Communicated by John Hertz

Bayesian Self-Organization Driven by Prior Probability Distributions

Alan L. Yuille
Stelios M. Smirnakis
Lei Xu
Division of Applied Sciences, Harvard University, Cambridge, MA 02138, USA

Recent work by Becker and Hinton (1992) shows a promising mechanism, based on maximizing mutual information assuming spatial coherence, by which a system can self-organize to learn visual abilities such as binocular stereo. We introduce a more general criterion, based on Bayesian probability theory, and thereby demonstrate a connection to Bayesian theories of visual perception and to other organization principles for early vision (Atick and Redlich 1990). Methods for implementation using variants of stochastic learning are described.

1 Introduction

The input intensity patterns received by the human visual system are typically complicated functions of the object surfaces and light sources in the world. It seems probable, however, that humans perceive the world in terms of surfaces and objects (Nakayama and Shimojo 1987). Thus the visual system must be able to extract information from the input intensities that is relatively independent of the actual intensity values. Such abilities may not be present at birth and hence must be learned. It seems, for example, that binocular stereo develops at about the age of 2 to 3 months (Held 1987).

Becker and Hinton (1992) describe an interesting mechanism for self-organizing a system to achieve this. The basic idea is to assume spatial coherence of the structure to be extracted and to train a neural network by maximizing the mutual information between neurons with spatially disjoint receptive fields (see Fig. 1). For binocular stereo, for example, the surface being viewed is assumed flat (see Becker and Hinton 1992, for generalizations of this assumption) and hence has spatially constant disparity. The intensity patterns, however, do not have any simple spatial behavior. Adjusting the synaptic strengths of the network to maximize the mutual information between neurons with nonoverlapping receptive fields, for an ensemble of images, causes the neurons to extract features that are spatially coherent, thereby obtaining the disparity.

Neural Computation 7, 580–593 (1995)

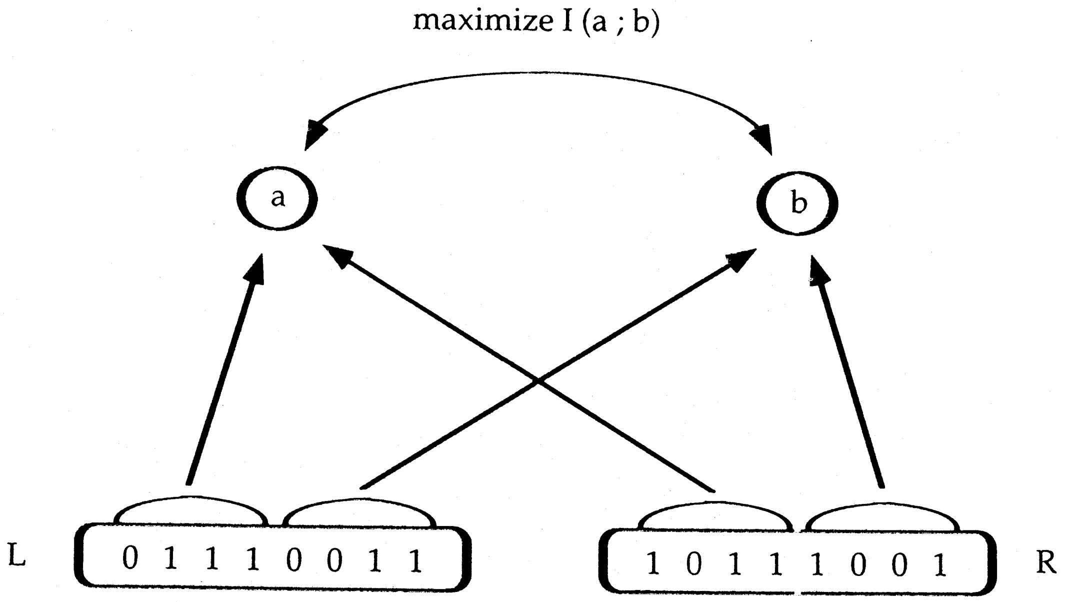

Figure 1: In Hinton and Becker's initial scheme, maximization of mutual information between neurons with spatially disjoint receptive fields leads to disparity tuning, provided they train on spatially coherent patterns (i.e., those for which disparity changes slowly with spatial position).

We argue that this approach has three key ingredients:

1. It uses strong prior knowledge about the output variables, i.e., it assumes that the disparities are spatially constant. If this assumption is not valid then the performance of the system will degrade.
2. It represents the desired outputs as functions of the inputs by a multilayer perceptron with adjustable weights.
3. It proposes a criterion, mutual information maximization, motivated by the prior knowledge (see point 1) to determine the weights.

The approach relies heavily on prior assumptions about the form of the outputs. This is similar to Bayesian theories of visual perception that also rely (Clark and Yuille 1990) on prior assumptions about properties of the world, such as binocular disparities. Such priors are needed because of the ill-posed nature of vision (Poggio *et al.* 1985) and can be thought of as *natural constraints* (Marr 1982).

This similarity motivates the following questions. Can we reformulate Becker and Hinton's theory so that it can be applied directly to learning Bayesian theories of vision? More precisely, assuming a prior of the type commonly used in vision, can we find an optimization criterion and learning algorithm such that we can learn the corresponding Bayesian theory?

This note shows that it is indeed possible to reformulate Becker and Hinton to make it compatible with Bayesian theories. In particular, their algorithm for stereo corresponds to one of the standard priors used for Bayesian stereo theories (see Section 3). The key idea is to force the activity distribution of the outputs, $\mathbf{S}$, to be close to a prespecified prior distribution $P_p(\mathbf{S})$. Our approach is general and is related to the work performed by Atick and Redlich (1990) for modeling the early visual system. In previous work (Yuille *et al.* 1993) we proved that applying our approach to linear filtering problems leads to a solution that is the square root of the Wiener filter in Fourier space. A similar result has been derived (Redlich, private communication) from the principles described in Atick and Redlich (1990).

