

The notion of growing clusters suggests a natural partition of the image. At any instant, the sites belonging to the same cluster are tightly related, but they are independent of the members of other clusters. The addition of a new member to a cluster may change the commitments of the old members, but the changes are expected to be small due to the way the clusters are constructed. Therefore, it makes sense to compute the MAP estimates exactly for small clusters early in the construction to reduce the possibility of early mistakes without being affected by the large-scale characteristics of the fields. We believe that this technique generally leads to better results than are found with nonadaptive strip partitioning [18].

The process of growing clusters is similar to annealing in the sense that it responds to large energy differences earlier than small ones. Nondeterminism can be introduced to those sites that stay "unstable"—the sites on or exterior to the border of the clusters—late in the process, since more contextual information is required for them to reach a globally satisfactory agreement. Cohen and Cooper's postprocessing procedure [26] can similarly be incorporated.

The highest confidence first estimation can be implemented with a set of cooperative computing units. Consider a *winner-take-all* network where each unit corresponds to a site of the image [39]. Only the units with the smallest stability measures can "fire" at one instant; each unit maintains the knowledge about the neighboring units so that its stability measure can be updated immediately should any neighbor change its state. The parallelism gained, however, is limited due to the sequential firing order.

The strict sequentiality of HCF can be relaxed. One possibility is the use of a global stability threshold. The (negative) threshold values increases in time and is broadcasted to all sites. Any site with a stability less than the threshold is allowed to change its state. The computation stops when the threshold reaches zero. It is not yet clear how this modification would affect the resulting estimates, and how to choose an appropriate schedule to increase the threshold.

It is interesting to note that Koch et al. [40] and Besag [17] have independently observed that better estimates result from using a sequence of weaker fields on previous cycles. In the case of surface reconstruction, Koch et al. strongly penalize the formation of lines at the beginning, and slowly decrease the penalty as the computation proceeds. Thus lines are formed only at very steep disparity gradients early in the process, and surface can break at smaller depth disparity gradients by paying a smaller price later. Similarly, Besag decreases the contribution from the prior field for the task of image restoration after each iteration of ICM. Their ideas are related to HCF in the sense they all try to avoid committing too early based on the unreliable initial estimate. However, the explicit uses of the uncommitted state and the stability measures by HCF have the advantages of efficient computation (least commitment), robust results (least commitment and graceful degradation), and easy implementation (no need to choose a proper schedule).

5 Probabilistic Boundary Detection

The important and frequently studied problem of intensity edge detection serves as the first illustrative application of our method. The labeling problem in this context is to assign to each site a label from the set {EDGE, NON-EDGE}, based on discrete intensity measures on the pixels of a square lattice-structured image.

An alternative to local edge detection is to detect discontinuities through the process of reconstructing global intensity functions. A priori knowledge is encoded via global energy functions. Edges are identified at the locations where the connections between pixels should be broken in order to reduce the energy measure related to the intensity configuration. Such schemes have demonstrated superior results in both robustness and localization [41]. However, the computation usually involves optimization of nonconvex functions with large state spaces. Since the reconstructed intensity data are, at least for now, of little use for higher-level tasks, the cost-benefit ratio of the method seems too high for our needs.

The probabilistic framework for edge detection has the advantages of the local and global schemes but few of the associated disadvantages. It uses the outputs of a set of local operators that relate the intensity observations to edge labels, and ignores the intensity values afterwards. Thus the global optimization is performed over a space much smaller than the ones for full reconstruction. On the other hand, the results are robust against local noise and consistent with spatial priors due to the use of a global energy measure. In addition, the probabilistic approach makes it easy to incorporate other image clues.

5.1 Local Edge Models

The edge sites are considered to be situated on the boundary between two pixels (see figure 3). We adopt a step-edge with white Gaussian noise model to compute the local likelihoods of a site s being *EDGE* or *NON-EDGE*—$P(O_s|\omega_s = EDGE)$ and $P(O_s|\omega_s = NON\text{-}EDGE)$. We choose to use a 1×4 or 4×1 window of brightness observations surrounding s as O_s, the observation of the site s. This window of intensity values is assumed to be a realization of one of the possible events depicted in figure 4, corrupted by independent Gaussian noise. Figure 4(a) shows the event E_1 of an edge occurring at the center of the window. Figure 4(b) indicates that the window corresponds to a uniform region (E_2), and 4(c) depicts the events (E_3 and E_4) that the window consists of two regions, however the boundary between the two regions is one pixel off (right or left) the center of the window. The events of 4(b) and 4(c) define the NON-EDGE event.

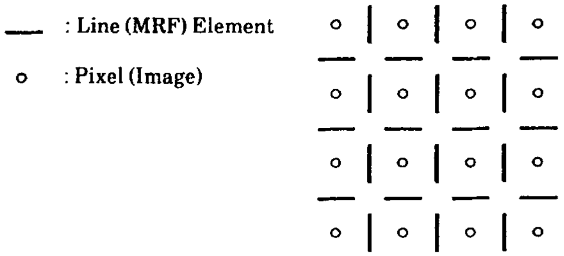


Fig. 3. Pixels and edges: Relationship between MRF (edge) sites and pixels.

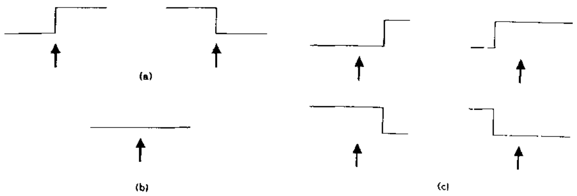


Fig. 4. Step edge model: Image events in a 4×1 window. (a) Edge occurring at center of window. (b) Homogeneous region: no edge occurs. (c) No edge at center, offset edge occurs (arrow indicates center of window).

There is no doubt that the above edge model can greatly be improved in many directions. The obvious ones include the modeling of different types of edges such as roof, line, and peaks, the use of more information by employing circular windows, and the modeling of image blurs. Since one primary goal of this work is to study the robustness of MRF modeling and HCF estimation in the presence of errors in sensory modeling, we feel that the above edge model serves our intention well, and further improvements can only result in better estimates. The computation of the likelihoods given a window of intensity observations and the above edge model is described next. It is based on the work of Sher [42]. The reader is referred to [14] for a complete treatment of probabilistic local edge detection.

5.2 Computing Edge Likelihoods

The computation of the likelihood of E_2 is relatively straightforward. Let W denote the vector of window observations (w_1, w_2, w_3, w_4), where the index indicates the spatial order of the pixels in the window. Since the noise is considered to be independent Gaussian additive, with zero mean and variance σ^2,

$$P(W|E_2) = \sum_{r=0}^{255} P(r)\ P(W|rT_0)$$

$$= \frac{1}{(2\pi\sigma^2)^2} \sum_{r=0}^{255} P(r)e^{-\frac{1}{2\sigma^2}(W-rT_0)(W-rT_0)^t}$$

where T_0 is vector consists of all 1's, and $P(r)$ is the prior probability of a pixel having intensity r. In our experiments, we assume that all intensity levels are equally likely, therefore $P(r) = 1/256\ \forall r$. More complex distributions may easily be incorporated. One possibility is to use a normalized intensity histogram of the input image as the prior distribution.

For the cases of edge occurrence, let r_1 and r_2 be the assumed intensities of the two regions on the left and right of the edge respectively. The window of assumed intensities can be represented as the vector $T = r_1T_1 + r_2T_2$, where T_1 and T_2 are the template vectors for the left and right regions. For example, for the region on the left of three-pixel wide (E_3), $T_1 = (1, 1, 1, 0)$, and the corresponding $T_2 = (0, 0, 0, 1)$. The likelihood of such cases (an edge present at a particular location in the window) can be computed by

$$\sum_{r_1,r_2} P(r_1, r_2)\ P(W|T)$$

$$= \frac{1}{(2\pi\sigma^2)^2} \sum_{r_1,r_2} P(r_1, r_2)e^{-\frac{1}{2\sigma^2}(W-r_1T_1-r_2\mathrm{T}_2)(W-r_1T_1-r_2T_2)^t}$$

Again we assume r_1 and r_2 are independently and uniformly distributed in the experiments. The *EDGE* event E_1 corresponds to the case with the pair of templates (1, 1, 0, 0) and (0, 0, 1, 1). The likelihood of *NON-EDGE* is

$$\begin{aligned} P(O|\textit{NON-EDGE}) &= P(O|E_2 \vee E_3 \vee E_4) \\ &= P(O|E_2)\ P(E_2|\textit{NON-EDGE}) \\ &\quad + P(O|E_3)\ P(E_3|\textit{NON-EDGE}) \\ &\quad + P(O|E_4)\ P(E_4|\textit{NON-EDGE}) \end{aligned}$$

We set

$$\begin{aligned} P(E_2|\textit{NON-EDGE}) &= 0.8 \quad \text{and} \\ P(E_3|\textit{NON-EDGE}) &= P(E_4|\textit{NON-EDGE}) = 0.1 \end{aligned}$$

in our experiments.

