PROBABILISTIC INFORMATION FUSION FOR MULTI-MODAL IMAGE SEGMENTATION

Paul B. Chou and Christopher M. Brown
Computer Science Department
The University of Rochester
Rochester, New York 14627

ABSTRACT

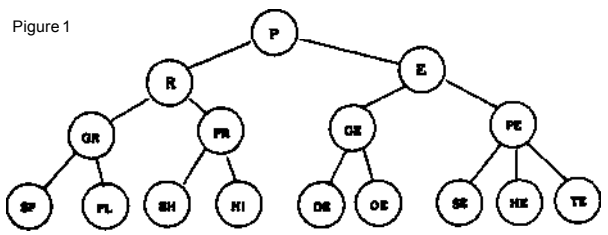
Observable evidence from disparate sources are combined coherently and consistently through a hierarchically structured knowledge tree. Prior knowledge of spatial interactions is modeled with Markov Random Fields. *A posteriori* probabilities of segmentations are maintained incrementally. This paper is a shortened version of [Chou and Brown 87], which contains many more references.*

1. A Probabilistic View of the Segmentation Problem

Represent an image as a set of primitive elements $S = \{s_1, s_2, \ldots, s_N\}$. A segmentation w of the image with respect to a label set $L = \{l_1, l_2, \ldots, l_Q\}$ is a mapping from S to L. Let $\omega_i = \omega(s_i) \in L$ represent the label attached to a, in segmentation a. Let Ω be the set of all segmentations. The image segmentation problem with respect to L can be loosely described as to find the $\omega \in \Omega$ that "best fits" the information collected subject to the limitation of the computational resources. This section describes the uses of probability as the representation for various kinds of information and the corresponding criteria for finding the "best fit"

It is frequently desirable to organise segmentation labels as a hierarchical tree (Figure 1). Each internal node in a tree of labels represents the disjunction of its sons. Each cross-section is a mutually exclusive and exhaustive label set; i.e., a segmentation problem can be denned with respect to a cross-section in a label tree. Using such a tree, we can represent a particular piece of knowledge about the labels at whatever level of abstraction that is appropriate. We use L to denote a set of mutually exclusive and exhaustive set of labels in a label tree H.

Pigure 1



An example of label trees
P - primitive element, 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.

♦This work was supported by the Air Force Systems Command, Rome Air Development Center, Griffiss Air Force Base, New York 13441-5700, and the Air Force Office of Scientific Research, Bolting AFB, DC 20332 under Contract No. F30602-85-C-0008. This contract supports the Northeast Artificial Intel* ligence Consortium (NAIC). This work was also supported by the U.8. Army Engineering Topographic Laboratories under Contract No. DACA76-85-C-0O01.

A visual module could provide opinions on any mutually exclusive set of labels in a hierarchical knowledge tree. We develop an evidence aggregation method that combines consistently and coherently the opinions of the visual modules on a label tree. This method, based on the reasoning proposed in [Pearl 86], follows the Bayesian formalism. It requires only trivial computations. With this method, we are able to design individual experts for a subset of labels of the tree without having to know about the rest of the world. The combined opinion can thus be fused with the image knowledge represented by the *a priori* probabilistic distributions.

1.1. Global Prior Knowledge

Let $X = \{X_s, s \in S\}$ be a set of random variables indexed by S, with $X_s \in L$ for all s. A segmentation can be considered a realization, or a *configuration,* of this random field and Ω can be considered the configuration space of X. Ideally the prior knowledge about Ω can be represented by a probability distribution over Ω. In practice this distribution is either unobtainable or unmanageable due to the immense size (Q^N) of the sample space. In many image understanding applications, however, some restricted classes of distributions can model the image adequately due to the local behavior of the image phenomena. In this paper, we will exclusively use Markov Random Fields (MRFs) as the *a priori* models for Ω, but the work illustrated here can be extended to other image models as well. Section 2 will discuss the MRF model in detail.

1.2. Local Visual Observations

In our treatment, the *opinions* of independent vision modules are expressed as likelihood ratios. For example, module A is an expert on the label set $L_A = \{l_i, l_j\} \subseteq L$. After observing O_s, the module reports one likelihood ratio for each label in L_A. 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 /, is:

$$\lambda_i^A = \frac{P(O_s|l_i)}{P(O_s|\neg(\bigcup_{l \in L_A} l))} \tag{1.0}$$

The methods for designing such modules are well known. Interested readers can consult [Bolles 77] and [Sher 87].

