ARTICLE Communicated by Pietro Perona

Minimax Entropy Principle and Its Application to Texture Modeling

Song Chun Zhu
Division of Applied Mathematics, Brown University, Providence, RI 02912, U.S.A.

Ying Nian Wu
Department of Statistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.

David Mumford
Division of Applied Mathematics, Brown University, Providence, RI 02912, U.S.A.

This article proposes a general theory and methodology, called the minimax entropy principle, for building statistical models for images (or signals) in a variety of applications. This principle consists of two parts. The first is the maximum entropy principle for feature binding (or fusion): for a given set of observed feature statistics, a distribution can be built to bind these feature statistics together by maximizing the entropy over all distributions that reproduce them. The second part is the minimum entropy principle for feature selection: among all plausible sets of feature statistics, we choose the set whose maximum entropy distribution has the minimum entropy. Computational and inferential issues in both parts are addressed; in particular, a feature pursuit procedure is proposed for approximately selecting the optimal set of features. The minimax entropy principle is then corrected by considering the sample variation in the observed feature statistics, and an information criterion for feature pursuit is derived. The minimax entropy principle is applied to texture modeling, where a novel Markov random field (MRF) model, called FRAME (filter, random field, and minimax entropy), is derived, and encouraging results are obtained in experiments on a variety of texture images. The relationship between our theory and the mechanisms of neural computation is also discussed.

1 Introduction

This article proposes a general theory and methodology, the minimax entropy principle, for statistical modeling in a variety of applications. This section introduces the basic concepts of the minimax entropy principle after a discussion of the motivation of our theory and a brief review of some relevant theories and methods previously studied in the literature.

1.1 Motivation and Goal. In a variety of disciplines ranging from computational vision, pattern recognition, and image coding to psychophysics, an important theme is to pursue a probability model to characterize a set of images (or signals) $\mathbf{I}$. This is often posed as a statistical inference problem: we assume that there exists a joint probability distribution (or density) $f(\mathbf{I})$ over the image space; $f(\mathbf{I})$ should concentrate on a subspace that corresponds to the ensemble of images in the application; and the objective is to estimate $f(\mathbf{I})$ given a set of observed (or training) images.

$f(\mathbf{I})$ plays significant roles in the following areas:

1. *Visual coding*, where the goal is to take advantage of the regularity or redundancy in the input images to produce a compact coding scheme. This involves measuring the efficiency of coding schemes in terms of entropy (Watson, 1987; Barlow, Kaushal, & Mitchison, 1989), where the computation of the entropy and thus the choice of the optimal coding schemes depend on the estimation of $f(\mathbf{I})$. For example, two kinds of coding schemes are compared in the recent work of Field (1994): the compact coding and the sparse coding. The former assumes gaussian distributions for $f(\mathbf{I})$, whereas the latter assumes nongaussian ones.

2. *Pattern recognition, neural networks, and statistical decision theory*, where one often needs to find a probability model $f(\mathbf{I})$ for each category of images of similar patterns. Thus, an accurate estimation of $f(\mathbf{I})$ is a key factor for successful classification and recognition.

3. *Computational vision*, where $f(\mathbf{I})$ is often adopted as a prior model in terms of Bayesian theory, and it provides a language for visual computation ranging from images segmentation to scene understanding (Zhu, 1996).

4. *Texture modeling*, where the objective is to estimate $f(\mathbf{I})$ by a probability model $p(\mathbf{I})$ for each set of texture images that have perceptually similar texture appearances. $p(\mathbf{I})$ is important not only for texture analysis such as texture segmentation and classification, but also plays a role in texture synthesis since texture images can be synthesized by sampling $p(\mathbf{I})$. Furthermore, the nature of the texture model helps us understand the mechanisms of human texture perception (Julesz, 1995).

However, making inferences about $f(\mathbf{I})$ is much more challenging than many of the learning problems in neural modeling (Dayan, Hinton, Neal, & Zernel, 1995; Xu, 1995) for the following reasons. First, the dimension of the image space is overwhelmingly large compared with the number of available training examples. In texture modeling, for instance, the size of images is often about 200×200 pixels, and thus $f(\mathbf{I})$ is a function of $40,000$ variables, whereas we have access to only one or a few training images. This make it inappropriate to use nonparametric inference methods, such

as kernel methods, radial basis functions (Ripley, 1996), and mixture of gaussian models (Jordan & Jacobs, 1994).

Second, $f(\mathbf{I})$ is often far from being gaussian; therefore some popular dimension-reduction techniques, such as the principal component analysis (Jolliffe, 1986) and spectral analysis (Priestley, 1981), do not appear to be directly applicable. As an illustration of the nongaussian property, Figure 1a shows the empirical marginal distribution (or histogram) of the intensity differences of horizontally adjacent pixels of some natural images (Zhu & Mumford, 1997). As a comparison, the gaussian distribution with the same mean and variance is plotted as a dashed curve in Figure 1a. Similar nongaussian properties are also observed in Field (1994). Another example is shown in Figure 1b, where the solid curve is the histogram of $F * \mathbf{I}$, with $\mathbf{I}$ being a texton image shown in Figure 8a, and F is a filter with the same texton (see section 4.5 for details). It is clear that the solid curve is far from being gaussian, and as a comparison, the dotted curve in Figure 1b is the histogram of $F * \mathbf{I}$, with $\mathbf{I}$ being a white noise image. The outliers in the histogram are perceptual features, not noise!

1.2 Previous Methods. A key issue in building a statistical model is the balance between generality and simplicity. The model should include rich structures to describe real-world images adequately and should be capable of modeling complexities due to high dimensionality and nongaussian property, and at the same time, it should be simple enough to be computationally feasible and give simple explanation to what we observe. To reduce complexity, it is often necessary to impose structures on the distribution. In the past, two main methods have been adopted in applications.

The first method adopts some parametric Markov random field (MRF) models in the forms of Gibbs distributions—for example, the general smoothness models in image restoration (Geman & Geman, 1984; Mumford & Shah, 1989) and the conditional autoregression models in texture modeling (Besag, 1973; Cross & Jain, 1983). This method involves only a small number of parameters and thus constructs concise distributions for images. However, they do not achieve adequate generality for the following reasons. First, these MRF models can afford only small cliques; otherwise the number of parameters will explode. But these small cliques can hardly capture image features at relatively large scales. Second, the potential functions are of very limited and prespecified forms, whereas in practice it is often desirable for the forms of the distributions to be determined or learned from the observed images.

The second method is widely used in visual coding and image reconstruction, where the high-dimensionality problem is avoided by representing the images with a relatively small set of feature statistics, and the latter are usually extracted by a set of well-selected filters. Examples of filters include the frequency and orientation selective Gabor filters (Daugman, 1985) and some wavelet pyramids based on various coding criteria (Mallat, 1989;

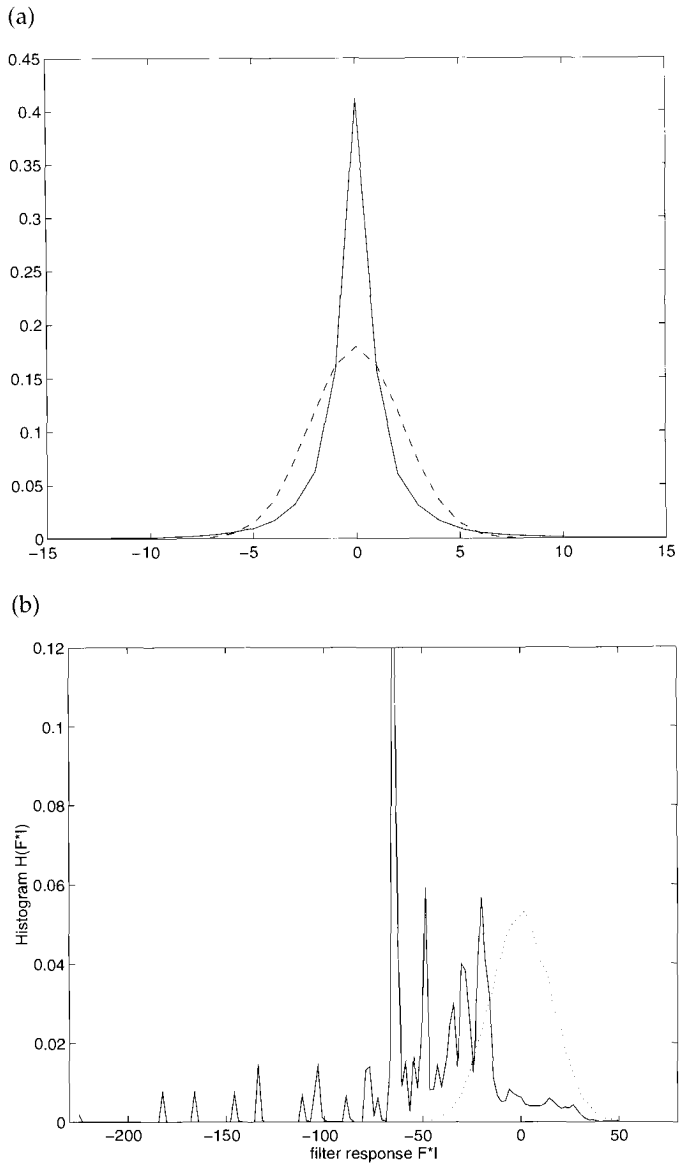

Figure 1: (a) The histogram of intensity difference at adjacent pixels and gaussian curve (dashed) of same mean and variance in domain $[-15, 15]$. (b) Histogram of the filtered texton image (solid curve) and a filtered noise image (dotted curve).