We should clarify what we mean by "learning a Bayesian theory." A Bayesian theory for estimating a scene property $\mathbf{S}$ from input $\mathbf{D}$ consists of three elements: (1) a prior for the property $P_p(\mathbf{S})$, (2) a likelihood function $P_l(\mathbf{D} \mid \mathbf{S})$, and (3) an algorithm for estimating $\mathbf{S}^*(\mathbf{D}) = \arg\max_{\mathbf{S}} P_l(\mathbf{D} \mid \mathbf{S}) P_p(\mathbf{S})$.[1] Because we assume that the prior is known we are essentially learning the likelihood function and the algorithm. Our approach, after training, will yield a neural net, or some other function approximation scheme, that computes $\mathbf{S}^*(\mathbf{D})$. In related work (Smirnakis and Yuille 1994) we assume that both prior and likelihood are known and train a network to learn the algorithm.

This can be contrasted to alternative ways for learning Bayesian theories. Hidden Markov models (Paul 1990) (see Section 5) learn both the priors and the likelihood functions. A general purpose optimization algorithm, dynamic programming, is then used to compute the MAP, or some alternative, estimator. This approach can be highly effective, though dynamic programming is efficient only for one-dimensional problems and functional forms for the prior and likelihood are required. Kersten *et al.* (1987) describe Bayesian learning with a teacher that yields the algorithm $\mathbf{S}^*(\mathbf{D}) = \arg\max_{\mathbf{S}} P_l(\mathbf{D} \mid \mathbf{S}) P_p(\mathbf{S})$. But as Becker and Hinton have shown, a teacher is not always necessary.

We will take the viewpoint that the prior $P_p(\mathbf{S})$ is assumed known in advance by the visual system (perhaps by being specified genetically) and will act as a self-organizing principle. Later we will discuss ways that this might be relaxed.

2 Theory

We assume that the input $\mathbf{D}$ is a function $F(\mathbf{n}, \alpha)$ of a *signal* α that the system wants to determine and a *distractor* $\mathbf{n}$. These quantities are vectors indexed by spatial location (see Fig. 2). For example, α might correspond to the disparities of a pair of binocular stereo images and $\mathbf{n}$ to the intensity

[1] This corresponds to the commonly used maximum a posteriori (MAP) estimator. Other estimators may be preferable, but we will consider only MAP in this paper.

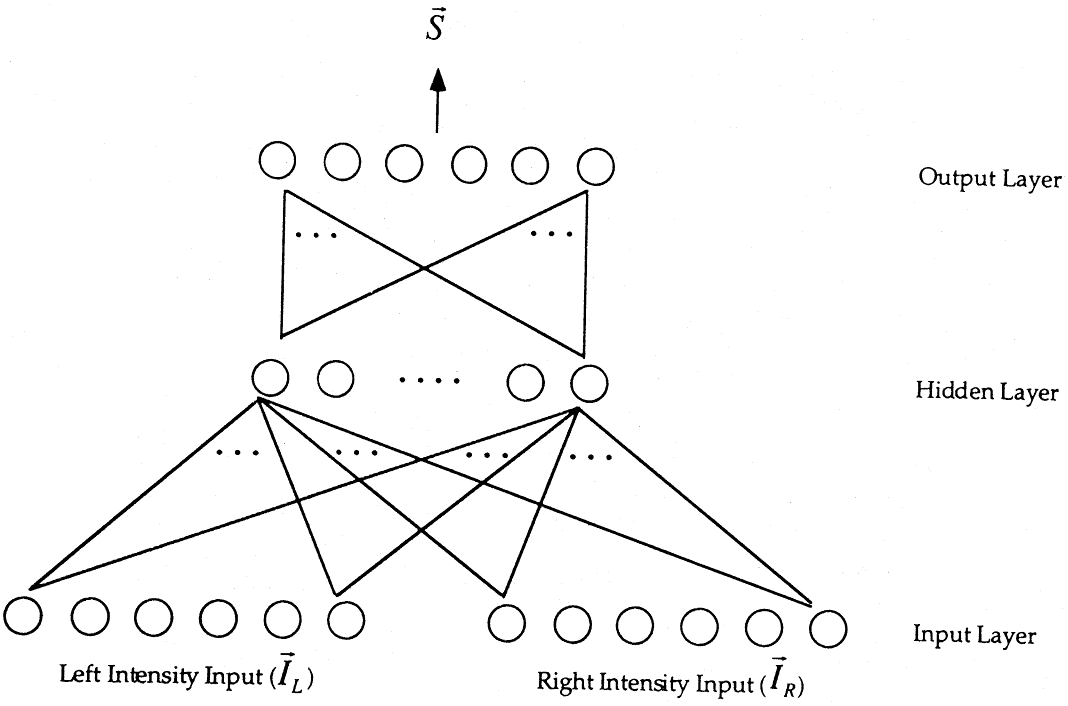
$\vec{S}$
Output Layer
Hidden Layer
Input Layer
Left Intensity Input ($\vec{I}_L$)
Right Intensity Input ($\vec{I}_R$)

Figure 2: Note that the vectors $\mathbf{I}_\mathrm{L}$ and $\mathbf{I}_\mathrm{R}$ represent the intensities falling on the left and right retinas respectively, and are indexed by spatial location. **S** represents the vector of the disparities to be extracted. That is, the output S_i of output unit i represents the disparity at spatial location i. By setting some of the synapses to zero we obtain the disjoint receptive fields of the Becker and Hinton paradigm (Fig. 1).

patterns. The variables have distributions $P_\mathrm{n}(\mathbf{n})$ and $P_\mathrm{p}(\alpha)$, respectively. Note that **D** and $P_\mathrm{p}(\alpha)$ are assumed to be known but $P_\mathrm{n}(\mathbf{n})$ and the functional form of $F(\mathbf{n}, \alpha)$ are unknown.

The input distribution is given by

$$P_\mathrm{D}(\mathbf{D}) = \int\int \delta[\mathbf{D} - F(\mathbf{n}, \alpha)] P_\mathrm{n}(\mathbf{n}) P_\mathrm{p}(\alpha) [d\alpha][d\mathbf{n}]$$

and can be observed by the system.