For a purpose that will soon become clear, we impose the following assumption of conditional independence between spatially distinct observations:

$$P(O^A|\omega) = \prod_{s \in S} P(O_s^A|\omega_s) \tag{1.1}$$

where the superscript *A* indicates the observations of the module A. This assumption has been used implicitly in numerous applications and is valid whenever the noise processes are spatially independent [Derin and Cole 86][Marroquin *et al* 85].

1.3. Posterior Probability and Bayesian Estimation

Following the Bayesian formalism, the goodness of a segmentation can be evaluated in terms of its *a posteriori* expected loss,

$$E(Loss(\omega|O)) = \sum_{f\in\Omega} loss(\omega,f)P(f|O) \tag{1.2}$$

where the $P(f\backslash 0)$ denotes the posterior probability of *f* given the observation *0.* Bayes' rule can then be used to derive the a *posteriori* probability

$$P(\omega|O) = \frac{P(\omega)P(O|\omega)}{\sum_{f\in\Omega} P(f)P(O|f)} \tag{1.3}$$

From (1.1), observe that scaling all $P(O_s|l)$ by a constant factor for fixed s does not change the posterior distribution in (1.3). This fact allows us to combine the likelihoods without having to normalize the results

The choice for the loss function depends on the characteristics of a particular application In [Geman and Geman 84], the Maximum A Posteriori (MAP) estimation is used. A simulated annealing procedure with a stochastic sampler (Gibbs sampler) carries out the computation. In [Marroquin *et al* 85], the Maximixer of the Posterior Marginals (MPM) estimation is proposed. This approach, computing the *a posteriori* probabilities for the segmentations given the set of opinions from the early modules and the *a priori* probability distribution, can support both the MAP and MPM estimation methods as well as other Bayesian estimations.

2. Markov Random Fields

Markov Random Fields have been used for image modeling in many applications for the past few years [Geman and Geman 84] [Marroquin *et al* 85] [Derin and Cole 86]. One of the most successful applications of MRFs is to model the spatial interactions of image features. In this section, we review the properties of MRFs and describe how to encode prior knowledge in this formalism.

2.1. Definition

Let $X = \{X_s, s\in S\}$ be a set of random variables indexed by S and E a set of unordered 2-tuple (s_i, s_j)'s representing the connections between the elements in S. The set E defines a neighborhood system $N = \{N_s|s\in S\}$, 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 $(s, r)\in E$

Let $\omega = \{X_s = \omega_s, s\in S\}$ be a configuration $X, \omega_s\in L$, and Ω the set of all possible configurations. We say X is a *Markov Random Field* with respect to N and *P,* where *P* is a probability function, if and only if

$$P(X=\omega) > 0 \text{ for all } \omega\in\Omega \tag{2.1}$$

$$P(X_s=\omega_s|X_r=\omega_r, r\in S, r\neq s) = P(X_s=\omega_s|X_r=\omega_r, r\in N_s) \tag{2.2}$$

The conditional probabilities in the right-hand side of (2.2) are called the local characteristics that characterize the random field. An intuitive interpretation of (2.2) is that the contextual information provided by *S* - s to *a* 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. A very desirable property of MRFs that makes them attractive to scientists in many disciplines is the MRF-Gibbs equivalence described in the following theorem.

2.2. MRF-Gibbs Equivalence

Hammersley-Clifford Theorem: A random field X is an MRF with respect to the neighborhood system N if and only if there exists a function V such that

$$P(\omega) = \frac{e^{-\frac{1}{T}U(\omega)}}{Z} \quad \text{for all } \omega\in\Omega \tag{2.3}$$

where

$$U(\omega) = \sum_{c\in C} V_c(\omega) \tag{2.4}$$

C is the set of totally connected subgraphs (cliques) with respect to N. Z is a normalizing constant, so that the probabilities of all realizations sum to one.