Simoncelli, Freeman, Adelson, & Weeger, 1992; Coifman & Wickerhauser, 1992; Donoho & Johnstone, 1994). The feature statistics extracted by a certain filter is usually the overall histogram of filtered images. These histograms are used for pattern classification, recognition, and visual coding (Watson, 1987; Donoho & Johnstone, 1994). Despite the excellent performances of this method, there are two major problems yet to be solved. The first is the feature binding or feature fusion problem: given a set of filters and their histograms, how to integrate them into a single probability distribution. This problem becomes much more difficult if the filters used are not all linear and are not independent of each other. The second problem is feature selection: for a given model complexity, how to choose a set of filters or features to characterize best the images being modeled.

1.3 Our Theory and Methodology. In this article, a minimax entropy principle is proposed for building statistical models, and it provides a new strategy to balance between model generality and model simplicity by two seemingly contrary criteria: maximizing entropy and minimizing entropy.

(I). The Maximum Entropy Principle (Jaynes 1957). Without loss of generality, any image feature can be expressed as $\phi^{(\alpha)}(\mathbf{I})$, where $\phi^{(\alpha)}(\ )$ can be a vector-valued functions of the image intensities and α is the index of the features. The statistic of the feature $\phi^{(\alpha)}(\mathbf{I})$ is $E_f[\phi^{(\alpha)}(\mathbf{I})]$, which is the expectation of $\phi^{(\alpha)}(\mathbf{I})$ with respect to $f(\mathbf{I})$ and is estimated by the sample mean computed from the training images. Then, given a set of features $S = \{\phi^{(\alpha)}, \alpha = 1, 2, \ldots, K\}$, a model $p(\mathbf{I})$ is constructed such that it reproduces the feature statistics as observed; that is, $E_p[\phi^{(\alpha)}(\mathbf{I})] = E_f[\phi^{(\alpha)}(\mathbf{I})]$, for $\alpha = 1, 2, \ldots, K$. Among all model $p(\mathbf{I})$ satisfying such constraints, the maximum entropy principle favors the simplest one in the sense that it has the maximum entropy. Since entropy is a measure of randomness, a maximum entropy (ME) model $p(\mathbf{I})$ is considered as the simplest fusion or binding of the features and their statistics.

(II). The Minimum Entropy Principle. The goodness of $p(\mathbf{I})$ constructed in (I) is measured by $KL(f, p)$, that is, the Kullback-Leibler divergence from $f(\mathbf{I})$ to $p(\mathbf{I})$ (Kullback & Leibler, 1951), and it depends on the feature set S that we selected. As we will show in the next section, $KL(f, p)$ is, up to a constant, equal to the entropy of $p(\mathbf{I})$. Thus, to estimate $f(\mathbf{I})$ closely, we need to minimize the entropy of the ME distribution $p(\mathbf{I})$ with respect to S, which often means that we should use as many features as possible to specify $p(\mathbf{I})$. In this sense, a minimum entropy principle favors model generality. When the model complexity or the number of features K is limited, the minimum entropy principle provides a criterion for selecting the feature set S that best characterizes $f(\mathbf{I})$.

Computational procedures are proposed for parameter estimation and

feature selection. The minimax entropy principle is further studied in the presence of sample variation of feature statistics.

As an example of application, the minimax entropy principle is applied to texture modeling, where the features are extracted by filters that are selected from a general filter bank, and the feature statistics are the empirical marginal distributions (usually further reduced to the histograms) of the filtered images. The resulting model, called FRAME (filters, random fields, and minimax entropy), is a new class of MRF model. Compared with previous MRF models, the FRAME model employs a much more enriched vocabulary and hence enjoys a much stronger descriptive ability, and at the same time, the model complexity is still under check. Texture images are synthesized by sampling the estimated models, and the correctness of estimated models is thus verified by checking whether the synthesized texture images have similar visual appearances to the observed images.

The rest of the article is arranged as follows. Section 2 studies the minimax entropy principle, where algorithms are proposed for parameters estimation and feature selection. In Section 3 we study the minimax entropy principle in depth by correcting it in the presence of estimation error and addressing the issue of variance estimation in homogeneous random fields. Section 4 applies the minimax entropy principle to texture modeling. Section 5 concludes with a discussion of the texture model and the relationship between minimax entropy and neural modeling.

2 The Minimax Entropy Principle

To fix notation, let $\mathbf{I}$ be an image defined on a domain $\mathcal{D}$; for example, $\mathcal{D}$ can be an $N \times N$ lattice, for each point $\vec{v} \in \mathcal{D}$, $\mathbf{I}(\vec{v}) \in \mathcal{L}$, and $\mathcal{L}$ is the range of image intensities. For a given application, we assume that there exists an underlying probability distribution (or density) $f(\mathbf{I})$ defined on the image space $\mathcal{L}^{|\mathcal{D}|}$, where $|\mathcal{D}|$ is the size of the image domain. Then the objective is to estimate $f(\mathbf{I})$ based on a set of observed images $\{\mathbf{I}_i^{\text{obs}}, i = 1, \ldots, M\}$ sampled from $f(\mathbf{I})$.

2.1 The Maximum Entropy Principle. At the initial stage of studying the regularity and variability of the observed images $\mathbf{I}_i^{\text{obs}}$, $i = 1, 2, \ldots, M$, one often starts from exploring a finite set of essential features that are characteristic of the observations. Without loss of generality, such features are extracted by $S = \{\phi^{(\alpha)}(\,), \alpha = 1, 2, \ldots, K\}$, where $\phi^{(\alpha)}(\mathbf{I})$ can be a vector-valued function of the image intensities. The statistics of these features are estimated by the sample means,

$$\mu_{\text{obs}}^{(\alpha)} = \frac{1}{M} \sum_{i=1}^{M} \phi^{(\alpha)}(\mathbf{I}_i^{\text{obs}}), \quad \text{for } \alpha = 1, \ldots, K.$$

If the large sample effect takes place (usually a necessary condition for modeling), then the sample averages $\{\mu_{\text{obs}}^{(\alpha)}, \alpha = 1, \ldots, K\}$ make reasonable estimates for the expectations $\{E_f[\phi^{(\alpha)}(\mathbf{I})], \alpha = 1, \ldots, K\}$, where E_f denotes the expectation with respect to $f(\mathbf{I})$. We call $\{\mu_{\text{obs}}^{(\alpha)}, \alpha = 1, \ldots, K\}$ the *observed statistics* and $\{E_f[\phi^{(\alpha)}(\mathbf{I})], \alpha = 1, \ldots, K\}$ the *expected statistics* of $f(\mathbf{I})$.

To approximate $f(\mathbf{I})$, a probability model $p(\mathbf{I})$ is restricted to reproduce the observed statistics, that is, $E_p[\phi^{(\alpha)}(\mathbf{I})] = \mu_{\text{obs}}^{(\alpha)}$ for $\alpha = 1, \ldots, K$. Let

$$\Omega_S = \{p(\mathbf{I}): E_p[\phi^{(\alpha)}(\mathbf{I})] = \mu_{\text{obs}}^{(\alpha)}, \quad \forall \phi^{(\alpha)} \in S\} \tag{2.1}$$

be the set of distributions that reproduce the observed statistics of feature set S; then we need to select a $p(\mathbf{I}) \in \Omega_S$ provided that $\Omega_S \neq \emptyset$.

As far as the observed feature statistics $\{\mu_{\text{obs}}^{(\alpha)}, \alpha = 1, \ldots, K\}$ are concerned, all the distributions in Ω_S explain them equally well, and they are not distinguishable from $f(\mathbf{I})$. The ME principle (Jaynes, 1957) suggests that we should choose $p(\mathbf{I})$ that achieves the maximum entropy to obtain the purest and simplest fusion of the observed features and their statistics. The underlying philosophy is that while $p(\mathbf{I})$ satisfies the constraints along some dimensions, it should be made as random (or smooth) as possible in other unconstrained dimensions, that is, $p(\mathbf{I})$ should represent information no more than that is available and in this sense, the ME principle is often called the minimum prejudice principle.