Let the output of the system be $\mathbf{S} = G(\mathbf{D}, \gamma)$ where **G** is a function of a set of parameters γ to be determined. For example, the function $\mathbf{G}(\mathbf{D}, \gamma)$ could be represented by a multilayer perceptron with γ being the synaptic weights. By approximation theory, it can be shown that a large variety of neural networks can approximate any input–output function arbitrarily well given enough hidden nodes (Hornik *et al.* 1991). We can combine these formulas to give

$$\mathbf{S} = \mathbf{G}[\mathbf{F}(\mathbf{n}, \alpha), \gamma] \tag{2.1}$$

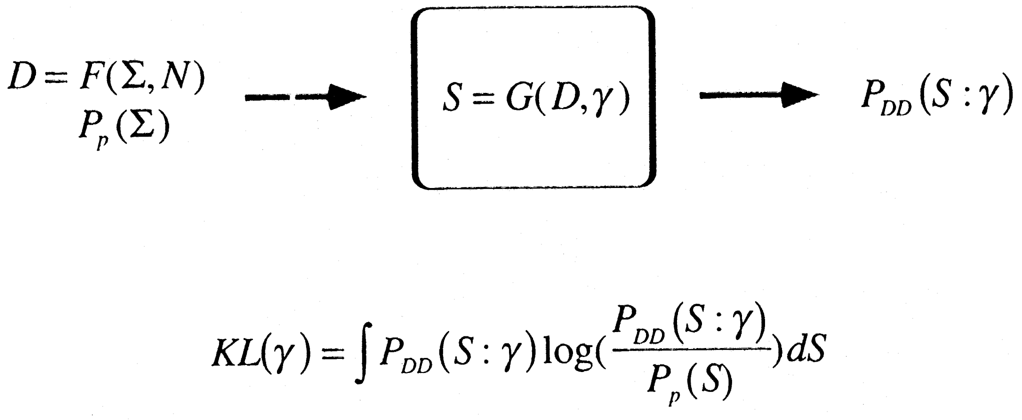
$D = F(\Sigma, N)$
$P_p(\Sigma)$
$S = G(D,\gamma)$
$P_{DD}(S:\gamma)$
$KL(\gamma) = \int P_{DD}(S:\gamma)\log(\frac{P_{DD}(S:\gamma)}{P_p(S)})dS$

Figure 3: The parameters γ are adjusted to minimize the Kullback–Leibler distance between the prior (P_{p}) distribution of the true signal (Σ) and the derived distribution (P_{DD}) of the network output (S).

The aim of self-organizing the network is to ensure that the parameters γ are chosen so that the outputs $\mathbf{S}$ are as close to the α (or some simple transformation of the αs) as possible. We claim that this can be achieved by adjusting the parameters γ so as to make the derived distribution of the outputs $P_{\mathrm{DD}}(\mathbf{S}:\gamma) = \int \delta[\mathbf{S} - \mathbf{G}(\mathbf{D},\gamma)]P_{\mathrm{D}}(\mathbf{D})[d\mathbf{D}]$ as close as possible to $P_{\mathrm{p}}(\mathbf{S})$.

This can be seen to be a consistency condition for a Bayesian theory. From Bayes's formula we obtain the condition:

$$\int P(\mathbf{S} \mid \mathbf{D})P_{\mathrm{D}}(\mathbf{D})[d\mathbf{D}] = \int P(\mathbf{D} \mid \mathbf{S})P_{\mathrm{p}}(\mathbf{S})[d\mathbf{D}] = P_{\mathrm{p}}(\mathbf{S}) \tag{2.2}$$

This is equivalent to our condition provided we identify $P(\mathbf{S} \mid \mathbf{D})$ with $\delta[\mathbf{S} - \mathbf{G}(\mathbf{D},\gamma)]$.

To make this more precise we must define a measure of similarity between the two distributions $P_{\mathrm{p}}(\mathbf{S})$ and $P_{\mathrm{DD}}(\mathbf{S}:\gamma)$. An attractive measure is the Kullback–Leibler distance (the entropy of P_{DD} relative to P_{p}):

$$KL(\gamma) = \int P_{\mathrm{DD}}(\mathbf{S}:\gamma)\log\frac{P_{\mathrm{DD}}(\mathbf{S}:\gamma)}{P_{\mathrm{p}}(\mathbf{S})}[d\mathbf{S}] \tag{2.3}$$

Thus our theory (see Fig. 3) corresponds to adjusting the parameters γ to minimize the Kullback–Leibler distance between $P_{\mathrm{p}}(\mathbf{S})$ and $P_{\mathrm{DD}}(\mathbf{S}:\gamma)$. This measure can be divided into two parts: (1) $-\int P_{\mathrm{DD}}(\mathbf{S}:\gamma)\log P_{\mathrm{p}}(\mathbf{S})[d\mathbf{S}]$ and (2) $\int P_{\mathrm{DD}}(\mathbf{S}:\gamma)\log P_{\mathrm{DD}}(\mathbf{S}:\gamma)[d\mathbf{S}]$. As we now show both terms have very intuitive interpretations.

Suppose that $P_{\mathrm{p}}(\mathbf{S})$ can be expressed as a Markov random field [i.e., the spatial distribution of $P_{\mathrm{p}}(\mathbf{S})$ has a local neighborhood structure, as is commonly assumed in Bayesian models of vision]. Then, by the

Hammersely–Clifford theorem, we can write $P_p(\mathbf{S}) = e^{-\beta E_p(\mathbf{S})}/Z$ where $E_p(\mathbf{S})$ is an energy function with local connections [for example, $E_p(\mathbf{S}) = \sum_i (S_i - S_{i+1})^2$], β is an inverse temperature, and Z is a normalization constant.