Several terminologies from Physics can provide intuition about the Gibbs measure - the right-hand side of (2.3). T is the *temperature* of the field that controls the flatness of the distribution of the configurations. A *potential V* is a way to assign a number $V_c(\omega)$ to every subconfiguration ω_c of a configuration ω, where $c\in C$. $U(\omega)$, the sum of the local potentials, is the *energy* of the configuration ω . A system is in *thermal equilibrium* when the probabilities of its configurations follows the Gibbs measure.

2.3. Encoding Prior Knowledge

For the image segmentation problem, we must choose an appropriate neighborhood system and a potential function for the random field X over the image S to represent prior knowledge about the image. The neighborhoods should be large enough to capture the interactions between the primitive elements but still small enough for a machine to carry out the computations required to make an estimation. The higher the energy measure of a configuration, the less likely it is to occur.

3. Combining Opinions of Early Visual Modules

Most research on evidence combination has focused on updating the "belief in a given hypothesis about an individual element when a piece of new evidence becomes available [Pearl 86]. This approach, however, is not suitable for our purpose. We believe that an information fusion mechanism should constantly maintain a representation of knowledge to reflect the total information available, except possibly for transient periods of time for aggregating evidence locally. Maintaining "marginal belief requires the effects of updating local "belief to be spatially propagated, thus violating such a requirement.

In this section, we limit our attention to an individual element s of S. We show how the opinions about s can be combined and provide the probabilistic justification for the proposed method. In Section 4 we show how the updating of this joint probability distribution given a new set of opinions about a set of primitive elements can be carried out with simple operations.

3.1. Representations and Combination Rules

As in Pearl's construction [Pearl 86], we assume the segmentation labels can be organized as a hierarchical tree H (e.g. Figure 1). Node / denotes the hypothesis that the corresponding primitive element is of label /, i.e., $X_s = l$. The numbers maintained in our method indicate the degrees of hypothesis confirmation or disconfirmation provided by the collected evidence.

Let α_l denote the current degree of confirmation/ disconfirmation for node /. The probabilistic interpretations for the a's will be given in Section 3.2. Initially, a/ is set to unity for every / indicating "neither confirmed nor disconfirmed". Besides *a,* each internal node / keeps one value, w^1 for each son *i.* Initially, w_l^i is set to the *a priori* probability of *i* given /. Obviously, the w_l^i's of each node sum up to unity initially.

Suppose a module A reports its opinion as a set of likelihood ratios $\{\lambda_l^A | l \in L_A\}$ where L_A is a set of mutually exclusive labels contained in H as described in Section 1.2. The corresponding a's are updated according to the rule:

$$\alpha_l \leftarrow \lambda_l^A \alpha_l \quad \text{for each } l \in L_A \tag{3.1}$$

To maintain the coherence of the a's, the effect of this opinion has to be propagated throughout the label tree by the following process:

(1) Every node l, $l \in L_A$, sends a message, $m = \lambda_l^A$, to its father and each of its sons.

(2) Any node k that receives a message m from its father, passes m to all its sons and replaces α_k by $m\alpha_k$, that is,

$$\alpha_k \leftarrow \lambda_k^A \alpha_k \tag{3.2}$$

(3) Any node j that receives a message m from one of its sons (say i), updates w_i^j by mw_i^j, i.e.,

$$w_i^j \leftarrow m w_i^j \tag{3.3}$$

and sends a message m' to its father, where

$$m' = \sum_i w_i^j \tag{3.4}$$

then updates a_j and all the $w^{j'}$s according to

$$a_j \leftarrow m' a_j \tag{3.5a}$$

$$w_k^j \leftarrow \frac{w_k^j}{m'} \quad \text{for all } k \tag{3.5b}$$

where the summation in (3.4) is taken over all the sons of *J-*

The combination and propagation procedures are commutative and associative, so their order is irrelevant.