Thus we have the following constrained optimization problem,

$$p(\mathbf{I}) = \arg\max \left\{ -\int p(\mathbf{I}) \log p(\mathbf{I}) d\mathbf{I} \right\}, \tag{2.2}$$

subject to

$$E_p[\phi^{(\alpha)}(\mathbf{I})] = \int \phi^{(\alpha)}(\mathbf{I}) p(\mathbf{I}) d\mathbf{I} = \mu_{\text{obs}}^{(\alpha)}, \quad \alpha = 1, \ldots, K,$$

and

$$\int p(\mathbf{I}) d\mathbf{I} = 1.$$

By an application of the Lagrange multipliers, it is well known that the solution for $p(\mathbf{I})$ has the following Gibbs distribution form:

$$p(\mathbf{I}; \Lambda, S) = \frac{1}{Z(\Lambda)} \exp\left\{ -\sum_{\alpha=1}^{K} \langle \lambda^{(\alpha)}, \phi^{(\alpha)}(\mathbf{I}) \rangle \right\}, \tag{2.3}$$

where $\Lambda = (\lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(K)})$ is the parameter, $\lambda^{(\alpha)}$ is a vector of the same dimension as $\phi^{(\alpha)}(\mathbf{I})$, $\langle \cdot, \cdot \rangle$ denotes inner product, and

$$Z(\Lambda) = \int \exp\left\{ -\sum_{\alpha=1}^{K} \langle \lambda^{(\alpha)}, \phi^{(\alpha)}(\mathbf{I}) \rangle \right\} d\mathbf{I}$$

is the partition function, which normalizes $p(\mathbf{I}; \Lambda)$ into a probability distribution.

2.2 Estimation and Computation. Equation 2.3 specifies an exponential family of distributions (Brown, 1986),

$$\Theta_S = \{p(\mathbf{I}; \Lambda, S) : \Lambda \in R^d\}, \tag{2.4}$$

where d is the total number of parameters, and Λ is solved at $\hat{\Lambda}$, which satisfies the constraints $p(\mathbf{I}; \hat{\Lambda}, S) \in \Omega_S$, that is,

$$E_{p(\mathbf{I}; \hat{\Lambda}, S)}[\phi^{(\alpha)}(\mathbf{I})] = \mu_{\text{obs}}^{(\alpha)}. \quad \alpha = 1, \ldots, K. \tag{2.5}$$

However, analytical solution of equation 2.5 is in general unavailable; instead, we solve for $p(\mathbf{I}; \hat{\Lambda}, S)$ iteratively from Θ_S by maximum likelihood estimator.

Let $L(\Lambda, S) = \frac{1}{M} \sum_{i=1}^{M} \log p(\mathbf{I}_i^{\text{obs}}; \Lambda, S)$ be the log-likelihood function for any $p(\mathbf{I}; \Lambda, S) \in \Theta_S$, and it has the following properties:

$$\frac{\partial L(\Lambda, S))}{\partial \lambda^{(\alpha)}} = -\frac{1}{Z} \frac{\partial Z}{\partial \lambda^{(\alpha)}} - \mu_{\text{obs}}^{(\alpha)} = E_{p(\mathbf{I}; \Lambda, S)}[\phi^{(\alpha)}] - \mu_{\text{obs}}^{(\alpha)}, \quad \forall \alpha, \tag{2.6}$$

$$\frac{\partial^2 L(\Lambda, S)}{\partial \lambda^{(\alpha)} \lambda^{(\beta)'}} = E_{p(\mathbf{I}; \Lambda, S)}[(\phi^{(\alpha)}(\mathbf{I}) - \mu_{\text{obs}}^{(\alpha)})(\phi^{(\beta)}(\mathbf{I}) - \mu_{\text{obs}}^{(\beta)})'], \quad \forall \alpha, \beta. \tag{2.7}$$

Following equation 2.6, maximizing the log likelihood by gradient ascent gives the following equation for solving Λ iteratively:

$$\frac{d\lambda^{(\alpha)}}{dt} = E_{p(\mathbf{I}; \Lambda, S)}[\phi^{(\alpha)}(\mathbf{I})] - \mu_{\text{obs}}^{(\alpha)}, \quad \alpha = 1, \ldots, K. \tag{2.8}$$

Obviously equation 2.8 converges to $\Lambda = \hat{\Lambda}$. Moreover, equation 2.7 means that the Hessian matrix of $L(\Lambda, S)$ is the covariance matrix $(\phi^{(1)}(\mathbf{I}), \ldots, \phi^{(K)}(\mathbf{I}))$ and thus is positive definite under the condition that $a^{(0)} + \sum_{\alpha=1}^{K} a^{(\alpha)} \phi^{(\alpha)}(\mathbf{I}) \equiv 0 \Longrightarrow a^{(\alpha)} = 0$ for $\alpha = 0, \ldots, K$, which is usually satisfied. So $L(\Lambda, S)$ is strictly concave with respect to Λ, and the solution for Λ uniquely exists.

Following equations 2.5, 2.6, and 2.7, we have the following conclusion.

Proposition 1. *Given a feature set S, $\Omega_S \cap \Theta_S = \{p(\mathbf{I}; \hat{\Lambda}, S)\}$ where $\hat{\Lambda}$ is both the maximum entropy estimator and the maximum likelihood estimator.*

At each step t of equation 2.8, the computation of $E_{p(\mathbf{I}; \Lambda, S)}[\phi^{(\alpha)}(\mathbf{I})]$ is in general difficult, and we adopt the stochastic gradient method (Younes 1988) for approximation. For a fixed Λ, we synthesize some typical images

$\{\mathbf{I}_i^{\text{syn}},\ i = 1, \ldots, M'\}$ by sampling $p(\mathbf{I}; \Lambda, S)$ with the Gibbs sampler (Geman & Geman, 1984) or other Markov chain Monte Carlo (MCMC) methods (Winkler, 1995), and approximate $E_{p(\mathbf{I};\Lambda,S)}[\phi^{(\alpha)}(\mathbf{I})]$ by the sample means; that is,

$$E_{p(\mathbf{I};\Lambda,S)}[\phi^{(\alpha)}(\mathbf{I})] \approx \mu_{\text{obs}}^{(\alpha)}(\Lambda) = \frac{1}{M'}\sum_{i=1}^{M'} \phi^{(\alpha)}(\mathbf{I}_i^{\text{syn}}),\quad \alpha = 1, \ldots, K. \tag{2.9}$$

Therefore the iterative equation for computing Λ becomes

$$\frac{d\lambda^{(\alpha)}}{dt} = \Delta^{(\alpha)}(\Lambda) = \mu_{\text{syn}}^{(\alpha)}(\Lambda) - \mu_{\text{obs}}^{(\alpha)},\quad \alpha = 1, \ldots, K. \tag{2.10}$$

For the accuracy of the approximation in equation 2.9, the sample size M' should be large enough. The data flow for parameter estimation is shown in Figure 2, and the details of the algorithm can be found in (Zhu, Wu, & Mumford, 1996).

2.3 The Minimum Entropy Principle. For now, suppose that the sample size M is large enough so that the expected feature statistics $\{E_f[\phi^{(\alpha)}(\mathbf{I})], \alpha = 1, \ldots, K\}$ can be estimated exactly by neglecting the estimation errors in the observed statistics $\{\mu_{\text{obs}}^{(\alpha)}, \alpha = 1, \ldots, K\}$. Then an ME distribution $p(\mathbf{I}; \Lambda^\star, S)$ is computed so that it reproduces the expected statistics of a feature set $S = \{\phi^{(\alpha)}, \alpha = 1, 2, \ldots, K\}$; that is,

$$E_{p(\mathbf{I};\Lambda^\star,S)}[\phi^{(\alpha)}(\mathbf{I})] = E_f[\phi^{(\alpha)}(\mathbf{I})],\quad \alpha = 1, \ldots, K.$$

Since our goal is to make an inference about the underlying distribution $f(\mathbf{I})$, the goodness of this model can be measured by the Kullback-Leibler (Kullback & Leibler, 1951) divergence from $f(\mathbf{I})$ to $p(\mathbf{I}; \Lambda^\star, S)$,

$$\begin{aligned} KL(f, p(\mathbf{I}; \Lambda^\star, S)) &= \int f(\mathbf{I}) \log \frac{f(\mathbf{I})}{p(\mathbf{I}; \Lambda^\star, S)} d\mathbf{I} \\ &= E_f[\log f(\mathbf{I})] - E_f[\log p(\mathbf{I}; \Lambda^\star, S)]. \end{aligned}$$

For $KL(f, p(\mathbf{I}; \Lambda^\star, S))$, we have the following conclusion:

Theorem 1. *In the above notation,* $KL(f, p(\mathbf{I}; \Lambda^\star, S)) = entropy(p(\mathbf{I}; \Lambda^\star, S)) - entropy(f)$.

See the appendix for a proof.