The calculation of edge likelihoods is efficiently carried out with a set of local convolutions followed by a table-lookup operation [42].

Based on the fact that scaling $P(O_s|l)$ for every $l \in L$ by a constant factor for fixed s in equation (13) does not change the a posteriori distribution, we can use the log likelihood ratios—

$$\log P(O_s|\omega_s = \textit{EDGE}) - \log P(O_s|\omega_s = \textit{NON-EDGE})$$

as the only input data, thus simplifying the computation of the stability measures. A binary configuration may be constructed by declaring an edge present only where this log likelihood ratio is greater than log $P(\textit{NON-EDGE})/P(\textit{EDGE})$, the logarithm of prior (local) odds. Call this binary configuration the *threshold log likelihood ratio* (TLR) configuration. This configuration can be considered as a MAP estimate obtained without using contextual information, because

$$\frac{P(\textit{EDGE}|O)}{P(\textit{NON-EDGE}|O)} = \frac{P(\textit{EDGE})}{P(\textit{NON-EDGE})} \frac{P(O|\textit{EDGE})}{P(O|\textit{NON-EDGE})}$$

therefore,

$$P(\textit{EDGE}|O) > P(\textit{NON-EDGE}|O)$$

$$\Leftrightarrow \log \frac{P(O|\textit{EDGE})}{P(O|\textit{NON-EDGE})} - \log \frac{P(\textit{NON-EDGE})}{P(\textit{EDGE})} > 0$$

In our experiments, we use TLRs as the initial estimates whenever possible.

5.3 Markov Random Field of Line Process

The MRF model used is similar to the "Line Process" MRF used by both Geman and Geman [24] and Marroquin et al. [23]. Each edge site is modeled as a random variable of the field. The field is binary, with $2(N^2 - N)$ entities where the image is a $N \times N$ rectangular pixel array. Notice one major difference between our setup and the existing MRF segmentation work: the line process of the latter is implicit. There are no external observations directly associated with it, and the formation of the lines depends on the configurations of a coupled MRF of the intensity process. In our setup, the intensity values are used only to calculate the local likelihoods for the edge sites, and the likelihoods constitute the input of the standalone line process.

5.4 Construction of Potential Functions to Encode Prior Knowledge

The spatial relationships between edge sites we wish to enforce have the following effects: (1) To encourage the growth of continuous line segments; (2) to discourage abrupt breaks in line segments; (3) to discourage close parallel lines (competitions); and (4) to discourage sharp turns in line segments. The potential values we assign to the configurations of the line cliques are shown in figure 5.

The sensitivity of the results obtained to changes in the parameters specifying the potential functions depends upon the parameter in question. Our experience is that changing the potential function associated with the 1-clique had the greatest effect on the final result, followed by the 2-clique and 4-clique potential functions, in that order. In this context it may be worth noting explicitly that the singleton clique controls first-order statistics and the larger cliques higher-order statistics.

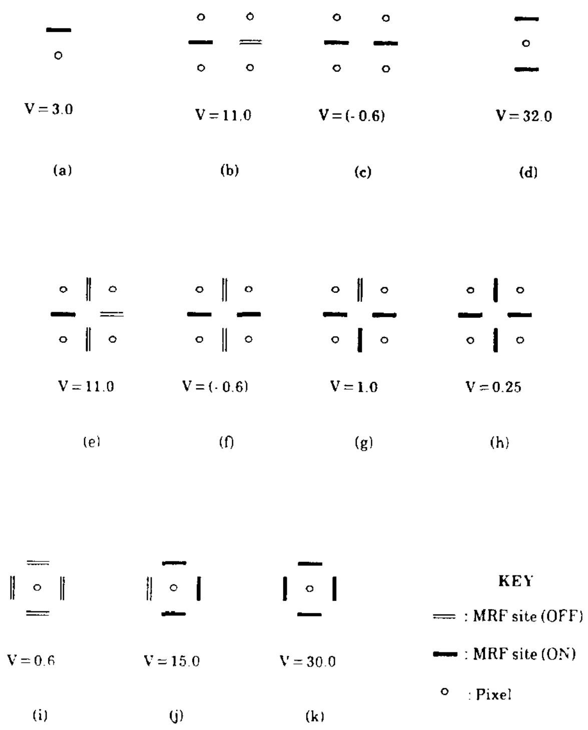


Fig. 5. Potential assignments used in experiments. Configurations not shown have potential 0. Sample semantics: (a) Edge presence. (b) Line termination. (c) Continuous line. (d) Parallel lines.

5.5 Experimental Results

We focus upon algorithms based on MRF modeling, including the proposed highest confidence first (HCF) (section 4), iterative conditional modes estimation (ICM) [17], stochastic MAP (simulated annealing with Gibbs Sampler [24]), and stochastic MPM (Monte Carlo approximation to the MPM estimate [23]). The edge results obtained by applying 3×3 Kirsch operators with nonmaximum suppression are also presented for the sake of completeness of comparisons. The annealing schedule for the stochastic MAP follows the one suggested in [24], i.e., $T_k = c/(\log(1 + k))$ where T_k is the temperature for the kth iteration, with $c = 4.0$. The stochastic MAP was run for 1000 iterations and the stochastic MPM for 500 (300 to reach equilibrium, 200 to collect statistics).

Here we show the results of four sets of experiments (figures 6 through 9). The figures for each set contain the original image, the result from the Kirsch operators (except for the artificially generated (figure 6), the TLR configuration, and the results obtained by using stochastic MAP, stochastic MPM, ICM (scan-line visiting order), ICM (random visiting order), and HCF algorithms. Except in the case of the HCF algorithm, where the MRF is initialized to all null (uncommitted) states, the MRF is initialized to the TLR configuration. The MRF specification is the same throughout.

Since the final estimates of the stochastic MAP, MPM, and the ICM with random visiting order technically depend on the "seeds" for a pseudo random-number generator in addition to the input images, we

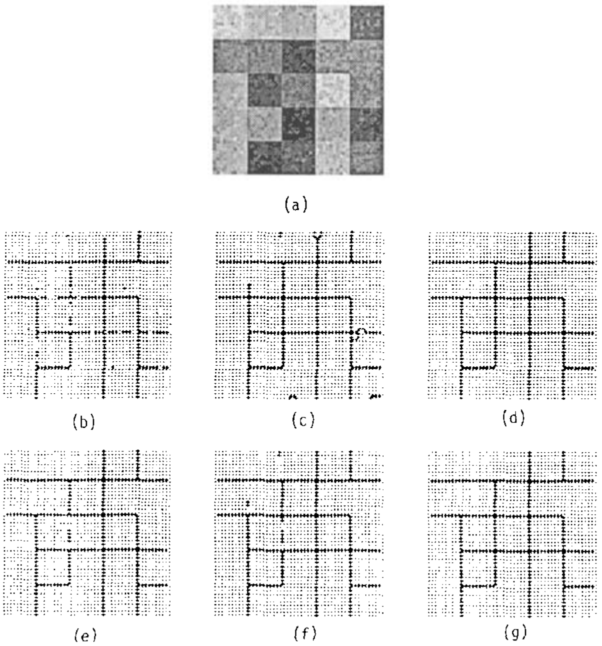


Fig. 6. Boundary detection experiment set (I). (a) Synthetic 50×50 checkerboard image corrupted by independent unbiased Gaussian noise with std. 16. (b) TLR configuration. (c) Stochastic MAP estimate. (d) Stochastic MPM estimate. (e) ICM (scan-line visiting order) estimate. (f) ICM (random visiting order) estimate. (g) HCF result.

have observed large variations among the results in repeated runs using the same setup. Theoretically, this should not be true for the stochastic methods because their results asymptotically converge to the true MAP and MPM values. In practice, the results are highly dependent on the configurations of the early iterations. Once a large-scale region or line segment is formed, it is seldom altered in a limited period of time. We have subjectively chosen to show the most typical results in our figures. There are better and worse results, as one might expect.

The HCF algorithm consistently produces superior results in the experiments. For example, figures 6(g) and 7(h) produce clean, connected edges, and figure 8(h) has the clearest letter outlines and also is alone in detecting the entire bottom edge of the "R" block. The completion of the horizontal line at the top of the "S" block arises from the small edge segment at the extreme right of the line noticeable in the TLR image. Edge propagation occurs in this direction because the likelihoods of the necessary collinear edges are only slightly below the threshold.