3.2. Probabilistic Justification

The above method fits in the Bayesian formalism if we maintain two notions of conditional independence. First, evidence 0^A that bears directly on a label / says nothing about the descendants of/:

$$P(O^A | l, l_i) = P(O^A | l), \quad l_i \text{ descendant of } l, \tag{3.6}$$

$$P(O^A | \neg l, l_j) = P(O^A | \neg l) \quad l_j \text{ descendant of } \neg l$$

Second, the observations of different modules are conditionally independent.

$$\frac{P(O|l)}{P(O|\neg l)} = \prod_A \frac{P(O^A|l)}{P(O^A|\neg l)} \tag{3.7}$$

where the product on the right-hand side is over a set of modules and *0* is the union of their observation 0^A's

As suggested by Pearl, (3.6) states that when the observation O^A is a unique property of /, common to all its descendants, once we know / is true/false, the identity of l_i or l_j does not make O^A more or less likely. (3.7), implicitly used in Pearl's scheme, states that each piece of evidence observed by the early modules provides independent information about a label. We believe that the disparate types of image clues in vision applications satisfy this assumption.

We define *consistent states* of a's as the states in which for each available opinion, all of the a's are either updated according to rules (3.1) - (3.5), or none of the a's have been changed with respect to this opinion. We say that a set of opinions *derives* a consistent state if all opinions in this set, and no other opinions, have been used to update the a's. The following theorem relates the a's to the likelihood probabilities at consistent states.

Theorem 1 [Chou and Brown 87]: Let a_l denote the *a* value for / at the consistent state *t,* and $P(O_t|l)$ be the probability of O_t given the label l, where O_t denotes the union of those observations that form the set of opinions that derives the state t. If $O_t \neq \emptyset$, then

$$\alpha_l^t = c_t P(O_t|l) \quad \text{for all } l \in H \tag{3.8}$$

where c, is a constant depending only on *t,* given (3.6) and (3.7).

Applying Theorem 1 and Bayes' rule, we have:

Corollary 1: Let a/ denote the a value for / at the consistent state t. If P_θ is the prior *p.d.f.* of a set of mutually exclusive and exhaustive labels L, then the posterior probability $P_t(l)$ of $l \in L$ at the consistent state *t* is

$$P_t(l) = \frac{\alpha_l^t P_\theta(l)}{\sum_{l \in L} \alpha_l^t P_\theta(l)}.$$

To summarize: We have developed an evidence combination method for a hierarchy of hypotheses based on the notions of conditional independence given by (3.6) and (3.7). This scheme, besides having all the characteristics listed in [Pearl, 86], has the following advantages:

(1) The computations involved are extremely simple. Simpler and fewer messages must be passed. Normalizations are never needed since relative degrees of confirmation/disconfirmation are maintained instead of probabilities (Theorem 1).

(2) This scheme decouples the notion of evidence and *a priori* belief. In the next section we show this characteristic is very helpful when the prior knowledge is represented as an MRF.

4. Combining Prior Knowledge with Observations

In this section, we move our attention to the relationships of segments of the image S. Recall that in the last section, each primitive element is associated with a set of a's to maintain the opinions of the early visual modules. Let β_s denote the set of a's associated with s€S, and $\beta_s(l)$ be the a value for label / in β_s. Define a *global consistent state* to be a state of the B'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, *X,EL* - a mutually exclusive and exhaustive label set in H, with respect to a neighborhood system N. (1.1), Theorem 1, Bayes rule, and the Hammersley-Clifford Theorem lead to the conclusion that the *a posteriori* Gibbs measure of a configuration to at a global consistent state *t*

$$P_t(\omega) = \frac{e^{-\frac{1}{T}\sum_{c \in C} V_c(\omega) + \sum_{s \in S} \ln(\beta_s^t(\omega_s))}}{Z_t} \tag{4.1}$$

Only simple local operations are needed to update the energy measure and local characteristics as new opinions from the early visual modules become available. Therefore, MAP and MPM estimations can easily be implemented in the proposed framework (Section 5). We believe that based on this property, novel estimation algorithms can ultimately be designed that incrementally improve their estimations as more and more information arrives. For now, the existing Bayesian estimation methods can be invoked at any global consistent state to provide the up-to-date estimations.

5. Experimental Results

We demonstrate the method using two images of overlapping rectangular patches (Fig 2a, 3a). Each patch in the first image corresponds to a geometrically identical patch in the second. The intensities of the patches in each image are randomly selected from the range [0, 255], with no intensity correlation between images. Gaussian zero mean noise is added, with standard deviation 16 and 12 respectively. These two images

can be considered as two different sources of information about the same set of rectangular objects.