In the above result, $entropy(f)$ is fixed, and $entropy(p(\mathbf{I}; \Lambda^\star, S))$ depends on the set of features S included in the distribution $p(\mathbf{I}; \Lambda^\star, S)$. Thus minimizing $KL(f, p(\mathbf{I}; \Lambda^\star, S))$ is equivalent to minimizing the entropy of $p(\mathbf{I}; \Lambda^\star, S)$. We call this the minimum entropy principle, and it has the following intuitive interpretations. First, in information theory, $p(\mathbf{I}; \Lambda^\star, S)$ defines a coding scheme with each $\mathbf{I}$ assigned a coding length $-\log p(\mathbf{I}; \Lambda^\star, S)$ (Shannon, 1948), and $entropy\,(p(\mathbf{I}; \Lambda^\star, S)) = E_p[-\log p(\mathbf{I}; \Lambda^\star, S)]$ stands for the

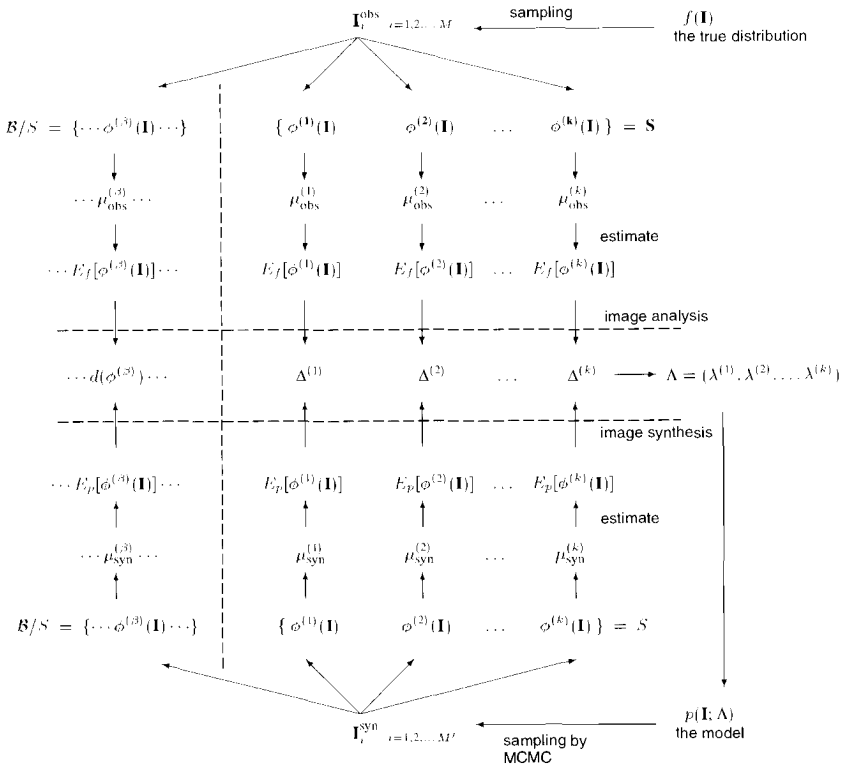

Figure 2: Data flow of the algorithm for model estimation and feature selection.

expected coding length. Therefore, a minimum entropy principle chooses the coding system with the shortest average coding length. The shortest average coding length in the actually estimated model $p(\mathbf{I}; \hat{\Lambda}, S)$ is minus its log likelihood in view of Proposition 2; hence minimizing entropy is the same as maximizing the likelihood of the data:

Proposition 2. *Given a feature set S and* $p(\mathbf{I}; \hat{\Lambda}, S)$, *then* $L(\hat{\Lambda}, S) = -$*entropy* $(p(\mathbf{I}; \hat{\Lambda}, S))$ *where* $\hat{\Lambda}$ *is the ML estimator.*

Proof. Since

$$E_{p(\mathbf{I};\hat{\Lambda},S)}[\phi^{(\alpha)}(\mathbf{I})] = \mu_{\text{obs}}^{(\alpha)}, \ \forall \phi^{(\alpha)} \in S,$$

$$L(\hat{\Lambda}, S) = \frac{1}{M} \sum_{i=1}^{M} \left\{ -\log Z(\hat{\Lambda}) - \sum_{\alpha=1}^{K} \langle \hat{\lambda}^{(\alpha)}, \phi^{(\alpha)}(\mathbf{I}_i^{\text{obs}}) \rangle \right\}$$

$$= -\log Z(\hat{\Lambda}) - \sum_{\alpha=1}^{K} \langle \hat{\lambda}^{(\alpha)}, \mu_{\text{obs}}^{(\alpha)} \rangle$$

$$= -\log Z(\hat{\Lambda}) - \sum_{\alpha=1}^{K} \langle \hat{\lambda}^{(\alpha)}, E_{p(\mathbf{I}; \hat{\Lambda})}[\phi^{(\alpha)}(\mathbf{I})] \rangle$$

$$= -entropy(p(\mathbf{I}; \hat{\Lambda}, S)).$$

However, to keep the model complexity under check, one often needs to fix the number of features K. To be precise, let $\mathcal{B}$ be the set of all possible features and $S \subset \mathcal{B}$ an arbitrary set of K features. Therefore entropy minimization provides a criterion for choosing the optimal set of features; that is,

$$S^* = \arg \min_{|S|=K} entropy(p(\mathbf{I}; \Lambda^*, S)). \tag{2.11}$$

According to the maximum entropy principle,

$$p(\mathbf{I}; \Lambda^*, S) = \arg \max_{p \in \Omega_S} entropy(p). \tag{2.12}$$

Combining equations 2.11 and 2.12, we have

$$S^* = \arg \min_{|S|=K} \{ \max_{p \in \Omega_S} entropy(p) \}. \tag{2.13}$$

We call equation 2.13 the *minimax entropy principle*. We have demonstrated that this principle is consistent with the goal of modeling: Finding the best estimate for the underlying distribution $f(\mathbf{I})$, and the relationship between minimax entropy and maximum likelihood estimator is addressed by Propositions 1 and 2.

2.4 Feature Pursuit. Enumerating all possible sets of features $S \subset \mathcal{B}$ and comparing their entropies is certainly impractical. Instead, we propose a greedy procedure to pursue the features in the following way.[1] Start from an empty feature set $\emptyset$ and $p(\mathbf{I})$ a uniform distribution, add to the model one feature at a time such that the added feature leads to the maximum decrease in the entropy of ME model $p(\mathbf{I}; \Lambda^*, S)$, and keep doing this until the entropy decrease is smaller than a certain value. To be precise, let $S = \{\phi^{(\alpha)}, \ \alpha = 1, \ldots, K\}$ be the currently selected set of features, and let

$$p = p(\mathbf{I}; \Lambda, S) = \frac{1}{Z(\Lambda)} \exp \left\{ -\sum_{\alpha=1}^{K} \langle \lambda^{(\alpha)}, \phi^{(\alpha)}(\mathbf{I}) \rangle \right\} \tag{2.14}$$

[1] We use the word *pursuit* to represent the stepwise method and distinguish it from selection.

be the ME distribution fitted to $f(\mathbf{I})$ (we omit * from Λ for notational simplicity in this subsection). For any new feature $\phi^{(\beta)} \in \mathcal{B}/S$, let $S_+ = S \cup \{\phi^{(\beta)}\}$ be a new feature set. The new ME distribution is

$$\begin{aligned} p_+ &= p(\mathbf{I}; \Lambda_+, S_+) \\ &= \frac{1}{Z(\Lambda_+)} \exp\left\{ -\sum_{\alpha=1}^{K} \langle \lambda_+^{(\alpha)}, \phi^{(\alpha)}(\mathbf{I}) \rangle - \langle \lambda_+^{(\beta)}, \phi^{(\beta)}(\mathbf{I}) \rangle \right\}. \end{aligned} \tag{2.15}$$

$E_{p_+}[\phi^{(\alpha)}(\mathbf{I})] = E_f[\phi^{(\alpha)}(\mathbf{I})]$ for $\alpha = 1, 2, \ldots, K, \beta$, and in general, $\lambda_+^{(\alpha)} \neq \lambda^{(\alpha)}$ for $\alpha = 1, \ldots, K$.

According to the above discussion, we choose feature $\phi^{(K+1)}$ to maximize the entropy decrease over the remaining features; that is,

$$\phi^{(K+1)} = \arg \max_{\phi^{(\beta)} \in \mathcal{B}/S} d(\phi^{(\beta)}),$$

where

$$d(\phi^{(\beta)}) = KL(f, p) - KL(f, p_+) = entropy(p) - entropy(p_+) = KL(p_+, p)$$

is the entropy decrease. Let $\Phi(\mathbf{I}) = (\phi^{(1)}(\mathbf{I}), \ldots, \phi^{(K)}(\mathbf{I}))$, since $E_{p_+}[\Phi(\mathbf{I})] = E_p[\Phi(\mathbf{I})] = E_f[\Phi(\mathbf{I})]$, $d(\phi^{(\beta)})$ is a function of the difference between p_+ and p on feature $\phi^{(\beta)}$. By second-order Taylor expansion, $d(\phi^{(\beta)})$ can be expressed in a quadratic form.