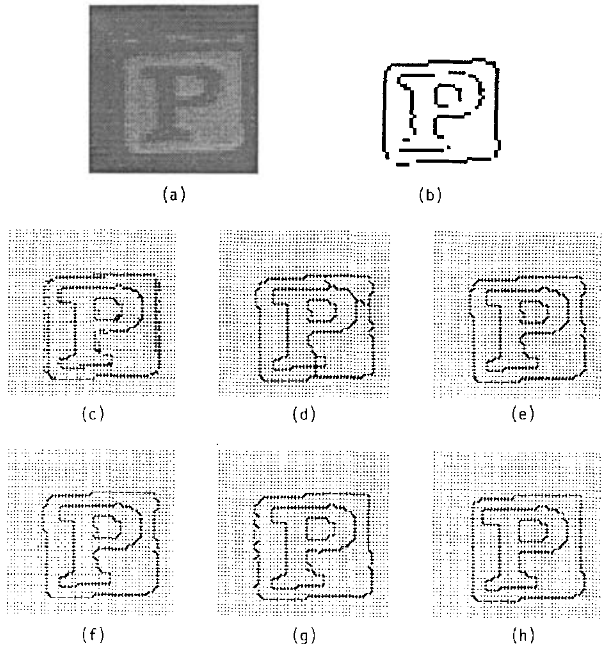


Fig. 7. Boundary detection experiment set (II). (a) Natural 50×50 image of wooden block. (b) Thinned and thresholded output of Kirsch operators. (c) TLR configuration. (d) Stochastic MAP estimate. (e) Stochastic MPM estimate. (f) ICM (scan-line visiting order) estimate. (g) ICM (random visiting order) estimate. (h) HCF result.

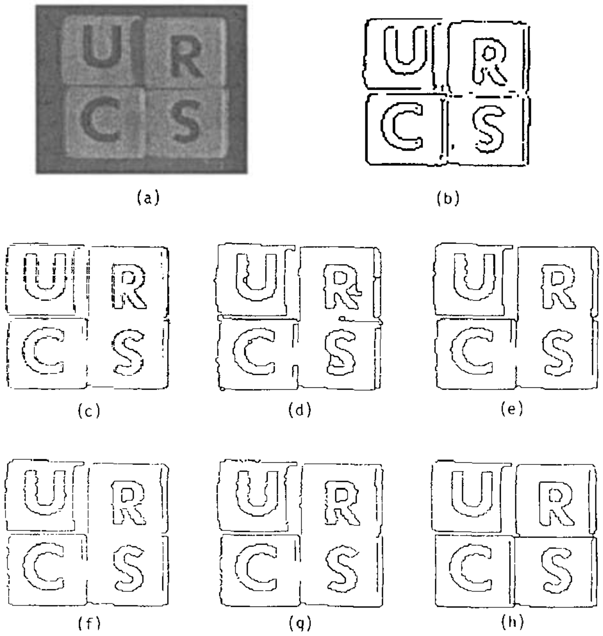


Fig. 8. Boundary detection experiment set (III). (a) Natural 100×124 image of four plastic blocks. (b) Thinned and thresholded output of Kirsch operators. (c) TLR configuration. (d) Stochastic MAP estimate. (e) Stochastic MPM estimate. (f) ICM (scan-line visiting order) estimate. (g) ICM (random visiting order) estimate. (h) HCF result.

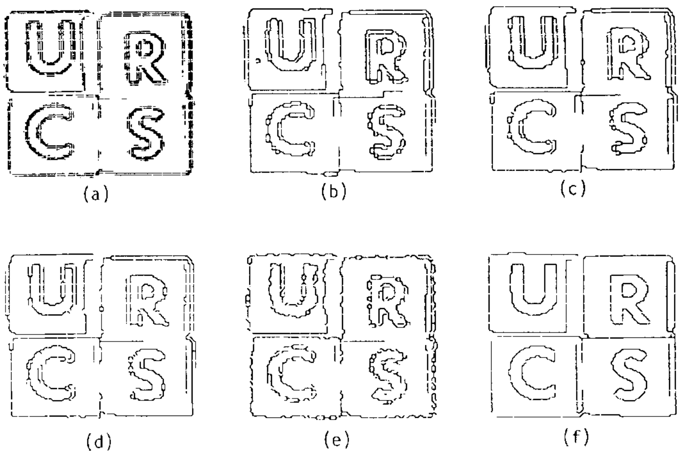


Fig. 9. Boundary detection experiment set (IV). Experiments with incomplete edge model—original image as in figure (a) TLR configuration. (b) Stochastic MAP estimate. (c) Stochastic MPM estimate. (d) ICM (scan-line visiting order) estimate. (e) ICM (random visiting order) estimate. (f) HCF result.

To test the robustness of the algorithms, we conduct further experiments further experiments using a likelihood generator with a less complete edge model. Since offset edges (figure 5(c)) are not considered here, multiple responses become significant as can be seen from the TLR configuration shown in figure 9(a). This change adversely affects the estimates produced by all the algorithms except the HCF, as can be seen from comparing corresponding pictures in figures 8 and 9.

We compare convergence times only for deterministic schemes, since stochastic schemes have no true convergence criterion, but rather need subjective judgments as to when equilibrium has been reached, and as to when we have gathered enough statistics to estimate accurately the joint (or marginal) probabilities. Typically, several hundred iterations are needed. The deterministic algorithms (HCF and ICM (scan-line)) have been timed on images of various sizes using a Sun 3/260 with floating point acceleration. The results are shown in figure 10.

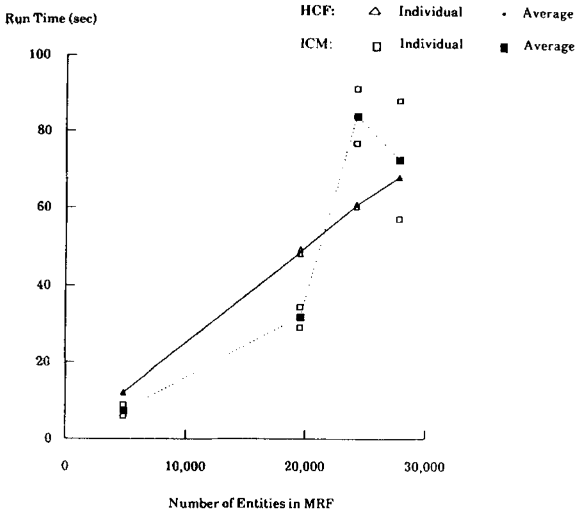


Fig. 10. Timing test results. The HCF and ICM algorithms are earch run on two images of the same size, for four image sizes. Individual and average run-times are shown.

5.6 Analysis of Experimental Results

Further consideration of some of these points may be found in the final discussion section.

Goodness of Estimates

1. The HCF algorithm consistently outperforms all other algorithms, giving superior results both with synthetic and real-image data. The results from this algorithm all fit well with our model of the world, which consists of smoothly continuous boundaries, and are consistent with the observations.
2. The HCF algorithm also appears to be robust, in that it produces an estimate consistent with the observations even when the MRF model used is inadequate, as in the experiment using the less sophisticated edge detector. Since our MRF model does not take into account multiple responses, the MAP criterion may not lead to the "best" results.

In this case, the local minimum found by the HCF algorithm is clearly better than the results produced by other methods as it is based on the strength of external evidence.

3. The ICM algorithm performs inconsistently and its results depend to a large extent upon the initialization of the MRF and the visiting order. It is also not clear which, if any, of the visiting orders studied is better than the other. The scan-line visiting order performs better in some of our experiments, but that is probably due to the horizontal and vertical characteristics of the boundaries. HCF does not rely on any predefined order, and thus is not biased for any boundary shape.
4. The stochastic MAP algorithm with simulated annealing gets stuck in undesirable local minima, suggesting that our annealing schedule might have lowered the temperature too fast. However, an appropriate annealing schedule seems hard to obtain a priori. We have conducted further experiments with simulated annealing with varying annealing constant c for 1000 iterations each. It appears that starting with temperature too high will destroy the TLR initial estimate, resulting in estimates inconsistent with the input data. The Monte Carlo MPM estimates are more reliable than simulated annealing results in most cases. However, we occasionally observed large-scale mistakes.
5. In addition to the qualitative comparisons, we have evaluated the results in terms of energy measures (table 1). Recall that, as discussed in section 4, the energy measures may not reflect the correctness of the estimates in the presence of significant modeling inaccuracy. The comparisons based on these measures serve the purposes of verifying the validity of our potential assignments, and, more importantly, identifying the effectiveness of HCF as an energy minimization strategy for similar applications. It is worth mentioning that in our many trials, HCF consistently found better local minima in all but one case when the Monte Carlo MPM beat HCF by 0.1%.