Then the first term can be written as

$$\begin{aligned} -\int P_{DD}(\mathbf{S}:\gamma)\log P_p(\mathbf{S})[d\mathbf{S}] & \\ = \int\int \delta[\mathbf{S}-\mathbf{G}(\mathbf{D},\gamma)]P_D(\mathbf{D})\beta E_p(\mathbf{S})[d\mathbf{D}][d\mathbf{S}] + \log Z & \\ = \int \beta E_p[\mathbf{G}(\mathbf{D},\gamma)]P_D(\mathbf{D})[d\mathbf{D}] + \log Z & \\ = \beta\langle E_p[\mathbf{G}(\mathbf{D},\gamma)]\rangle_D + \log Z & \qquad (2.4) \end{aligned}$$

We can ignore the $\log Z$ term since it is a constant (independent of γ). Minimizing the first term with respect to γ will therefore try to minimize the energy of the outputs averaged over the inputs—$\langle E_p[\mathbf{G}(\mathbf{D},\gamma)]\rangle_D$—which is highly desirable [since it has a close connection to the minimal energy principles in Poggio *et al.* (1985), and Clark and Yuille (1990)]. It is important, however, to avoid the trivial solution $\mathbf{G}(\mathbf{D},\gamma) = \textit{constant}$ or solutions where $\mathbf{G}(\mathbf{D},\gamma)$ is very small for most inputs. Fortunately these solutions will be discouraged by the second term.

The second term $\int P_{DD}(\mathbf{D},\gamma)\log P_{DD}(\mathbf{D},\gamma)[d\mathbf{D}]$ can be interpreted as the negative of the entropy of the derived distribution of the output. Minimizing it with respect to γ is a maximum entropy principle that will encourage variability in the outputs $G(\mathbf{D},\gamma)$ and hence prevent the trivial solutions.

The two terms combine to determine the γ so that the energy of the output variables is minimized while maximizing their variability. This is closely related to Becker and Hinton's method of maximizing the mutual information between pairs of output variables—essentially assuming a spatially constant prior distribution for $\mathbf{S}$. At the same time it is reminiscent of other organizational principles for early vision based on information theory (Atick and Redlich 1990).

How can one guarantee that the optimal solution to our criteria will indeed extract the signal? This will depend on a number of factors: (1) the forms of the functions $\mathbf{F}$ and $\mathbf{G}$, (2) the forms of the probability distributions $P_n(\mathbf{n})$ and $P_p(\alpha)$, and (3) whether the prior P_p is indeed correct or not.

It is straightforward to write down the conditions for the derived distribution to be equal to the prior distribution (assuming that the prior is correct). This is a stronger condition than requiring the Kullback–Leibler distance to be minimal (though, if equality is possible, minimizing Kullback–Leibler would lead to it). It is

$$P_p(\mathbf{S}) = \int\int \delta\{\mathbf{S}-\mathbf{G}[\mathbf{F}(\mathbf{n},\alpha),\gamma]\}P_n(\mathbf{n})P_p(\alpha)[d\alpha][d\mathbf{n}] \qquad (2.5)$$

If one could find γ^* so that $\mathbf{G}[\mathbf{F}(\mathbf{n},\alpha),\gamma^*] = \alpha, \forall\mathbf{n},\alpha$ then the equation could be solved exactly. The condition $\mathbf{G}[\mathbf{F}(\mathbf{n},\alpha),\gamma^*] = \alpha$, however, is too strong. It requires that the function $\mathbf{G}$, which can be thought of as a nonlinear filter, is able to completely eliminate the dependence on $\mathbf{n}$.

We have assumed that the correct prior is known by the system, perhaps by being specified genetically. An alternative possibility is that the prior itself is learned by a method reminiscent of Occam's razor: the goodness of the prior is evaluated based on the Kullback–Leibler distance after self-organization, and a more complex prior is chosen if this distance is large (see also Mumford 1992).

3 Connection to Becker and Hinton

In this section, we show that the case of disparity extraction implemented by Becker and Hinton based on their principle of mutual information maximization arises as a special case of our formalism, by choosing a particular prior. The Becker and Hinton method (Becker and Hinton 1992) for extracting the disparity involves maximizing the mutual information between two network output units S_1, S_2 with spatially disjoint receptive fields, under the assumption that disparity is spatially coherent. S_1 and S_2 denote the scalar values of two units in the output layer of a neural network, indexed by spatial location. The mutual information between S_1, S_2 is given by

$$\begin{aligned} I(S_1, S_2; \gamma) &= -\langle \log P_{\rm DD}(S_1; \gamma)\rangle - \langle \log P_{\rm DD}(S_2; \gamma)\rangle \\ &\quad + \langle \log P_{\rm DD}(S_1, S_2; \gamma)\rangle \\ &= H(S_1; \gamma) - H(S_1 \mid S_2; \gamma) \end{aligned} \tag{3.1}$$

From this equation we see that we want to maximize the entropy, $H(S_1; \gamma)$, of S_1 while minimizing the conditional entropy, $H(S_1 \mid S_2; \gamma)$, of S_1 given S_2, which forces S_1 to be a deterministic function of S_2 (alternatively, by symmetry, we can interchange the roles of S_1 and S_2). For the discussion below we will use our criterion to reproduce the case in which this last term forces $S_1 \approx S_2$.

By contrast, in our version (see Fig. 4) we propose to minimize the expression $\langle \log P_{\rm DD}(S_1, S_2; \gamma)\rangle - \int \log P_{\rm p}(S_1, S_2) P_{\rm DD}(S_1, S_2; \gamma)[d\mathbf{S}]$. If we ensure that the prior $P_{\rm p}(S_1, S_2) \propto e^{-\tau(S_1 - S_2)^2}$, then, for large τ, our second term will force $S_1 \approx S_2$ and our first term will maximize the entropy of the joint distribution of S_1, S_2. We argue that this is effectively the same as Becker and Hinton (1992), since maximizing the joint entropy of S_1, S_2 with S_1 constrained to equal S_2 is equivalent to maximizing the individual entropies of S_1 and S_2 with the same constraint.

To be more concrete, we consider Becker and Hinton's implementation of the mutual information maximization principle in the case of units with continuous outputs. They assume that the outputs of units 1, 2 are gaussian[2] and perform steepest descent to maximize the symmetrized

[2] We assume for simplicity that these gaussians have zero mean.