Proposition 3. *In the above notation,*

$$\begin{aligned} d(\phi^{(\beta)}) = \frac{1}{2} &(E_p[\phi^{(\beta)}(\mathbf{I})] - E_f[\phi^{(\beta)}(\mathbf{I})])' \\ &\times V_{p'}^{-1}(E_p[\phi^{(\beta)}(\mathbf{I})] - E_f[\phi^{(\beta)}(\mathbf{I})]), \end{aligned} \tag{2.16}$$

where p' is a distribution such that $E_{p'}[\Phi(\mathbf{I})] = E_f[\Phi(\mathbf{I})]$, and $E_{p'}[\phi^{(\beta)}(\mathbf{I})]$ lies between $E_p[\phi^{(\beta)}(\mathbf{I})]$ and $E_f[\phi^{(\beta)}(\mathbf{I})]$. $V_{p'} = V_{22} - V_{21}V_{11}^{-1}V_{12}$, where $V_{11} = Var_{p'}[\Phi(\mathbf{I})]$, $V_{22} = Var_{p'}[\phi^{(\beta)}(\mathbf{I})]$, $V_{12} = Cov_{p'}[\Phi(\mathbf{I}), \phi^{(\beta)}(\mathbf{I})]$, and $V_{21} = V_{12}'$.

See the appendix for proof.

The $V_{p'}$ can be interpreted as follows. Let $C = -V_{12}V_{11}^{-1}$, and let $\phi_\perp^{(\beta)}(\mathbf{I}) = \phi^{(\beta)}(\mathbf{I}) + C\Phi(\mathbf{I})$ be the linear combination of $\phi^{(\beta)}(\mathbf{I})$ and $\Phi(\mathbf{I})$; then under p', it can be shown that $\phi_\perp^{(\beta)}(\mathbf{I})$ is uncorrelated with $\Phi(\mathbf{I})$, thus $V_{p'} = Var_{p'}[\phi_\perp^{(\beta)}(\mathbf{I})]$ is the variance of $\phi^{(\beta)}(\mathbf{I})$ with its dependence on $\Phi(\mathbf{I})$ being eliminated.

In practice, $E_f[\phi^{(\beta)}(\mathbf{I})]$ is estimated by the observed statistic $\mu_{\text{obs}}^{(\beta)}$, and $E_p[\phi^{(\beta)}(\mathbf{I})]$ by $\mu_{\text{syn}}^{(\beta)}$—the sample mean computed from synthesized images

sampled from the current model p. If the intermediate distribution p' is approximated using the current distribution p, the distance $d(\phi^{(\beta)})$ is approximated by

$$d(\phi^{(\beta)}) \approx \frac{1}{2}\left(\mu_{\text{obs}}^{(\beta)} - \mu_{\text{syn}}^{(\beta)}\right)' V_p^{-1}\left(\mu_{\text{obs}}^{(\beta)} - \mu_{\text{syn}}^{(\beta)}\right), \tag{2.17}$$

where $V_p = Var_p(\phi_\perp^{(\beta)})$ is the variance estimated from the synthesized images. We will further study the estimation of V_p in section 3.2.

The feature pursuit procedure governed by equation 2.17 has the following intuitive interpretation. Under the current model p, for any new feature $\phi^{(\beta)}$, $\mu_{\text{syn}}^{(\beta)}$ is the statistic we observe from the image samples following p. If $\mu_{\text{syn}}^{(\beta)}$ is close to $\mu_{\text{obs}}^{(\beta)}$, then adding this new feature to $p(\mathbf{I}; \Lambda)$ leads to little improvement in estimating $f(\mathbf{I})$. So we should look for the most salient new feature $\phi^{(\beta)}$ such that $\mu_{\text{syn}}^{(\beta)}$ is very different from $\mu_{\text{obs}}^{(\beta)}$. The saliency of the new feature is measured by $d(\phi^{(\beta)})$, which is the discrepancy between $\mu_{\text{syn}}^{(\beta)}$ and $\mu_{\text{obs}}^{(\beta)}$ scaled by V_p, where V_p is the variance of the new feature compensated for dependence of the new feature on the old ones under the current model.

As a summary, Figure 2 illustrates the data flow for both the computation of the model and the pursuit of features.

3 More on Minimax Entropy

3.1 Correcting the Minimum Entropy Principle. In previous sections, for a set of features $S = \{\phi^{(\alpha)}, \alpha = 1, \ldots, K\}$, we have studied two ME distributions. One is $p(\mathbf{I}; \hat{\Lambda}, S)$, which reproduces the observed feature statistics, that is,

$$E_{p(\mathbf{I}; \hat{\Lambda}, S)}[\phi^{(\alpha)}(\mathbf{I})] = \mu_{\text{obs}}^{(\alpha)}, \quad \text{for } \alpha = 1, \ldots, K,$$

and the other is $p(\mathbf{I}; \Lambda^*, S)$, which reproduces the expected feature statistics, that is,

$$E_{p(\mathbf{I}; \Lambda^*, S)}[\phi^{(\alpha)}(\mathbf{I})] = E_f[\phi^{(\alpha)}(\mathbf{I})], \quad \text{for } \alpha = 1, \ldots, K.$$

In the previous derivations, we assume that $\{E_f[\phi^{(\alpha)}(\mathbf{I})], \alpha = 1, \ldots, K\}$ can be estimated exactly by the observed statistics $\{\mu_{\text{obs}}^{(\alpha)}, \alpha = 1, \ldots, K\}$, which is not true in practice since only a finite sample is observed. Taking the estimation errors into account, we need to correct the minimum entropy principle and the feature pursuit procedure.

First, let us consider the minimum entropy principle, which relates the Kullback-Leibler divergence $KL(f, p(\mathbf{I}; \Lambda, S))$ to the entropy of the model $p(\mathbf{I}; \Lambda, S)$ for $\Lambda = \Lambda^*$. Since in practice Λ is estimated at $\hat{\Lambda}$, the goodness of the actual model should be measured by $KL(f, p(\mathbf{I}; \hat{\Lambda}, S))$ instead of $KL(f, p(\mathbf{I}; \Lambda^*, S))$, for which we have:

Proposition 4. *In the above notation,*

$$KL(f, p(\mathbf{I}; \hat{\Lambda}, S)) = KL(f, p(\mathbf{I}; \Lambda^*, S)) + KL(p(\mathbf{I}; \Lambda^*, S), p(\mathbf{I}; \hat{\Lambda}, S)). \quad (3.1)$$

See the appendix for proof.

That is, because of the estimation error, $p(\mathbf{I}; \hat{\Lambda}, S)$ does not come as close to $f(\mathbf{I})$ as $p(\mathbf{I}; \Lambda^*, S)$ does, and the extra noise is measured by $KL(p(\mathbf{I}; \Lambda^*, S), p(\mathbf{I}; \hat{\Lambda}, S))$. In fact, $\hat{\Lambda}$ in model $p(\mathbf{I}; \hat{\Lambda}, S)$ is a random variable depending on the random sample $\{\mathbf{I}_i^{\text{obs}}, i = 1, \ldots, M\}$; so is $KL(f, p(\mathbf{I}; \hat{\Lambda}, S))$. Let E_{obs} stands for the expectation with respect to the training images. Applying E_{obs} to both sides of equation 3.1, we have,

$$\begin{aligned} & E_{\text{obs}}[KL(f, p(\mathbf{I}; \hat{\Lambda}, S))] \\ & = KL(f, p(\mathbf{I}; \Lambda^*, S)) + E_{\text{obs}}[KL(p(\mathbf{I}; \Lambda^*, S), p(\mathbf{I}; \hat{\Lambda}, S))] \\ & = entropy(p(\mathbf{I}; \Lambda^*, S)) - entropy(f) \\ & \quad + E_{\text{obs}}[KL(p(\mathbf{I}; \Lambda^*, S), p(\mathbf{I}; \hat{\Lambda}, S))]. \end{aligned} \quad (3.2)$$

The following proposition relates $entropy(p(\mathbf{I}; \Lambda^*, S))$ to $entropy(p(\mathbf{I}; \hat{\Lambda}, S))$.

Proposition 5. *In the above notation,*

$$\begin{aligned} entropy(p(\mathbf{I}; \Lambda^*, S)) = & E_{\text{obs}}[entropy(p(\mathbf{I}; \hat{\Lambda}, S))] \\ & + E_{\text{obs}}[KL(p(\mathbf{I}; \hat{\Lambda}, S), p(\mathbf{I}; \Lambda^*, S))]. \end{aligned} \quad (3.3)$$

See the appendix for proof.

According to Proposition 5, the entropy of $p(\mathbf{I}; \hat{\Lambda}, S)$ is on average smaller than the entropy of $p(\mathbf{I}; \Lambda^*, S)$; this is because $\hat{\Lambda}$ is estimated from specific training data, and hence $p(\mathbf{I}; \hat{\Lambda}, S)$ does a better job than $p(\mathbf{I}; \Lambda^*, S)$ in fitting the training data.