Table 1. Energy values

Fig.	TLR	MAP	MPM	ICM(s)	ICM(r)	HCF
6	−3952	−4282	−4392	−4364	−4334	−4392
7	−572	−680	−723	−693	−715	−740
8	4785	−349	−503	−503	−513	−629
9	59719	−5303	−5296	−4954	−3728	−9587

Quantitative comparisons between results of simulated annealing (MAP), Monte Carlo (MPM), ICM (scan-line order), ICM (random order), and HCF. Column Fig. lists the figure numbers of the input images. Some of the values (of MAP, MPM, ICMs) are the averages of the results from several runs. (The smaller the energy the better the estimate.)

Convergence Times

1. The HCF algorithm makes a perhaps surprisingly small number of visits before converging. Clearly, due to the initialization, it must visit every site at least once. What is surprising is that it visits each site on the *average* less than 1.01 times before converging. What this implies is that the first decision made by a site is nearly always the best one. Also, the HCF algorithm takes almost the same time on different images of the same size.
2. The convergence times of the ICM algorithms are unpredictable—they vary with visiting order, MRF initialization, and even with the particular image given as input.
3. The time taken by the HCF algorithm includes the time taken to set up the heap initially. This may, in some circumstances, be a little unfair. For instance, if one has to process data online from various information sources (section 2) [29], the heap setting up cost can be treated as a preprocessing cost rather than a run-time one. In theory, the time taken by the HCF algorithm should be given by $c_1N + c_2V \log_2 N$, where c_1 and c_2 are positive constants, N the number of sites to be labeled and V the number of visits. V here is at least N and we conjecture that on the average it is cN for some small $(1 < c < 2)$ constant c. Since the latter term should dominate, one would expect to see a nonlinear curve in a plot of run time vs. number of sites. However, the curve is very nearly a straight line, indicating either that the constant c_2 is very small, or that the changed stability values do not propagate very far up the heap on the average. The former does not appear to be true, as our experiences suggest that the initial heap construction takes far less time than the rest of the algorithm.

6 Segment and Reconstruct Depth Maps by Incorporating Intensity Edge Information with Sparse Depth Measures

The second domain we treat with our method is the reconstruction of three-dimensional scene parameters (intrinsic images) from visual information. Such reconstruction often depends on a smoothness assumption to regularize the computation. The problem is that smoothing is not wanted across object boundaries, and reliable reconstruction cannot be achieved without the detection of the discontinuities [43]. On the other hand, discontinuities are best described as boundaries between surface patches defined by the corresponding scene parameters, and thus cannot be detected directly from sparse, noisy data. The cooperation of reconstruction and discontinuity detection has been of interest for some time: The challenge is to develop a unified treatment for reconstruction and segmentation. The mechanism we use is *coupled MRFs*, in which MRFs, one for the depth process and one for the discontinuity process, work in parallel and interact.

In fusing depth and intensity information, Gamble and Poggio [30] use the intensity edges detected with the Canny operator [44] to constrain the locations of the depth discontinuities while reconstructing a depth map. Their rule is that no depth discontinuity is allowed without a corresponding intensity discontinuity. The results of combining the two information modalities are encouraging and better than either modality operating alone, but the uncompromising relation between depth and intensity discontinuities means that depth discontinuities within regions of little intensity variation will be lost even if the depth information is good. The problem is thus how to assign a general a priori relation between depth and intensity information.

6.1 Coupled Markov Random Fields

Represent a pixel image $S = \{s_1, s_2, \ldots, s_N\}$ as a set of lattice-structured sites, and the discontinuity image, D, as the set of sites placed midway between each vertical and horizontal pair of pixel sites. Let $F = \{f_s, s \in S\}$ be the set of random variables indexed by S, with $f_s \in R$ representing the depth value at location s, and $L = \{l_d, d \in D\}$ be the set of random variables indexed by D, with $l_d \in \{0, 1\}$ representing the absence or presence of a depth discontinuity at site d. F and L correspond to the depth process and the line (depth discontinuity) process respectively of the coupled Markov random fields introduced by Geman and Geman [24] and used in [23, 30], and this paper. A configuration of (F, L) corresponds to an admissible solution to our problem. The observation models we use are the following.

6.2 Observation Models

Early Depth Measurements. The depth measurements are considered sparse and independently measured. Denote by $\hat{S}$ the set of sites in S at which depth measurements are available and $G = \{g_s, s \in \hat{S}\}$ the measurement process. We assume

$$P(G = g | F = f) = \prod_{s \in \hat{S}} P_s(g_s | f_s) \qquad (16)$$

where g_s denotes the measurement at pixel site s, and $\boldsymbol{g}$ represents these measurements. Often the noise can be adequately modeled by unbiased Gaussian distributions. That is,

$$P_s(g_s | f_s) = \frac{1}{Z_s} e^{\frac{-(f_s - g_s)^2}{2\sigma_s^2}} \qquad (17)$$

Discontinuity Observations Based on Intensity. Instead of treating intensity edges as constraints on the locations of depth discontinuities, we consider them as partial evidence supporting or refuting the hypotheses about depth discontinuities. The motivation is simple. The intensity images are the results of many confounding factors—lighting, surface geometry, surface reflectance, and camera characteristics. Intensity discontinuities may reflect sudden changes of depth values, but depth discontinuities do not necessarily imply large intensity variations.

Figure 11 shows the conceptual hierarchy that consists of the interesting events involved here. At the first level, only *EDGE* or *NON-EDGE* is of concern. Node *EDGE* represents the event that the site of interest corresponds to some sort of discontinuity in the world; *NON-EDGE* represents the event that the site is within a homogeneous region. At the next level, whether a particular intensity discontinuity is due to depth discontinuity becomes interesting. With the edge operator of section 5.1, intensity observations yield likelihood ratios for intensity edges and thus provide information about the events in the first level. They say nothing about the events in the second level, which are important to the depth segmentation problem. Our approach is to incorporate prior experience and knowledge (represented

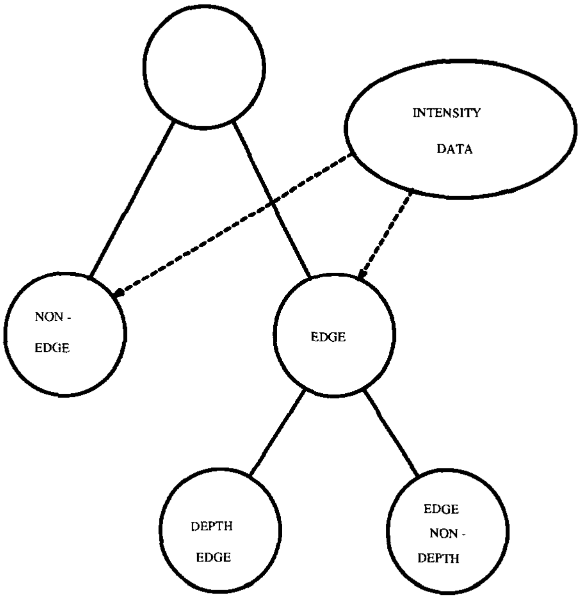


Fig. 11. Relation between intensity and depth edges.

by conditional probabilities) to infer the amount of support (represented as likelihood ratios) that the intensity observations provide for the existence of depth discontinuities. To be more precise, we are interested in computing the likelihood ratio of a site being a *DEPTH-EDGE* given the numbers α_0, α_1, and α_2 constantly proportional to the likelihoods $P(O|\mathit{NON\text{-}EDGE})$, $P(O|\mathit{DEPTH\text{-}EDGE})$, and $P(O|\mathit{EDGE\text{-}NON\text{-}DEPTH})$. This computation is effected by the fast hierarchical evidence-combination algorithm mentioned in (section 2). The α values are maintained in the hierarchical label tree, and are calculated from the outputs of edge likelihood operators acting on intensity data. We also need the conditional probability $p = P(\mathit{NON\text{-}EDGE}|\mathit{NON\text{-}DEPTH\text{-}EDGE})$, where *NON-DEPTH-EDGE* stands for the joint event *NON-EDGE* ∨ *EDGE-NON-DEPTH*, which is determined empirically for a particular domain. We have

$$\frac{P(O|\mathit{DEPTH\text{-}EDGE})}{P(O|\mathit{NON\text{-}DEPTH\text{-}EDGE})} = \frac{\alpha_1}{p\alpha_0 + (1-p)\alpha_2} \tag{18}$$

In the rest of the paper, λ_d denotes the edge likelihood ratio of site d given the intensity observation O_d, where $d \in D$.

$$\lambda_d(l_d) = \frac{P(O_d|l_d)}{P(O_d|\neg l_d)} \tag{19}$$

Again, we consider the spatially distinct intensity observations are conditionally independent:

$$P(O|l) = \prod_{d \in D} P_d(O_d|l_d) \tag{20}$$

where O denotes the collection of intensity observations.