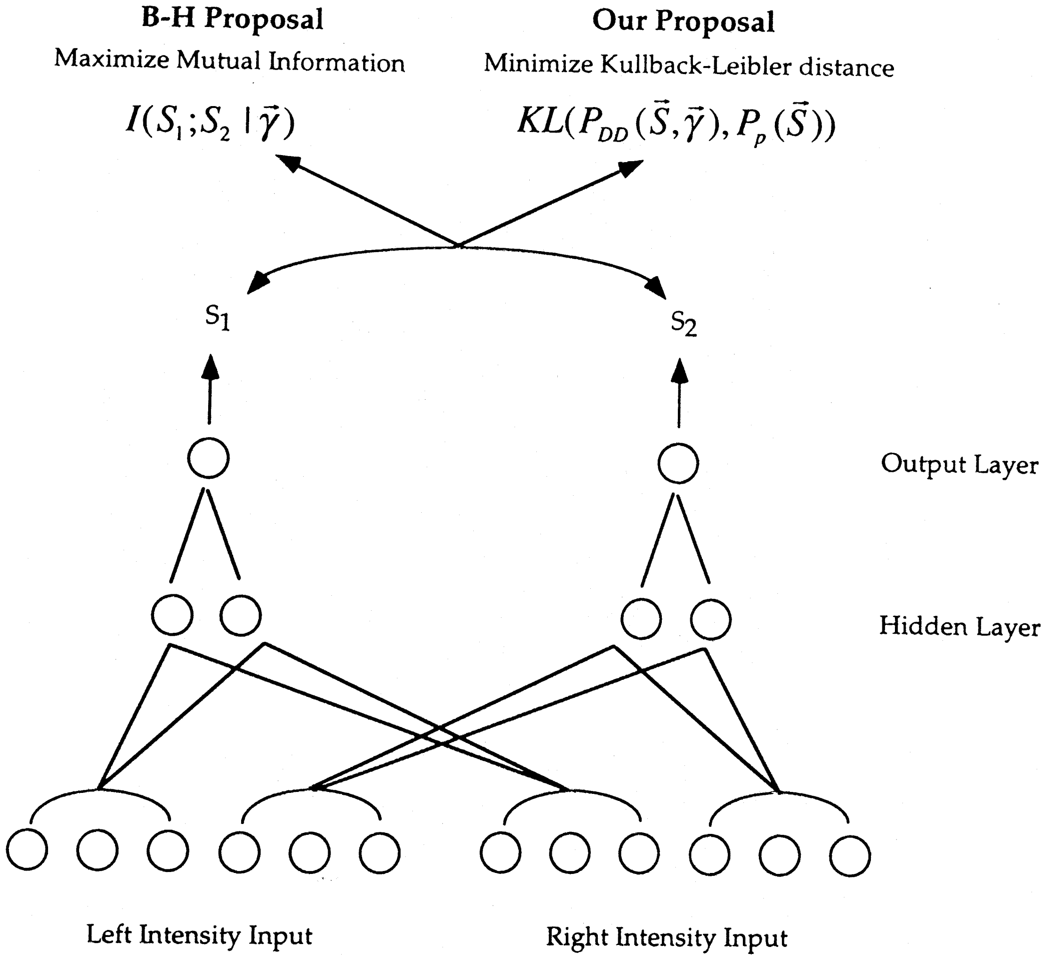
B-H Proposal
Maximize Mutual Information
$I(S_1;S_2 \mid \vec{\gamma})$
Our Proposal
Minimize Kullback-Leibler distance
$KL(P_{DD}(\vec{S},\vec{\gamma}), P_p(\vec{S}))$
S1
S2
Output Layer
Hidden Layer
Left Intensity Input
Right Intensity Input

Figure 4: Comparing our theory with Becker and Hinton's. Observe that setting $P_{\mathrm{p}}(S_1, S_2) \propto e^{-\tau(S_1-S_2)^2}$ forces $S_1 \approx S_2$ for large τ, implementing their assumption that the disparity is spatially coherent.

form of the mutual information between S_1 and S_2:

$$\begin{aligned} I(S_1, S_2) &= \log \frac{V(S_1)}{V(S_1 - S_2)} + \log \frac{V(S_2)}{V(S_1 - S_2)} \\ &= \log V(S_1) + \log V(S_2) - 2 \log V(S_1 - S_2) \end{aligned} \tag{3.2}$$

where $V(\cdot)$ stands for variance over the set of inputs. They assume that the difference between the two outputs can be expressed as uncorrelated additive noise, $S_1 = S_2 + N$. Therefore, their criterion amounts to maximizing

$$E_{\mathrm{BH}}[V(S_2), V(N)] = \log\{V(S_2) + V(N)\} + \log V(S_2) - 2 \log V(N) \tag{3.3}$$

For our scheme we make similar assumptions about the distributions of S_1 and S_2. We then see that, up to additive constants independent of γ,

$\langle \log P_{\mathrm{DD}}(S_1, S_2)\rangle = -1/2 \log\{\langle S_1^2\rangle\langle S_2^2\rangle - \langle S_1 S_2\rangle^2\} = -1/2 \log\{V(S_2)V(N)\}$ [since $\langle S_1 S_2\rangle = \langle (S_2 + N)S_2\rangle = V(S_2)$ and $\langle S_1^2\rangle = V(S_2) + V(N)$]. We now observe that if we choose the prior distribution $P_{\mathrm{p}}(S_1, S_2) \propto e^{-\tau(S_1-S_2)^2}$ our criterion corresponds to minimizing $E_{YSX}[V(S_2), V(N)]$ where

$$E_{YSX}[V(S_2), V(N)] = -\log V(S_2) - \log V(N) + \tau V(N) \tag{3.4}$$

It is easy to see that maximizing $E_{\mathrm{BH}}[V(S_2), V(N)]$ will try to make $V(S_2)$ as large as possible and force $V(N)$ to zero [recall that, by definition, $V(N) \geq 0$]. On the other hand, minimizing our energy will try to make $V(S_2)$ as large as possible and will force $V(N)$ to $1/\tau$. Since τ appears as the inverse of the variance of the gaussian prior for $\mathbf{S} = (S_1, S_2)$, making τ large will force the prior distribution to approach $\delta(S_1 - S_2)$. Thus, in the case of large τ, our method has the same effect as the Becker and Hinton algorithm.