Combining equations 3.2 and 3.3, we have

$$\begin{aligned} E_{\text{obs}}[KL(f, p(\mathbf{I}; \hat{\Lambda}, S))] = & E_{\text{obs}}[entropy(p(\mathbf{I}; \hat{\Lambda}, S))] \\ & - entropy(f) + C_1 + C_2, \end{aligned} \quad (3.4)$$

where the two correction terms are

$$C_1 = E_{\text{obs}}[KL(p(\mathbf{I}; \Lambda^*, S), p(\mathbf{I}; \hat{\Lambda}, S))],$$

$$C_2 = E_{\text{obs}}[KL(p(\mathbf{I}; \hat{\Lambda}, S), p(\mathbf{I}; \Lambda^*, S))].$$

Following Ripley (1996, sec. 2.2), both C_1 and C_2 can be approximated by

$$\frac{1}{2M} tr(Var_f[\Phi(\mathbf{I})] Var_{p^*}^{-1}[\Phi(\mathbf{I})]) + O(M^{-3/2}),$$

where $tr(\ )$ is the trace of matrix. Therefore, we arrive at the following form of the Akaike information criterion (Akaike, 1977):

$$E_{\text{obs}}[KL(f, p(\mathbf{I}; \hat{\Lambda}, S))] \approx E_{\text{obs}}[entropy(p(\mathbf{I}; \hat{\Lambda}, S))] - entropy(f) + \frac{1}{M} tr(Var_f[\Phi(\mathbf{I})] Var_{p^*}^{-1}[\Phi(\mathbf{I})]),$$

where we drop the higher-order term $O(M^{-3/2})$. The optimal set of features should be chosen to minimize $E_{\text{obs}}[KL(f, p(\mathbf{I}; \hat{\Lambda}, S))]$, which leads to the following correction of the minimum entropy principle:

$$S^* = \arg\min_{|S|=K} \{entropy(p(\mathbf{I}; \hat{\Lambda}, S)) + \frac{1}{M} tr(Var_f[\Phi(\mathbf{I})] Var_{p^*}^{-1}[\Phi(\mathbf{I})])\}. \quad (3.5)$$

In practice, $Var_f[\Phi(\mathbf{I})]$ and $Var_{p^*}[\Phi(\mathbf{I})]$ can be estimated from the observed images and synthesized images, respectively. If $Var_f[\Phi(\mathbf{I})] \approx Var_{p^*}[\Phi(\mathbf{I})]$, then $tr(Var_f[\Phi(\mathbf{I})] Var_{p^*}^{-1}[\Phi(\mathbf{I})])$ is approximately the number of free parameters in the model. This provides another reason for restricting the model complexity besides scientific parsimony and computational efficiency. Another perspective for this issue is the minimum description length (MDL) principle (Rissanen, 1989).

Now let us consider correcting the feature pursuit procedure. Following the notation in section 2.4, at each step $K+1$, suppose we choose a new feature $\phi^{(\beta)}$, and let $\Phi_+(\mathbf{I}) = (\Phi(\mathbf{I}), \phi^{(\beta)}(\mathbf{I}))$; the decrease of the expected Kullback-Leibler divergence is:

$$\begin{aligned} & E_{\text{obs}}[KL(f, p)] - E_{\text{obs}}[KL(f, p_+)] \\ &= d(\phi^{(\beta)}) - \frac{1}{M}[tr(Var_f[\Phi_+(\mathbf{I})] Var_{p_+^*}^{-1}[\Phi_+(\mathbf{I})]) \\ &\qquad - tr(Var_f[\Phi(\mathbf{I})] Var_{p^*}^{-1}[\Phi(\mathbf{I})])]. \end{aligned}$$

By linear algebra, we can show that

$$\begin{aligned} & tr(Var_f[\Phi_+(\mathbf{I})] Var_{p_+^*}^{-1}[\Phi_+(\mathbf{I})]) - tr(Var_f[\Phi(\mathbf{I})] Var_{p^*}^{-1}[\Phi(\mathbf{I})]) \\ &= tr(Var_f[\phi_\perp^{(\beta)}(\mathbf{I})] Var_{p_\perp^*}[\phi_\perp^{(\beta)}(\mathbf{I})]). \end{aligned} \quad (3.6)$$

See the appendix for the proof of equation 3.6.

Therefore, at every step of the (corrected) feature pursuit procedure, we should choose $\phi^{(\beta)}$ to maximize

$$d'(\phi^{(\beta)}) = d(\phi^{(\beta)}) - \frac{1}{M} tr\left(Var_f[\phi_\perp^{(\beta)}] Var_{p_+^*}^{-1}[\phi_\perp^{(\beta)}]\right).$$

In practice, we approximate $Var_{p_+^*}[\phi_\perp^{(\beta)}]$ by $Var_p[\phi_\perp^{(\beta)}]$, and estimate the variances from the observed and synthesized images. Let $\mu_{\perp obs}^{(\beta)}$ and $\hat{V}_{\text{obs}}$ be

the sample mean and variance of $\{\phi_\perp^{(\beta)}(\mathbf{I}_i^{\text{obs}}), i = 1, 2, \ldots, M\}$ and let $\hat{V}_{\text{syn}}$ be the sample variance of $\{\phi_\perp^{(\beta)}(\mathbf{I}_i^{\text{syn}}), i = 1, 2, \ldots, M'\}$. Thus we have

$$d'(\phi^{(\beta)}) \approx \frac{1}{2}(\mu_{\text{syn}}^{(\beta)} - \mu_{\text{obs}}^{(\beta)})' \hat{V}_{\text{syn}}^{-1}(\mu_{\text{syn}}^{(\beta)} - \mu_{\text{obs}}^{(\beta)}) - \frac{1}{M} tr(\hat{V}_{\text{obs}}\hat{V}_{\text{syn}}^{-1}). \quad (3.7)$$

We note that in equation 3.7,

$$\begin{aligned} tr(\hat{V}_{\text{obs}}\hat{V}_{\text{syn}}^{-1}) &= \frac{1}{M} tr\left(\sum_{i=1}^{M}(\phi_\perp^{(\beta)}(\mathbf{I}_i^{\text{obs}}) - \mu_{\perp obs}^{(\beta)})(\phi_\perp^{(\beta)}(\mathbf{I}_i^{\text{obs}}) - \mu_{\perp obs}^{(\beta)})'\hat{V}_{\text{syn}}^{-1}\right) \\ &= \frac{1}{M}\sum_{i=1}^{M}(\phi_\perp^{(\beta)}(\mathbf{I}_i^{\text{obs}}) - \mu_{\perp obs}^{(\beta)})'\hat{V}_{\text{syn}}^{-1}(\phi_\perp^{(\beta)}(\mathbf{I}_i^{\text{obs}}) - \mu_{\perp obs}^{(\beta)}) \end{aligned}$$

is a measure of fluctuation in the observed images.

The intuitive meaning of equation 3.7 is the following. The first term is the distance between $\mu_{\text{syn}}^{(\beta)}$ and $\mu_{\text{obs}}^{(\beta)}$, and we call it the *gain* by introducing a new feature $\phi^{(\beta)}$. The second term measures the uncertainty in estimating $E_f[\phi^{(\beta)}(\mathbf{I})]$, and we call it the *loss* by adding $\phi^{(\beta)}$. If the loss term is large, it means the feature is less common to the observed images; thus $d'(\phi^{(\beta)})$ is small. When $\mu_{\text{syn}}^{(\beta)}$ comes very close to $\mu_{\text{obs}}^{(\beta)}$, $d'(\phi^{(\beta)})$ become negative, which provides a criterion for stopping the iteration in computing Λ in equation 2.10.

3.2 Variance Estimation in Homogeneous Random Field. In previous sections, we assume that we have M independent observations $\mathbf{I}_i^{\text{obs}}, i = 1, 2, \ldots, M$, and each $\mathbf{I}_i^{\text{obs}}$ is of the same size as the image domain $\mathcal{D}$ ($N \times N$ pixels). A feature $\phi^{(\beta)}(\mathbf{I}_i^{\text{obs}})$ is computed based on the intensities of an entire image, and the sample mean and variance are then computed from $\phi^{(\beta)}(\mathbf{I}_i^{\text{obs}})$ $i = 1, 2, \ldots, M$. The same is true for synthesized images. However, in many applications, such as texture modeling in the next section, it is assumed that the underlying distribution $f(\mathbf{I})$ is ergodic and images $\mathbf{I}$ are homogeneous; thus, $\phi^{(\beta)}(\mathbf{I})$ is often expressed as the average of local features $\psi^{(\beta)}(\ )$:

$$\phi^{(\beta)}(\mathbf{I}) = \frac{1}{|\mathcal{D}|}\sum_{v\in\mathcal{D}}\psi^{(\beta)}(\mathbf{I}|_{W+\vec{v}}),$$

where $\psi^{(\beta)}$ is a function defined on locally supported (filter) windows W centered at $\vec{v} \in \mathcal{D}$. Therefore by ergodicity we still estimate $E_f[\phi^{(\beta)}(\mathbf{I})]$ and $Var_f[\phi^{(\beta)}(\mathbf{I})]$ reasonably well even through only a single image is observed, provided that the observed image is large enough compared to the strength of autocorrelation.