A set of likelihood edge detectors, an early version of the detectors described in [Sher 86], provides a set of likelihood ratios - $\{\frac{P(O|E_i)}{P(O|NE)} | i = 1,2,3,4\}$ - for each pixel given the 3×3 window of intensities centered at it, where *NE* denotes the hypothesis that the given pixel is not an edge element, and *E,* denotes the hypothesis that the given pixel is an edge of one of the four (horisontal, vertical, and two diagonal) orientations. The likelihoods are computed using a model for step edges and a Gaussian model for additive noise. Figure 2.b and 3b show the Maximum Likelihood Estimation (MLE) for edges in Figure 2.a and 3.a respectively. That is, a pixel *s* is on if and only if max B
t

We use a homogeneous and isotropic MRF with a third order neighborhood system over the image lattice to encode a body of basic knowledge about edges. Cliques of size 3 are used to discourage parallel and competing edges, whereas cliques of size 2 are used to encourage line continuations, region homogeneity and to discourage breaks in the line forming process. The potential assignments for the cliques are chosen conservatively in the sense that estimation methods based on (4.1) make as few false detections of edges as possible while maintaining reasonable detectability.

A software package has been implemented to study the behavior of various estimation criteria and schemes [Chou and Raman 87]. Here we show the results of using this package to perform MPM estimations based on the Monte Carlo procedure proposed in [Marroquin *et al* 85]. Figure 2c and 3.c show the MPM estimations based on the statistics collected over 300 iterations. Considering Figures 2.a and 3.a to provide only partial evidence to support the *NE* and E,'s hypotheses, Figure 4.a and 4.b show the MLE and the MPM estimations resulting from applying the upward propagation rule (3.3H3.5). Here the initial $w_D = 0.75$.

Alternatively we can consider that Figure 2.a and 3a independently support the same set of hypotheses. By applying rule (3.1), we obtain Figure 5.a and 5b representing the MLE and the MPM estimations based on the combined information. Observe the lines detected in the lower left quadrant of Figure 5b that do not show up in either of Figure 2c and 3.c, and the false detections in Figure 2.c and 3c that are removed in Figure 5b. These sorts of results can not be achieved by multi-modal segmenters that rely on Boolean operations to combine evidence.

6. Future Research

We are currently improving the MRF model to handle curved lines. One of our ultimate goals is to encode all kinds of geometrical and photometrical constraints in terms of local clique potentials in an MRF that has general connectivity. The stochastic estimation methods are computationally very expensive. We are now designing a deterministic estimation algorithm that incremently improves its estimation as new evidence arrives. We believe this method of information fusion can be applied to problems other than image segmentation as well.

Acknowledgements

We would like to thank Dave Sher for providing his likelihood edge detectors, Rajeev Raman for designing and implementing software for the experiments, and Henry Kyburg for providing many valuable suggestions.

REFERENCES

1 Bolles, R. C. "Verification Vision for Programmable Assembly." In *Proc.* IJCA/-77. Aug. 1977, pp. 569-575.

2 Chou, P. B. and C. M. Brown, "Multi-Modal Segmentation Using Markov Random Fields." In *Proc. Darpa Image Understanding Workshop,* Feb. 1987, pp 663-670.

3 Chou, P. B. and R. Raman, "Relaxation Algorithms Based on Markov Random Fields." Technical Report 212, University of Rochester, Computer Science Department, Apr. 1987.

4 Derin, H. and W. S. Cole, "Segmentation of Textured Images Using Gibbs Random Fields." *Computer Vision, Graphics, and Image Processing,* 35, *1986, pp. 72-98.*

5 Geman, S. and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Patter Analysis and Machine Intelligence,* PAMI-6, No. 6, 1984.

6 Marroquin, J., S. Mitter, and T. Poggio, "Probabilistic Solution of Ill-Posed Problems in Computational Vision." *Proc. Darpa Image Understanding Workshop,* Dec. 1985.

7 Pearl, J. "On Evidential Reasoning in a Hierarchy of *Hypotheses" Artificial Intelligence,* 16, No. 2, Feb. 1986.

8 Sher, D. B., "Advanced Likelihood Generators for Boundary Detection." Technical Report 197, University of Rochester, Computer Science Department, Jan 1987.