For this to be true, it is important to choose a network architecture satisfying the requirement that the output units representing disparity have spatially disjoint receptive fields (see Fig. 4). If this were not the case, the output units would run the risk of getting entrained on the receptive field overlap, provided it has the right probability structure. Even though we did not pursue this issue in the above analysis, it is, in principle, possible to implement such architectural constraints by defining a prior distribution on the weights of the network.

Note that, in principle, maximizing the mutual information between S_1, S_2 can only determine the network output up to transformations that leave the mutual information invariant. Which solution the network will settle at depends on the specifics of the implementation and on initial conditions. For instance, in the Becker and Hinton example the network sometimes settles so that $S_1 \approx S_2$, and sometimes so that $S_1 \approx -S_2$. This may not be always desirable. In this context, the ability to choose a prior affords a natural way to restrict the possible space of solutions.

4 Reformulating for Implementation in a General Setting

Our proposal requires us to minimize the Kullback–Leibler distance (equation 2.3) with respect to γ. In the previous section, we showed that Becker and Hinton's implementation of the mutual information maximization principle for disparity extraction arose as a special case of our formalism, for a particular prior. Therefore, their simulation already represents a concrete example of how our scheme can be implemented. In the present section, we endeavor to expand further by outlining two general implementation strategies based on variants of stochastic learning:

First observe that by substituting the form of the derived distribution, $P_{\mathrm{DD}}(\mathbf{S} : \gamma) = \int \delta[\mathbf{S} - \mathbf{G}(\mathbf{D}, \gamma)] P_{\mathrm{D}}(\mathbf{D})[d\mathbf{D}]$, into equation 2.3 and integrating out the $\mathbf{S}$ variable we obtain

$$KL(\gamma) = \int P_{\mathrm{D}}(\mathbf{D}) \log \frac{P_{\mathrm{DD}}[\mathbf{G}(\mathbf{D}, \gamma) : \gamma]}{P_{\mathrm{p}}[\mathbf{G}(\mathbf{D}, \gamma)]}[d\mathbf{D}] \tag{4.1}$$

This is the form of the Kullback–Liebler distance that we assume in the implementation strategies we describe below:

1. Assuming a representative sample $\{\mathbf{D}^{\mu} : \mu \in \Lambda\}$ of inputs we can approximate $KL(\gamma)$ by $\sum_{\mu\in\Lambda} \log\{P_{\mathrm{DD}}[\mathbf{G}(\mathbf{D}^{\mu},\gamma) : \gamma]/P_{\mathrm{p}}[\mathbf{G}(\mathbf{D}^{\mu},\gamma)]\}$. We can now, in principle, perform stochastic learning using backpropagation: pick inputs $\mathbf{D}^{\mu}$ at random and update the weights γ using $\log\{P_{\mathrm{DD}}[\mathbf{G}(\mathbf{D}^{\mu},\gamma) : \gamma]/P_{\mathrm{p}}[\mathbf{G}(\mathbf{D}^{\mu},\gamma)]\}$ as the error function.

To do this, however, we need expressions for $P_{\mathrm{DD}}[\mathbf{G}(\mathbf{D}^{\mu},\gamma) : \gamma]$ and its derivative with respect to γ. If the function $\mathbf{G}(\mathbf{D},\gamma)$ can be restricted to being 1-1 (artificially increasing the dimensionality of the output space if necessary) then we can obtain analytic expressions $P_{\mathrm{DD}}[\mathbf{G}(\mathbf{D},\gamma) : \gamma] = P_{\mathrm{D}}(\mathbf{D})/|\det(\partial\mathbf{G}/\partial\mathbf{D})|$ and $\{\partial \log P_{\mathrm{DD}}[\mathbf{G}(\mathbf{D},\gamma) : \gamma]/\partial\gamma\} = -(\partial\mathbf{G}/\partial\mathbf{D})^{-1}$ $(\partial^2\mathbf{G}/\partial\mathbf{D}\partial\gamma)$, where -1 denotes the matrix inverse.

To see this we observe that

$$\begin{aligned} P_{\mathrm{DD}}(\mathbf{S} : \gamma) &= \int \delta[\mathbf{S} - \mathbf{G}(\mathbf{D},\gamma)] P_{\mathrm{D}}(\mathbf{D})[d\mathbf{D}] \\ &= \frac{P_{\mathrm{D}}(\mathbf{D}^*)}{|\det(\partial\mathbf{G}/\partial\mathbf{D})(\mathbf{D}^*,\gamma)|} \end{aligned} \tag{4.2}$$

where $\mathbf{D}^* = G^{-1}(\mathbf{S},\gamma)$ and we assume that the function G is 1-1. It follows directly that

$$P_{\mathrm{DD}}[\mathbf{G}(\mathbf{D},\gamma) : \gamma] = \frac{P_{\mathrm{D}}(\mathbf{D})}{|\det(\partial\mathbf{G}/\partial\mathbf{D})(\mathbf{D},\gamma)|} \tag{4.3}$$

Substituting back into the K–L measure (equation 4.1) means that we must minimize with respect to γ the cost function $E[\gamma,\mathbf{D}]$ averaged over a sample of $\mathbf{D}$ (where we have dropped terms that are independent of γ):

$$E[\gamma,\mathbf{D}] = -\log\left|\det\frac{\partial\mathbf{G}}{\partial\mathbf{D}}(\mathbf{D},\gamma)\right| + \beta E_{\mathrm{p}}[\mathbf{G}(\mathbf{D},\gamma)] \tag{4.4}$$

We implement this by stochastic learning. Pick an input $\mathbf{D}$ at random, set $\gamma_{\mathrm{new}} = \gamma_{\mathrm{old}} - \zeta(\partial E/\partial\gamma)$ (where ζ is the learning rate), and repeat.

This involves calculating $\partial E/\partial\gamma$. After some algebra we find that

$$\frac{\partial}{\partial\gamma_a}\log\left|\det\left\{\frac{\partial\mathbf{G}}{\partial\mathbf{D}}(\mathbf{D},\gamma)\right\}\right| = \sum_{j,k}\left(\frac{\partial G_j}{\partial D_k}\right)^{-1}\frac{\partial^2 G_k}{\partial D_j\partial\gamma_a} \tag{4.5}$$

where -1 denotes the matrix inverse.