In particular, we adopt the method recently proposed by Sherman (1996) for the estimation of $Var_f[\phi^{(\beta)}(\mathbf{I})]$. To fix notation, suppose we observe one

image $\mathbf{I}^{\text{obs}}$ on an $N_o \times N_o$ lattice $\mathcal{D}_{\text{obs}}$, and we define subdomains $\mathcal{D}_m^j \subset \mathcal{D}_{\text{obs}}$, $j = 1, 2, \ldots, \ell$. For simplicity, we assume all subdomains are square image patches of the same size of $m \times m$ pixels; m is usually chosen at the scale of $\sqrt{N_o}$, and the subdomains may overlap each other. Then for each subdomain $\mathcal{D}_m^j$, we compute $\phi^{(\beta)}(\mathcal{D}_m^j) = \frac{1}{m^2} \sum_{\vec{v} \in \mathcal{D}_m^j} \psi^{(\beta)}(\mathbf{I}|_{W+\vec{v}})$, and the sample mean and variance are computed over the subdomains:

$$\mu_{\text{obs}}^{(\beta)}(\mathcal{D}_m) = \frac{1}{\ell} \sum_{j=1}^{\ell} \phi^{(\beta)}(\mathcal{D}_m^j),$$

$$Var_{\text{obs}}^{(\beta)}(\mathcal{D}_m) = \frac{1}{\ell} \sum_{j=1}^{\ell} (\phi^{(\beta)}(\mathcal{D}_m^j) - \mu_{\text{obs}}^{(\beta)}(\mathcal{D}_m))(\phi^{(\beta)}(\mathcal{D}_m^j) - \mu_{\text{obs}}^{(\beta)}(\mathcal{D}_m))'.$$

Then, according to Sherman (1996), $Var_f[\phi^{(\beta)}(\mathbf{I})]$ can be estimated by

$$\frac{m^2}{N^2} Var_{\text{obs}}^{(\beta)}(\mathcal{D}_m).$$

Now let us consider the feature pursuit criterion in equation 3.7. For feature $\phi_\perp^{(\beta)}$ we define variance $\hat{V}_{\text{obs}}(\mathcal{D}_1) = m^2 \hat{V}_{\text{obs}}(\mathcal{D}_m)$, where $\hat{V}_{\text{obs}}(\mathcal{D}_m)$ is the sample variance of $\phi_\perp^{(\beta)}(\mathcal{D}_m^j)$, $j = 1, 2, \ldots, \ell$. Then from the above result, we can approximate $\hat{V}_{\text{obs}}$ in equation 3.7 by $\hat{V}_{\text{obs}}(\mathcal{D}_1)/N^2$. Similarly $\hat{V}_{\text{syn}}$ in equation 3.7 is replaced by $\hat{V}_{\text{syn}}(\mathcal{D}_1)/N^2$. Thus we have

$$d'(\phi^{(\beta)}) \approx N^2 \left[\frac{1}{2} (\mu_{\text{syn}}^{(\beta)} - \mu_{\text{obs}}^{(\beta)})' \hat{V}_{\text{syn}}^{-1}(\mathcal{D}_1)(\mu_{\text{syn}}^{(\beta)} - \mu_{\text{obs}}^{(\beta)}) - \frac{1}{c} tr(\hat{V}_{\text{obs}}(\mathcal{D}_1) \hat{V}_{\text{syn}}^{-1}(\mathcal{D}_1)) \right], \tag{3.8}$$

where c is $|\mathcal{D}_{\text{obs}}|$ minus the number of pixels around the boundary. From equation 3.8, we notice that $d'(\phi^{(\beta)})$ is proportional to N^2—the size of domain $\mathcal{D}$.

A more rigorous study is often complicated by phase transition, and we shall not pursue it in this article.

4 Application to Texture Modeling

This section applies the minimax entropy principle to texture modeling.

4.1 The General Problem. Texture is an important characteristic of surface property in visual scenes and a power cue in visual perception. A general model for textures has long been sought in both computational vision and psychology, but such a model is still far from being achieved because

of the vast diversity of the physical and chemical processes that generate textures and the large number of attributes that need to be considered. As an illustration of the diversity of textures, Figure 3 displays some typical texture images.

Existing models for textures can be roughly classified into three categories: (1) dynamic equations or replacement rules, which simulate specific physical and chemical processes to generate textures (Witkin & Kass, 1991; Picard, 1996), (2) the kth-order statistics model for texture perception, that is, the famous Julesz's conjecture (Julesz, 1962), and (3) MRF models. (For a discussion of previous models and methods, see Zhu et al., 1996.)

In our method, a texture is considered an ensemble of images of similar texture appearances governed by a probability distribution $f(\mathbf{I})$. As discussed in section 2, we seek a model $p(\mathbf{I}; \Lambda, S)$ given a set of observed images. $p(\mathbf{I}; \Lambda, S)$ should be consistent with human texture perception in the sense that if $p(\mathbf{I}; \Lambda, S)$ estimates $f(\mathbf{I})$ closely, the images sampled from $p(\mathbf{I}; \Lambda, S)$ should be perceptually similar to the training images.

4.2 Choosing Features and Their Statistics. As the first step of applying the minimax entropy principle, we need to choose image features and their statistics, that is, $\phi^{(\alpha)}(\mathbf{I})$ and $\mu_{\text{obs}}^{(\alpha)}$ $\alpha = 1, 2, \ldots, K$.

First, we limit our model to homogeneous textures; thus $f(\mathbf{I})$ is stationary with respect to location $\vec{v}$. We assume that features of texture images can be extracted by "filters" $F^{(\alpha)}$, $\alpha = 1, 2, \ldots, K$, where $F^{(\alpha)}$ can be a linear or nonlinear function of the intensities of the image $\mathbf{I}$. Let $\mathbf{I}^{(\alpha)}(\vec{v})$ denote the filter response at point $\vec{v} \in \mathcal{D}$, that is, $\mathbf{I}^{(\alpha)}(\vec{v}) = F^{(\alpha)}(\mathbf{I}|_{W+\vec{v}})$ is a function depending on the intensities inside window W centered at $\vec{v}$.

Second, recent psychophysical research on human texture perception suggests that two homogeneous textures are often difficult to discriminate when they produce similar marginal distributions (histograms) of responses from a bank of filters (Bergen & Adelson, 1991; Chubb & Landy, 1991). Motivated by the psychophysical research, we make the following assumptions to limit the number of filters and the window size of each filter for computational reason, though these assumptions are not necessary conditions for our theory to hold true:

1. All features that concern texture perception can be captured by "locally" supported filters. By "locally" we mean that the sizes of filters should be much smaller than the size of the image. For example, the size of image is 256×256 pixels, and the window sizes of filters are limited to be less than 33×33 pixels.

2. Only a finite set of filters are used.

Given a filter $F^{(\alpha)}$, we compute the histogram of the filtered image $\mathbf{I}^{(\alpha)}$ as the features of $\mathbf{I}$. Therefore in texture modeling, the notation $\phi^{(\alpha)}(\mathbf{I})$ is