Conditional Independence Between Intensity and Depth Observations. We assume that the depth (g) and intensity (O) observations are only related through the geometry of the surfaces in view. They are conditionally independent in the following sense:

$$P(g, O|f, l) = P(g|f, l)P(O|f, l) \tag{21a}$$

We further assume that the knowledge of depth discontinuities contributes no information to make one prefer the observation of g over others once the true depth values are known:

$$P(g|f, l) = P(g|f) \tag{21b}$$

and that the knowledge of surface depth does not make O more or less likely once the depth discontinuities are known:

$$P(O|f, l) = P(O|l) \tag{21c}$$

The scene depth, in many circumstances, affects the observed intensity values. The assumption (21c) is reasonable, however, since it is the indirect observations of intensity discontinuities but not the magnitude of the intensity that are actually used in this work. Thus in this work we discard intensities after computing likelihoods of discontinuities. An interesting research problem would be to use the intensity information (perhaps through the irradiance-orientation constraint [45]) more directly.

Summarizing (16–21), we assume

$$P(g, O|f, l) = \prod_{s \in S} P_s(g_s|f_s) \prod_{d \in D} P_d(O_d|l_d) \tag{22}$$

6.3 Markov Random Fields and Energy Measures

Within each F and L, spatially adjacent variables tend to have similar values. That is, surfaces and boundaries tend to be continuous and smooth. MRFs corresponding to F and L can be separately defined to model these properties. Section 5 has demonstrated some promising edge detection results using an MRF for the line process alone. The depth and line processes, however, are not independent of each other. The presence of a line at an edge site breaks the connection between the two variables at the adjacent pixel sites; a small change in the values of two adjacent depth variables suggests the absence of a discontinuity in between. This interdependence is the basis for the concept of coupled MRFs—a unified treatment of reconstruction and segmentation. Figure 12 shows a neighborhood system Γ of the MRF consisting of the depth and line processes. In addition to the depth and line processes, the concept of coupled MRFs can also be applied to model many other interdependent processes corresponding to various intrinsic parameters [29].

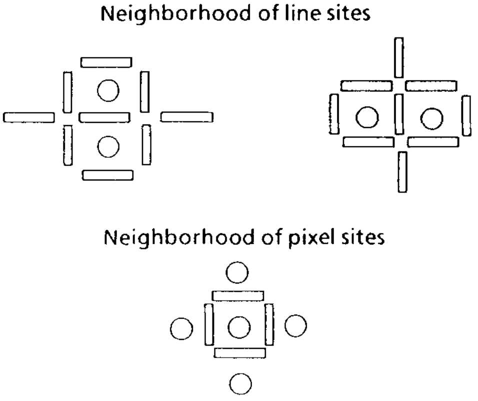


Fig. 12. Neighborhood system for coupled MRFs.

(F, L) is an MRF with respect to a neighborhood system Γ if and only if, according to the Hammersley-Clifford theorem, the joint probability distribution of the variables is a Gibbs distribution. That is,

$$P(f, l) = \frac{1}{Z} e^{\frac{-U(f, l)}{T}} \tag{23a}$$

where the energy functional

$$U(f, l) = \sum_{c \in C} V_c(f, l) \tag{23b}$$

where C is the set of cliques defined by Γ. Continuous surfaces can be modeled by setting the potential energy V for the cliques consisting of two adjacent depth sites, say i and j, and the line site in between them, say ij, proportional to $(1 - l_{ij})(f_i - f_j)^2$ [23, 30]. Using this potential function, minimizing the energy measure has the effect of fitting membrane patches to the lattice. Higher-order spline surfaces can similarly be encoded with larger neighborhood systems to account for higher-order derivatives. Since only depth discontinuities are concerned here, we use the neighborhood system depicted in figure 12 and the above potential function throughout our experiments. Other types of cliques that have nonzero potential functionals used in our experiments consist only of line sites. They are the same as the ones described in section 5.

6.4 A Posteriori Energy

Bayes' rule combines the a priori knowledge and the early visual observations to derive the a posteriori belief.

$$P(f, l|g, O) = \frac{P(f, l)P(g, O|f, l)}{\sum_{f,l} P(f, l)P(g, O|f, l)}$$

Note from (16) and (20) that a constant term may be taken out of the energy function for each site which appears in the constant term of the Gibbs distribution. Thus scaling all of the likelihoods for a fixed site by a constant does not change the posterior distribution of (F, L). From (22) and (23), and assuming (17), the posterior distribution is a Gibbs distribution, with the a posteriori energy functional proportional to

$$\begin{aligned} U(f, l|g, O) = {} & \sum_{c \in C} V_c(f, l) \\ & + T[\sum_{s \in S} \frac{(f_s - g_s)^2}{2\sigma_s^2} \\ & - \sum_{d \in D} \log\lambda_d(l_d)] \end{aligned} \tag{24}$$

6.5 HCF: Coping with Continuous Variables

Let ζ denote the uncommitted state, and $\bar{R} = R \cup \{\zeta\}$, $\bar{L} = \{\zeta, 0, 1\}$ denote the *augmented state spaces* for the depth and line processes respectively. Based on (24), define the *augmented local energy* measures with respect to an augmented configuration (f, l) as

$$E_s(f) = \sum_{c: s \in c} V_c'(f', l) + T\frac{(f_s - g_s)^2}{2\sigma_s^2} \quad \text{for } s \in \hat{S} \tag{25a}$$

$$E_s(f) = \sum_{c:s\in c} V_c'(f', l) \qquad \text{for } s \in S - \hat{S} \qquad (25b)$$

and

$$E_d(l) = \sum_{c:d\in c} V_c'(f, l') - T \log \lambda_d(l) \quad \text{for } d \in D \quad (25c)$$

where (f', l') agree with (f, l) everywhere except $f_s' = f$ and $l_d' = l$, with $(f, l) \in (R, L)$. $V_c' = 0$ if there is an uncommitted site in c, otherwise it is equal to V_c. Thus the cliques containing uncommitted sites have no effect on the augmented energy measures. Since the only cliques involved in (25a) and (25b) are those consisting of two neighboring pixel sites and a line site in between, the terms $\sum_{c:s\in c} V_c'(f', l)$ can be written as

$$\sum_{r\in N_s} \beta(1 - l_{rs})\,(f_r - f_s)^2$$

where N_s is the pixel neighborhood of s. The augmented local energy measures for pixel sites thus are quadratic; the shape of each quadratic depends on the constant parameter β, the number of active neighbors, and the variances of the noise in the early measurements. The temperature T is set to 1 throughout our experiments, therefore β decides the degree of smoothness relative to the magnitude of noise.

6.5.1 Stability Measures. The confidence of a site in a configuration (f, l) is evaluated in terms of the following *stability* measures:

$$\begin{aligned} G_s(f, l) &= \Delta E_s(f_{\min}, f_{\min} + \alpha) && \text{if } f_s = \zeta \quad (26a)\\ &= \Delta E_s(f_{\min}, f_s) && \text{otherwise} \end{aligned}$$

where $E_s(f_{\min}) = \min_{f\in R} E_s(f)$, and

$$\begin{aligned} G_d(f, l) &= \max_{l\in L, l\neq l_{\min}} \Delta E_d(l_{\min}, l) && \text{if } l_d = \zeta \quad (26b)\\ &= \Delta E_d(l_{\min}, l_d) && \text{otherwise} \end{aligned}$$

where $E_d(l_{\min}) = \min_{l\in L} E_d(l)$. The term $\Delta E_r(k, j)$ is *defined as* $E_r(k) - E_r(j)$ with respect to (f, l). It represents the change in local energy measure of r, thus the global energy (e.g., (24)), if r should switch its state from j to k. The stability of a site is nonnegative only when it is in its minimal energy state (i.e., $l_{\min}$ or $f_{\min}$) with respect to its current local energy measure. A large negative stability value signals high confidence in making a state change to the minimal energy state. The constant α determines the stability of uncommitted pixel sites: it gives the price paid in energy for remaining uncommitted. α has the semantics of offset along R from state of minimal energy. Using it, the stability measure for uncommitted states has the semantics "how much energy could be lost by committing." Large α encourages quicker commitment.

6.5.2 Convergence Properties. The HCF network behaves as follows. Every site starts in the uncommitted state. At any instant, only the sites with the highest confidence in changing their states, i.e., the least stable ones with respect to the current configuration, are allowed to change their states. The identities of the sites, pixel or discontinuity, are ignored in the process of comparing the stability measures. Thus the reconstruction and segmentation processes proceed simultaneously. Eventually, the network settles at a configuration when no further reduction of the global energy measure can be made at each site; i.e., when all local stability measures are nonnegative.