The contribution from the second term will simply be $\beta(\partial E/\partial\mathbf{G})$ $(\partial\mathbf{G}/\partial\gamma_a)$.

This analysis has assumed that $\mathbf{G}$ is a 1-1 function and requires, as a necessary condition, that the input and output spaces have the same dimension. This could often be ensured by adding additional output units or input units with fixed synaptic strengths.

2. Alternatively we can perform additional sampling to estimate $P_{\text{DD}}[\mathbf{G}(\mathbf{D},\gamma):\gamma]$ and $\{\partial \log P_{\text{DD}}[\mathbf{G}(\mathbf{D},\gamma):\gamma]/\partial\gamma\}$ directly from their integral representations. [This second approach is similar to Becker and Hinton (1992), though they are concerned with estimating only the first and second moments of these distributions.] The Kullback–Leibler measure corresponds to minimizing $KL(\gamma) = \sum_{\mu} E(\gamma, \mathbf{D}^{\mu})$, where $E(\gamma, \mathbf{D}^{\mu}) = \log P_{\text{DD}}[\mathbf{G}(\mathbf{D}^{\mu},\gamma):\gamma] + \beta E_{\text{p}}[\mathbf{G}(\mathbf{D}^{\mu},\gamma)]$.

Thus calculating the gradient of $E(\gamma, \mathbf{D}^{\mu})$ requires evaluating the expression $\{\partial P_{\text{DD}}[\mathbf{G}(\mathbf{D}^{\mu},\gamma):\gamma]/\partial\gamma\}/P_{\text{DD}}[\mathbf{G}(\mathbf{D}^{\mu},\gamma):\gamma]$. To estimate these quantities we make the approximation:

$$P_{\text{DD}}[\mathbf{G}(\mathbf{D}^{\mu},\gamma):\gamma] \approx \sum_{\nu} \frac{1}{\left[\sqrt{(2\pi)}\sigma\right]^{N}} e^{-(1/2\sigma^2)|\mathbf{G}(\mathbf{D}^{\mu},\gamma)-\mathbf{G}(\mathbf{D}^{\nu},\gamma)|^2} \tag{4.6}$$

where $\{\mathbf{D}^{\nu}\}$ are a representative set of samples from $P_{\text{D}}(\mathbf{D})$ and σ is a constant. This reduces to the previous expression, the first part of equation 4.2, in the limit as $\sigma \mapsto 0$ and as the size of the sample set tends to infinity.

A formula for $\{\partial P_{\text{DD}}[\mathbf{G}(\mathbf{D}^{\mu},\gamma):\gamma]/\partial\gamma\}$ can be obtained by differentiating (4.6) with respect to γ. This gives

$$\begin{aligned}&\frac{\partial P_{\text{DD}}[\mathbf{G}(\mathbf{D}^{\mu},\gamma):\gamma]}{\partial\gamma_a}\\ &\approx \sum_{\nu} \frac{1}{\left[\sqrt{(2\pi)}\sigma\right]^{N}} \times \left\{-\frac{1}{\sigma^2}\right\}\\ &\times \sum_{i} \left\{\frac{\partial G_i(\mathbf{D}^{\mu},\gamma)}{\partial\gamma_a} - \frac{\partial G_i(\mathbf{D}^{\nu},\gamma)}{\partial\gamma_a}\right\} \{G_i(\mathbf{D}^{\mu},\gamma) - G_i(\mathbf{D}^{\nu},\gamma)\}\\ &\times e^{-(1/2\sigma^2)|\mathbf{G}(\mathbf{D}^{\mu},\gamma)-\mathbf{G}(\mathbf{D}^{\nu},\gamma)|^2}\end{aligned} \tag{4.7}$$

The learning proceeds by picking a sample $\mathbf{D}^{\mu}$ from $P_{\text{D}}(\mathbf{D})$ and then an additional set of samples $\{\mathbf{D}^{\nu}\}$ to approximate the integrals 4.6 and 4.7 and hence enable us to calculate the gradient of $E(\gamma, \mathbf{D}^{\mu})$ and update the weights. Then the process repeats.

Note that this approach has the advantage of circumventing the demand that the dimensions of the input and output spaces be equal, i.e., that $\mathbf{G}$ be 1-1, and is more generally applicable.

5 Relationship to Hidden Markov Models and Maximum Likelihood Estimation

It is instructive to contrast our work to alternative learning approaches and, in particular, to hidden Markov models (HMMs)[3] (Paul 1990).

[3] Approaches closely related to HMMs are being used for learning stereo (Geiger, personal communication).

HMMs have been very successful in speech processing where models are trained for each recognizable speech segment. Here, however, we are considering training only a single HMM.

In an HMM there are hidden states and observables that, in our notation, correspond to $\mathbf{S}$ and $\mathbf{D}$, respectively. An HMM assumes (1) a prior model $P(\mathbf{S} \mid \boldsymbol{\beta})$, where the $\boldsymbol{\beta}$ are parameters to be learned, and (2) an imaging model $P(\mathbf{D} \mid \mathbf{S}, \boldsymbol{\alpha})$, where the $\boldsymbol{\alpha}$ are parameters to be learned. Together these generate probabilities $P(\mathbf{D} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_S P(\mathbf{D} \mid \mathbf{S}, \boldsymbol{\alpha}) P(\mathbf{S} \mid \boldsymbol{\beta})$ for the observables as functions of the parameters.[4] Similar expressions arise in MLE parameter estimation (Ripley 1992).

To learn the priors and likelihood functions we must estimate the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. This requires a set of data $\{\mathbf{D}^\mu\}$, indexed by μ, that we assume is a representative sample from the distribution $P(\mathbf{D})$ of the observables. We then train the system by maximum likelihood estimation (MLE). More precisely, we select the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that maximize $\prod_\mu P(\mathbf{D}^\mu \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$ or, equivalently, that maximize $\sum_\mu \log P(\mathbf{D}^\mu \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$. As the sample size tends to infinity this becomes equivalent to maximizing $\sum_{\mathbf{D}} P(\mathbf{D}) \log[P(\mathbf{D} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$ or, equivalently, to maximizing $\sum_{\mathbf{D}} P(\mathbf{D}) \log P(\mathbf{D} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) / P(\mathbf{D})]$ [since $P(\mathbf{D})$ is independent of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$]. Thus, in the infinite sample size limit, we are simply *minimizing* the Kullback–Leibler measure $(\sum_{\mathbf{D}} P(\mathbf{D}) \log[P(\mathbf{D})/P(\mathbf{D} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})])$ between the observed distribution $P(\mathbf{D})$ and the distribution $P(\mathbf{D} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$ derived by the model.

By contrast, we propose a Kullback–Leibler measure of similarity on the outputs, or hidden states, $\mathbf{S}$, rather than on the input states. The MLE justification for this leads to minimizing the Kullback–Leibler distance $\sum_S P(\mathbf{S}) \log[P(\mathbf{S})/P(\mathbf{S} \mid \gamma)]$, where γ represents the parameters of the network.

HMMs assume a class of prior probabilities, parameterized by $\boldsymbol{\beta}$, rather than the single model that we have assumed. However, we can readily generalize our model to deal with this case by replacing $P_{\mathrm{p}}(\mathbf{S})$ by a parameterized family of distributions $P_{\mathrm{p}}(\mathbf{S} \mid \boldsymbol{\tau})$. We must now minimize the Kullback–Leibler distance between $P_{\mathrm{p}}(\mathbf{S} \mid \boldsymbol{\tau})$ and the derived distribution $P_{\mathrm{DD}}(\mathbf{S} : \gamma)$ with respect to γ and $\boldsymbol{\tau}$ simultaneously.

6 Conclusion

The goal of this note was to introduce a Bayesian approach to self-organization using prior assumptions about the signal as an organizing principle. We argued that it was a natural generalization of the criterion of maximizing mutual information assuming spatial coherence (Becker and Hinton 1992). Using our principle it should be possible to

[4] HMMs have other important properties that are not directly relevant here. For example, the functional forms of $P(\mathbf{S} \mid \boldsymbol{\beta})$ and $P(\mathbf{D} \mid \mathbf{S}, \boldsymbol{\alpha})$ are chosen to ensure that highly efficient algorithms are available to perform these computations (Paul 1990).

self-organize Bayesian theories of vision, assuming that the priors are known, the network is capable of representing the appropriate functions, and the learning algorithm converges. There will also be problems if the probability distributions of the true signal and the distractor are too similar.

If the prior is not correct then it may be possible to detect this by evaluating the goodness of the Kullback–Leibler fit after learning.[5] This suggests a strategy whereby the system increases the complexity of the priors until the Kullback–Leibler fit is sufficiently good [this is somewhat similar to an idea proposed by Mumford (1992)]. This is related to the idea of competitive priors in vision (Clark and Yuille 1990). One way to implement this would be for the prior probability itself to have a set of adjustable parameters that would enable it to adapt to different classes of scenes.

Our approach differs from standard MLE by acting on the distributions of the output variables rather than the inputs. Unlike MLE our approach will directly yield an algorithm for computing the outputs. It is still unclear, however, for what class of problems our approach is applicable. For example, it seems unlikely to work if the dimensions of the outputs is a lot lower than that of the inputs.

We proposed two variants of stochastic learning that are suitable for implementing our theory. They relate, in particular, to Becker and Hinton's approach. As a further illustration of our approach we derived elsewhere (Yuille *et al.* 1993) the filter that our criterion would give for filtering out additive gaussian noise (possibly the only analytically tractable case). This turned out to be the square root of the Wiener filter in Fourier space.

Acknowledgments

We would like to thank ARPA for an Air Force contract F49620-92-J-0466. Conversations with Dan Kersten and David Mumford were highly appreciated. We would also like to thank the reviewers for their insightful comments.

References

Atick, J. J., and Redlich, A. N. 1990. Towards a theory of early visual processing. *Neural Comp.* **2**, 308–320.

Barlow, H. B. 1993. What is the computational goal of the neocortex? In *Large Scale Neuronal Theories of the Brain*, C. Koch, ed. MIT Press, Cambridge, MA.

[5] This is reminiscent of Barlow's suspicious coincidence detectors (Barlow 1993), where we might hope to determine if two variables x and y are independent or not by calculating the Kullback–Leibler distance between the joint distribution $P(x,y)$ and the product of the individual distributions $P(x)P(y)$.

Becker, S., and Hinton, G. E. 1992. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature (London)* **355**, 161–163.

Clark, J. J., and Yuille, A. L. 1990. *Data Fusion for Sensory Information Processing Systems*. Kluwer, Boston.

Held, R. 1987. Visual development in infants. In *The Encyclopedia of Neuroscience*, Vol. 2. Birkhauser, Boston.

Hornik, K., Stinchcombe, S., and White, H. 1991. Multilayer feed-forward networks are universal approximators. *Neural Networks* **4**, 251–257.

Kersten, D., O'Toole, A. J., Sereno, M. E., Knill, D. C., and Anderson, J. A. 1987. Associative learning of scene parameters from images. *Opt. Soc. Am.* **26**, 4999–5006.

Marr, D. 1982. *Vision*. W. H. Freeman, San Francisco.

Mumford, D. 1992. *Pattern Theory: A Unifying Perspective*. Mathematics Preprint. Harvard University.

Nakayama, K., and Shimojo, S. 1987. Experiencing and perceiving visual surfaces. *Science* **257**, 1357–1363.

Paul, D. B. 1990. Speech recognition using hidden Markov models. *Lincoln Lab. J.* **3**, 41–62.

Poggio, T., Torre, V., and Koch, C. 1985. Computational vision and regularization theory. *Nature (London)* **317**, 314–319.

Ripley, B. D. 1992. Classification and clustering in spatial and image data. In *Analyzing and Modeling Data and Knowledge*, M. Schader, ed. Springer-Verlag, Berlin.

Smirnakis, S. M., and Yuille, A. L. 1994. Neural implementation of Bayesian vision theories by unsupervised learning. *CNS Conf. Proc.*, in press.

Yuille, A. L., Smirnakis, S. M., and Xu, L. 1993. Bayesian self-organization. *NIPS Conf. Proc.*

Received October 14, 1992; accepted October 31, 1994.