The convergence property can be easily verified. It is possible, however, that the final configuration contains uncommitted depth sites, due to the sparseness of the depth data. In the initial stage of the computation, $E_s(f) \equiv 0$ if $s \in S - \hat{S}$. This local energy measure, and thus the stability measure, will remain zero until one of the pixel neighbors becomes committed and the line process indicates there is no discontinuity in between. If a region of pixel sites, consisting only of members of $S - \hat{S}$, was surrounded with discontinuities before any of them has nonzero stability measure, all of them will stay uncommitted. In the extreme case, if there is no depth measurement at all, this network will not produce an estimate of the depth map. This is an advantageous feature since in such degenerate cases, there are infinite number of configurations that have the minimal energy measure. It is important for the low-level process to indicate the lack of information to higher-level processes so that attention can be directed to acquire more information. This feature can be turned off, if desired, by assigning a priori estimates (e.g., expected range of the scene) to those sites.

Again we implemented the HCF network on a serial machine using a binary heap to decide the visiting order of the sites. The number of comparisons in maintaining the heap property for each change is limited by the height H of the heap ($H \approx \log_2 (3N)$ where N is the number of pixels). Ideally, the computation terminates when the top element has nonnegative stability since no more energy reduction would be possible afterwards. In practice, a small (negative) threshold is used to force termination without noticeable degradation of depth value (see below). Some threshold is necessary in any case because of limited precision in the calculations.

6.6 Experiments and Results

6.6.1 Synthetic Scenes. The enhanced HCF algorithm, which reconstructs depth and finds depth discontinuities from a pair of depth and intensity images, is first demonstrated on a synthetic scene. The scene consists of a range image and an irradiance image. Noise of a particular description is added independently to each image. Figures 13(a) and (b) show original full-resolution intensity and depth images, and figures 13(c) and (d) show the images with added noise. The intensity image has a range of pixel values in [0, 255], and is perturbed by zero-mean, signal-independent Gaussian-distributed noise $G(0, \sigma)$ with $\sigma = 16$. Perturbed values less than 0 are set to 0, greater than 255 are set to 255. The range image has an unusual noise pattern. Part of the motivation is to test the effects of spatially nonuniform noise (the noise model affects the stability calculations of elements). Another motivation is to reflect a range imaging system whose values are more accurate near its optic axis, perhaps as an effect of reduced resolution (averaging) in the periphery. The noise distribution for the synthetic range image is radially symmetric around the center of the image, with standard deviation of the additive, signal-independent, zero-mean Gaussian noise at a point increasing exponentially as the distance of the point from the center of the image. The exponential is scaled so the maximum noise has $\sigma = 20$ (in the corners of the image.)

Figure 14 illustrates the effect of sparse depth data on the final depth reconstruction and segmentation, and also shows the beneficial effects of incorporating intensity information. In these experiments α and β are held constant. We assume that 95% of *NON-DEPTH-EDGE* events are *NON-EDGE*. The range image is sampled randomly at reduced resolution. Figures 14(a) and (b) show the reconstruction and segmentation using just 50% of the original full-resolution depth data. Figures 14(c) and (d) show the improvement gained by allowing the MRF access to the irradiance image as well.

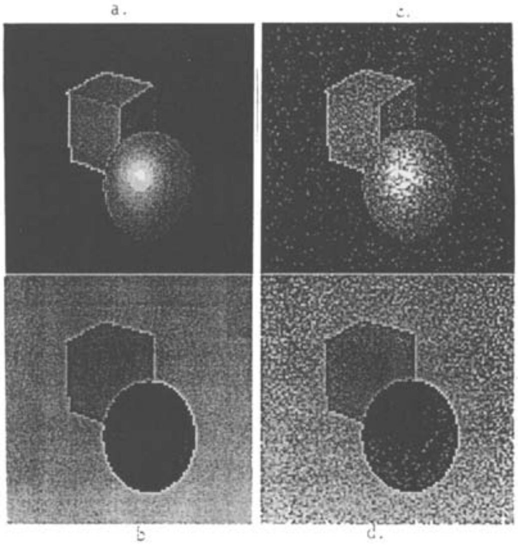


Fig. 13. Synthetic intensity and range data. (a) Original intensity image. (b) Original depth image. (c) Intensity image with $G(0, 20)$ additive noise, clipped to range [0, 255]. (d) Depth image with spatially varying Gaussian noise, maximum standard deviation of 20 (see text).

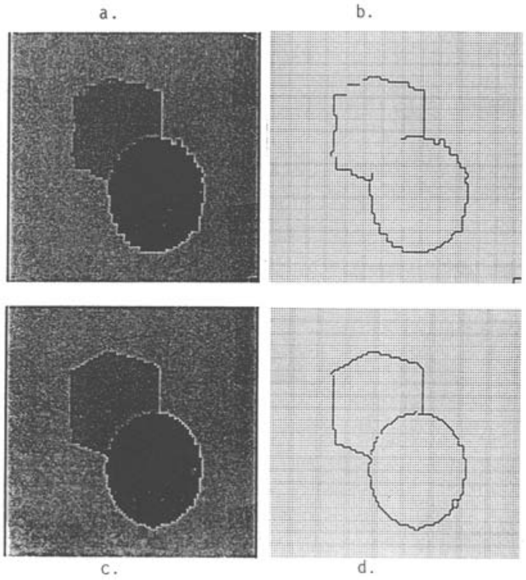


Fig. 14. Results with synthetic data. (a) Reconstructed depth with $\alpha = 50$ and $\beta = 0.001$, with 50% depth data randomly sampled and no intensity input. (b) Depth edges with conditions of (a). (c) Reconstructed depth with α and β and 50% depth sampling as in (a), but using intensity input in MRF. (d) Depth edges with conditions of (c).

6.6.2 Natural Scenes. Depth segmentation and reconstruction using HCF is also demonstrated with stereo disparity data of a scene consisting of table-top objects. The Cooper stereo algorithm [46] yields sparse disparity data associated with intensity contours in the input images. It has been demonstrated to be robust and to work with natural scenes and with structured light. Briefly, it uses a global goodness measure to decide the stereo correspondences between zero-crossing contours of Difference of Gaussian (DOG) using a dynamic programming procedure.

Figure 15(a) shows a beach ball and a rectangular box sitting in front of a cylindrical object, with a flat background. Vertical (with respect to the epipolar line) light strips were projected onto the scene to create artificial texture needed by the stereo system (figure 15(b)). The disparity observations are scaled and rounded to 256 levels, and are assumed to have independent unbiased Gaussian noise with standard deviation of 12. Three pairs of stereo images were taken under different setups of the light projector to increase the density of disparity observations (figure 15(c)). Observations at the same pixel location are combined using a conditional-independence assumption:

$$P(d_1, d_2|d) = P(d_1|d)\, P(d_2|d)$$

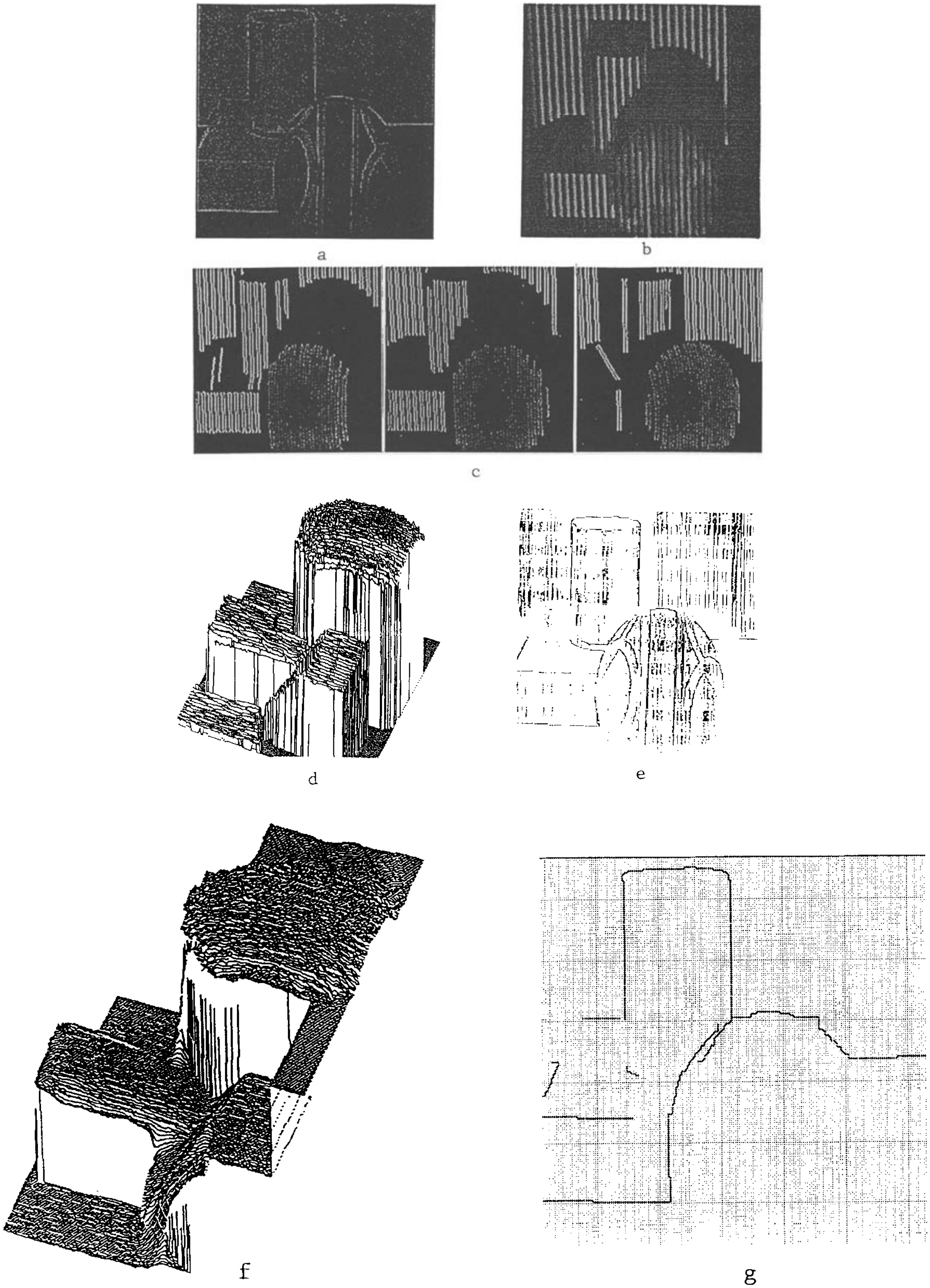
where d_1 and d_2 are two disparity observations obtained from two different setups of the structured light. Figure 15(d) shows the resulting disparity observations in perspective (0 represents no observation). Note that only 35% of the pixel locations have at least one disparity observation. Figure 15(e) shows a map of those locations (in black) overlaid with the TLR estimate of the intensity discontinuities (section 5).

A perspective view of the reconstructed disparity map along with the discontinuities detected, with $\alpha = 30$, $\beta = 0.003$, and $P(NON\text{-}EDGE|NON\text{-}DEPTH\text{-}EDGE) = 0.9$, is shown in figures 15(f) and (g). The surfaces corresponding to the sphere, cylinder, and the background planes are smoothly reconstructed, and significant disparity discontinuities are detected. There are regions that have no (e.g., the table top) or few (e.g., the top of the box) disparity observations. The result is the *leaking* effect: The disparity values of the neighboring regions leak through the holes of weak intensity gradients, resulting in under-segmentation and erroneous disparity estimates. A possible fix is to identify those regions prior to the reconstruction, possibly by building convex hulls of the "adjacent" pixels with disparity measures, and limit the reconstruction process outside of those regions.

6.7 Discussion

The Role of Intensity Discontinuities. Successful integration of multi-modal data requires knowledge about the characteristics of scene and the vision modules processing the data. Such knowledge affects the decisions that have to be made when different modalities provide conflicting information about particular events. Figure 13 shows an example: The strong intensity gradients across the cube edges suggest depth discontinuities at the face intersections but the relatively small depth differences at these locations refute such suggestions. Also, the self-shadowed face merges with the background in the intensity image while the depth information indicates clear separation of the two regions. If reliable depth observations are available, e.g., a noise-free depth map, it is bad practice to use the intensity observations for clues to depth discontinuities. On the other hand, when the depth observations are sparse and unreliable, the correlation of intensity information with depth information should be recognized and used. The probabilistic integration provided by HCF optimization is one coherent framework for such integration tasks. The HCF scheme, as a deterministic method, finds a local probability maximum. In so doing, its behavior is consistent with the natural evidence weighing described above. In particular, if the depth observations are less reliable than the intensity observations (i.e., they have larger variation from their expected true values), the line sites tend to have larger (negative) stability measures than the depth sites at the early stage of the computation. This means the line sites commit (since they are based on intensity information) earlier than the depth sites, resulting in a final configuration that is more consistent with the intensity discontinuities. At the locations where depth observations are missing, the intensity discontinuity information helps to localize the depth discontinuities before the depth information can be spatially propagated.

Fig. 15. Experiments with stereo disparity data. (a) 200×200 intensity image. (b) Scene with projected structured light. (c) Three disparity images. (d) Perspective view of the combined disparity image. (e) Locations of the disparity measurements overlaid with the TLR estimate of the intensity discontinuities. (f) Reconstructed disparity map with $p = 0.9$. (g) Disparity discontinuities.


a
b
c
d
e
f
g

Similarly, when the depth observations are more reliable (denser, less noisy), the depth sites commit earlier. The early depth commitments influence the stability measures, and thus the later commitments, of the line sites. The resulting configuration tends to be more consistent with the depth data.

The Computational Advantages of HCF. The performance of HCF in minimizing the energy of coupled MRFs is consistent with its performance on simple MRFs incorporating only the line process reported earlier (section 5). That is, the enhanced HCF algorithm behaves efficiently and predictably. The introduction of the continuous-valued depth process requires more visits to the depth sites than occurred in the optimization using only the binary line process. The MRF consisting of the line-process only stabilized after fewer than 1.01 visits per site (on the average). Our experiments take on average fewer than 3 visits per site to achieve reasonable estimates in the coupled intensity-depth MRFs. The average number of visits per site is very likely determined by the density of the data, going up as density goes down. The situation is complicated by the fact that the sizes of the regions affect the speed of convergence: Larger regions require on average more visits per site for results to propagate through them.

Since the energy functional is quadratic given a line configuration, in principle any deterministic minimization method would find the same minimal configuration of the depth process. However, there are some advantages to HCF over iterative relaxation schemes with predetermined visiting orders. HCF always visits the site that can reduce the energy measure the most. Thus early visits are far more important than the later ones with HCF. The rate of stabilizing is always maximized, and the most reliable decisions, which reduce energy most, are made first, and at some point the computation may be terminated with confidence of negligible future improvement. Fixed-order schemes cannot guarantee this property.

7 Summary and Discussion

We have given a framework, based on Bayesian-probability theory, for posing and simultaneously attacking the image segmentation and reconstruction problems as a labeling problem. The central issues addressed by the paper are the representation of knowledge, reasoning procedures for combining distinct bodies of knowledge, and inference methods for using available knowledge to infer scene properties.

The central idea of our approach to knowledge representation and reasoning is the decoupling of external evidence and a priori knowledge. A hierarchically structured label tree is used to accrue external evidence concerning the labels for each site, so that a particular piece of knowledge can be represented at the appropriate level of abstraction. The evidence-combination procedure accumulates evidence in terms of joint likelihood ratios rather than probability distributions. This feature enables the integration of the a priori knowledge, encoded in terms of a joint probability distribution of all sites (MRFs), with the pooled external evidence in a Bayesian formalism.

The basic computation performed in the labeling context is one of statistical estimation. The method developed in this research is neither the traditional maximum a posteriori (MAP) estimation (which is very difficult to compute) nor the traditional maximum likelihood estimation (MLE), but rather a new type of estimation that treats individual variables differently in accordance with the relative significance of the variable observations (section 4). The underlying intuition of the highest confidence first (HCF) estimation is simple: in deciding the identities of the variables, the use of contextual information becomes more important as the external observations become less informative. Traditional estimations treat all of the variables equally, thus their results are more likely to be affected by noise and the inaccuracy of prior models.

Why is the HCF computation effective in energy minimization? Take the boundary detection problem as an example (section 5). Let $X_s = 1$ denote the presence of an edge at site s and $X_s = 0$ otherwise. The admissible solution space of the problem consists of the corners of the hypercube $[0, 1]^N$, where N is number of sites. Many corners (solutions) are locally optimal in energy measure—no adjacent corners (of distant 1 away) have lower energy values. The computation of existing iterative relaxation methods consists of a sequence of steps through adjacent corners. The resulting path depends on a (predetermined) updating order and the energy values associated with the corners. It is easy for such computations to get stuck at minor local optima. To ensure that the global optimum can be reached, immensely more computations are required, as in the case of using stochastic simulated annealing techniques.

Although it does not guarantee global optimality, HCF effectively avoids minor local optima by a "greedy search" in an augmented space. The augmented space consists of 3^N elements, including the 2^N hypercube corners in the admissible solution space. The elements in the augmented space form $N + 1$ layers. Layer i consists of the elements with exactly i dimensions of values 0 or 1 ("committed"). The elements are conceptually connected in the sense that the computation starts at the 0th layer (all "uncommitted"), and terminates at some element in the Nth layer—the admissible solution space. The connections between layers are unidirectional; the computation goes through the layers one by one, from lower-numbered ones to higher-numbered ones. There are also some intralayer connections that allow the computation to fine tune its directions. The computation is reminiscent of traditional greedy search or steepest descent techniques in that every step is a move to the "best" adjacent element with respect to an augmented energy measure defined over the augmented space. The greedy search in the augmented space has the flavor of "flying over" minor local optima in the "best" directions observed in the air. It is also interesting to note that the use of the augmented space does not impose any significant amount of overhead. In fact, our preliminary implementation of HCF has consistently demonstrated fast and predictable results.

We plan to continue to explore the properties of the HCF algorithm and MRF modeling. We are interested in seeing this work applied to other domains. Large-scale optimization problems for which HCF could be applied are becoming common in many areas, from signal and speech processing through robotics. We believe that problems involving large numbers of variables, reasoning with multiple imperfect knowledge sources, statistical estimation, and energy minimization computations can all benefit from the approach presented in this work.

Acknowledgments

We thank Dana Ballard, Jerry Feldman, Henry Kyburg, Tomaso Poggio, Max Mintz, Mike Swain, and the referees for their valuable suggestions. Dave Sher provided edge-detection software for our experiments. Paul Cooper provided his stereo system and proofread this document. Rajeev Raman, Lata Narayanan, and Myra VanInwegen also contributed to the experiments.

This work was supported in part by U.S. Army Engineering Topographic Laboratories research contract no. DACA76-85-C-0001, in part by NSF Coordinated Experimental Research grant no. DCR-8320136, and in part by ONR/DARPA research contract no. N00014-82-K-0193. We thank the Xerox Corporation University Grants Program for providing equipment used in the preparation of this paper.

References

1. J. Pearl, "On evidential reasoning in a hierarchy of hypotheses," *Artificial Intelligence* 28 (1):9–15, 1986.
2. G. Reynolds, D. Strahman, N. Lehrer, and L. Kitchen, "Plausible reasoning and the theory of evidence," COINS Technical Report 86-11, April 1986.
3. A. Rosenfeld, R. Hummel, and S. Zucker, "Scene labeling by relaxation operations," *IEEE Trans. Syst., Man, Cybern.* 6:420, 1976.
4. S. Peleg, "A new probabilistic relaxation scheme," *IEEE Trans. PAMI* 2:362, 1980.
5. L.S. Davis and A. Rosenfeld, "Cooperating processes for low-level vision: A survey," *Artificial Intelligence* 17:245-263, 1981.
6. J. Kittler and J. Foglein, "On compatibility and support functions in probabilistic relaxation," *Comput. Vision, Graphics, Image Process.* 34:257-267, 1986.
7. R.M. Haralick, "An interpretation for probabilistic relaxation," *Comput. Vision, Graphics, Image Process.* 22:388–395, 1983.
8. J. Gordon and E.H. Shortliffe, "A method of managing evidential reasoning in a hierarchical hypothesis space," *Artificial Intelligence* 26:323–357, 1985.
9. P.B. Chou, "The theory and practice of Bayesian image labeling," Tech. Reprt. 258, Computer Science Department, University of Rochester, August 1988.
10. P.B. Chou and C.M. Brown, "Probabilistic information fusion for multi-modal image segmentation," *Proc. 10th Intern. Joint Conf. Artif. Intell.*, Milan, Italy, August 1987.
11. I.J. Good, *Probability and the Weighing of Evidence*, Hafner Publishing: New York, 1950.
12. R.O. Duda, P.E. Hart and N.J. Nilsson, "Subjective Bayesian methods for rule-based inference systems," SRI Tech. Note 124, SRI International, 1976.
13. R.C. Bolles, "Verification vision for programmable assembly," *Proc. 5th Intern. Joint Conf. Artif. Intell.*, Cambridge, MA, pp. 569–575, August 1977.
14. D.B. Sher, "A probabilistic approach to low-level vision," Tech. Reprt. 232, University of Rochester, October 1987.
15. R.M. Bolle and D.B. Cooper, "Bayesian recognition of local 3-D shape by approximating image intensity functions with quadric polynomials," *IEEE Trans. PAMI* 6 (4):418–429, 1984.

16. R.M. Bolle and D.B. Cooper, "Optimal statistical techniques for combining pieces of information applied to 3-D complex object position estimation." In *Pattern Recognition in Practice*, E.S. Gelsema and L.N. Kanal (eds.), Elsevier Science Publishers: North Holland, pp. 243-253, 1986.
17. J. Besag, "On the statistical analysis of dirty pictures," *J. Roy. Statis. Soc.* B48(3), 1986.
18. H. Derin and W.S. Cole, "Segmentation of textured images using Gibbs random fields," *Comput. Vision, Graphics, Image Process.* 35:72–98, 1986.
19. J.L. Marroquin, *Probabilistic Solution of Inverse Problems*, MIT Artificial Intelligence Laboratory, September 1985.
20. J. Besag, "Spatial interaction and the statistical analysis of lattice systems (with discussion)," *J. Roy. Statis. Soc.*, ser. B 36:192-326, 1974.
21. M. Hassner and J. Slansky, "The use of Markov random fields as models of texture." In *Image Modeling*, A. Rosenfeld (ed.), Academic Press: San Diego, CA, pp. 185-198, 1980.
22. G.R. Cross and A.K. Jain, "Markov random field texture models," *IEEE Trans. PAMI* 5(1):25-39, 1983.
23. J. Marroquin, S. Mitter, and T. Poggio, "Probabilistic solution of ill-posed problems in computational vision," *Proc. Image Understanding Workshop*, pp. 293-309, December 1985.
24. S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. PAMI* 6(6):721-741, 1984.
25. D.W. Murray and B.F. Buxton, "Scene segmentation from visual motion using global optimization," *IEEE Trans. PAMI* 9(2):220-228, 1987.
26. F.S. Cohen and D.B. Cooper, "Simple parallel hierarchical and relaxation algorithms for segmenting noncausal Markovian random fields," *IEEE Trans. PAMI* 9(2):195-219, 1987.
27. R. Szeliski, "Regularization uses fractal priors," *Proc. AAAI 87*, July 1987.
28. S. Geman and C. Graffigne, "Markov random field image models and their applications to computer vision," *Proc. Intern. Cong. Mathematicians*, Berkeley, CA, pp. 1496-1517, 1986.
29. T. Poggio, "Integrating vision modules with coupled MRFs," Working Paper No. 285, MIT A.I. Lab., December 1985.
30. E. Gamble and T. Poggio, "Visual integration and detection of discontinuities: The key role of intensity edges," MIT A.I. Memo No. 970, October 1987.
31. R. Kindermann and J.L. Snell, "Markov random fields and their applications." In *Contemporary Mathematics*, vol. 1, American Mathematical Society, 1980.
32. G.E. Hinton and T.J. Sejnowski, "Optimal perceptual inference," *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, Washington, D.C., 1983.
33. J. Besag, "Statistical analysis of non-lattice data," *The Statistician* 24:179-195, 1975.
34. H. Elliott and H. Derin, "Modeling and segmentation of noisy and textured images using Gibbs random fields," #ECE-UMASS-SE84-1, University of Massachusetts, 1984.
35. J.A. Feldman and Y. Yakimovsky, "Decision theory and artificial intelligence: I.A Semantics-Based Region Analyzer," *Artificial Intelligence* 5:349-371, 1974.
36. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, "Equations of state calculations by fast computing machines," *J. Chem. Physics* 21, 1953.
37. J.J. Hopfield and D.W. Tank, " "Neural" computation of decisions in optimization problems," *Biological Cybernetics* 52, 1985.
38. D. Marr, *VISION*, W.H. Freeman: New York, 1982.
39. J.A. Feldman and D.H. Ballard, "Computing with connections," Tech. Rept. 72, Computer Science Department, University of Rochester, 1981.
40. C. Koch, J. Marroquin, and A. Yuille, "Analog 'Neuronal' Networks in Early Vision," *Proc. Nat. Acad. Sci. USA* 83:4263–4267, 1986.
41. A. Blake and A. Zisserman, *Visual Reconstruction*, MIT Press: Cambridge, MA, 1987.
42. D.B. Sher, "Advanced likelihood generators for boundary detection," TR197, University of Rochester, Computer Science Dept., January 1987.
43. B.H. Stuth, D.H. Ballard and C.M. Brown, "Boundary conditions in multiple intrinsic images," *Proc. 8th Intern. Joint Conf. Artif. Intell.* Karlsruhe, pp. 1068-1072, 1983.
44. J.F. Canny, "Finding edges and lines in images," AI-Tech. Rept. 720, MIT Artificial Intelligence Laboratory, 1983.
45. B.K.P. Horn, "Obtaining shape from shading information." In *The Psychology of Computer Vision*, P.H. Winston (ed.). McGraw-Hill:New York, pp. 115-155, 1975.
46. P.R. Cooper, "Order and structure in correspondence by dynamic programming," Tech. Rept. 216, University of Rochester, June 1987.