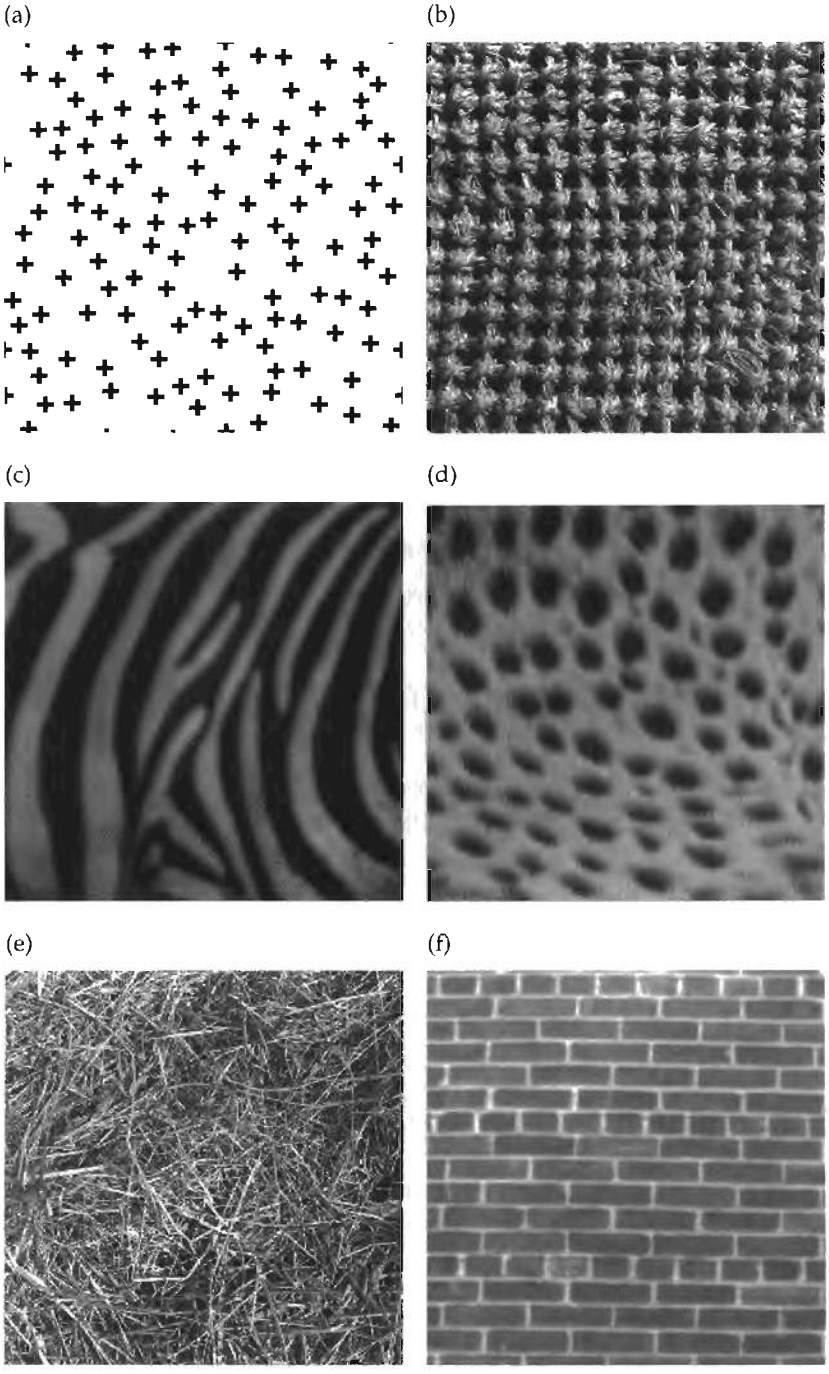

Figure 3: Some typical texture images.

replaced by

$$H^{(\alpha)}(\mathbf{I}, z) = \frac{1}{|\mathcal{D}|} \sum_{\vec{v} \in \mathcal{D}} \delta(z - \mathbf{I}^{(\alpha)}(\vec{v})), \qquad \alpha = 1, 2, \ldots, K, \quad z \in \mathbf{R}$$

where $\delta(\ )$ is the Dirac point mass function concentrated at 0. Correspondingly the observed statistics $\mu_{\text{obs}}^{(\alpha)}$ are defined as

$$\mu_{\text{obs}}^{(\alpha)}(z) = \frac{1}{M} \sum_{i=1}^{M} H^{(\alpha)}(\mathbf{I}_i^{\text{obs}}, z), \qquad \alpha = 1, 2, \ldots, K.$$

$H^{(\alpha)}(\mathbf{I}, z)$ and $\mu_{\text{obs}}^{(\alpha)}(z)$ are, in theory, continuous functions of z.[2] In practice, they are approximated by piecewise constant functions of a finite number L of bins and are denoted by $H^{(\alpha)}(\mathbf{I})$ and $\mu_{\text{obs}}^{(\alpha)}$ as L (e.g., $L = 32$) dimensional vectors in the rest of the article.

As the sample size M is large or the images $\mathbf{I}_i^{\text{obs}}$ are large so that the large sample effect takes place by ergodicity, then $\mu_{\text{obs}}^{(\alpha)}(z)$ will be a close estimate of the marginal distributions of $f(\mathbf{I})$:

$$f^{(\alpha)}(z) = E_f[H^{(\alpha)}(\mathbf{I}, z)].$$

Another motivation for choosing $\mu_{\text{obs}}^{(\alpha)}(z)$ as feature statistics comes from a mathematical theorem, which states that $f(\mathbf{I})$ is determined by all its marginal distributions $f^{(\alpha)}(z)$. Thus, if model $p(\mathbf{I})$ reproduces $f^{(\alpha)}(z)$ for all α, then $p(\mathbf{I}) = f(\mathbf{I})$ (Zhu et al., 1996).

Substituting $H^{(\alpha)}(\mathbf{I})$ for $\phi^{(\alpha)}(\mathbf{I})$ in equation 2.3, we obtain

$$p(\mathbf{I}; \Lambda, S) = \frac{1}{Z(\Lambda)} \exp\left\{ - \sum_{\alpha=1}^{K} \langle \lambda^{(\alpha)}, H^{(\alpha)}(\mathbf{I}) \rangle \right\}, \tag{4.1}$$

which we call the FRAME model. Here the angle brackets indicate that we are taking a sum over bin z: that is, $\langle \lambda^{(\alpha)}, H^{(\alpha)}(\mathbf{I}) \rangle = \sum_z \lambda_z^{(\alpha)} H^{(\alpha)}(\mathbf{I}, z)$.

The computation of the parameters Λ and the selection of filters $F^{(\alpha)}$ proceed as described in the last section. For detailed analysis of the texture modeling algorithm, see Zhu et al. (1996).

4.3 FRAME: A New Class of MRF Models. In this section, we derive a continuous form for the FRAME model in equation 4.1, and compare it with existing MRF models.

[2] Compared with the definitions of $\phi^{(\alpha)}(\mathbf{I})$ and $\mu_{\text{obs}}^{(\alpha)}$, $H^{(\alpha)}(\mathbf{I}, z)$ and $\mu_{\text{obs}}^{(\alpha)}(z)$ are considered vectors of infinite dimensions.

Since the histograms of an image are continuous functions, the constraint in the ME optimization problem is the following:

$$E_p\left[\frac{1}{|\mathcal{D}|}\sum_{\vec{v}\in\mathcal{D}}\delta(z-\mathbf{I}^{(\alpha)}(\vec{v}))\right]=\mu_{\text{obs}}^{(\alpha)}(z),\quad \forall z\in\mathbf{R},\ \forall\vec{v}\in\mathcal{D},\ \forall\alpha. \tag{4.2}$$

By an application of Lagrange multipliers, maximizing the entropy of $p(\mathbf{I})$ under the above constraints gives

$$\begin{aligned} p(\mathbf{I};\Lambda,S) &= \frac{1}{Z(\Lambda)}\exp\left\{-\sum_{\alpha=1}^{K}\sum_{\vec{v}\in\mathcal{D}}\int\lambda^{(\alpha)}(z)\frac{1}{|\mathcal{D}|}\sum_{\vec{v}\in\mathcal{D}}\delta(z-\mathbf{I}^{(\alpha)}(\vec{v}))dz\right\} \\ &= \frac{1}{Z(\Lambda)}\exp\left\{-\sum_{\alpha=1}^{K}\sum_{\vec{v}\in\mathcal{D}}\lambda^{(\alpha)}(\mathbf{I}^{(\alpha)}(\vec{v}))\right\}. \end{aligned} \tag{4.3}$$

Since z is a continuous variable, there is an infinite number of constraints. The Lagrange multipliers $\Lambda=(\lambda^{(1)}(\ ),\ldots,\lambda^{(K)}(\ ))$ take the form as one-dimensional potential functions. More specifically when the filters are linear, $\mathbf{I}^{(\alpha)}(\vec{v})=F^{(\alpha)}*\mathbf{I}(\vec{v})$, and we can rewrite equation 4.3 as,

$$p(\mathbf{I};\Lambda,S)=\frac{1}{Z(\Lambda)}\exp\left\{-\sum_{\alpha=1}^{K}\sum_{\vec{v}}\lambda^{(\alpha)}(F^{(\alpha)}*\mathbf{I}(\vec{v}))\right\}. \tag{4.4}$$

Clearly, equations 4.3 and 4.4 are MRF models or, equivalently, Gibbs distributions. But unlike the previous MRF models, the potentials are built directly on the filter response instead of cliques, and the forms of the potential functions $\lambda^{(\alpha)}(\ )$ are learned from the training images, so they can incorporate high-order statistics and thus model nongaussian properties of images. The FRAME model has much stronger expressive power than traditional clique-based MRF models. Every filter introduces the same number of L parameters regardless of its window size, which enables us to explore structures at large scales (e.g., the 33×33 pixel filters in modeling the fabric texture in section 4.5). It is easy to show that existing MRF models for texture are special cases of the FRAME model with the filters and their potential functions specified. Detailed comparison between the FRAME model and the MRF models is covered in Zhu et al. (1996).

4.4 Designing a Filter Bank. To describe a wide variety of textures, we need to specify a general filter bank, which serves as the "vocabulary" by analogy to language. We shall not discuss the rules for constructing an optimal filter bank; instead, we use the following five kinds of filters motivated

by the multichannel filtering mechanism discovered and generally accepted in neurophysiology (Silverman, Grosof, De Valois, & Elfar, 1989).

1. The intensity filter, $\delta(\ )$, for capturing the DC component.
2. The Laplacian of gaussian filters, which are isotropic center surrounded and are often used to model retinal ganglion cells. The impulse response functions are of the following form:

$$LG(x, y \mid T) = const \cdot (x^2 + y^2 - T^2)e^{-\frac{x^2+y^2}{T^2}}. \tag{4.5}$$

   We choose eight scales with $T = \sqrt{2}/2, 1, 2, 3, 4, 5$, and 6. The filter with scale T is denoted by $LG(T)$.
3. The Gabor filters, which are models for the frequency and orientation-sensitive simple cells. The impulse response functions are of the following form,

$$\begin{aligned} Gabor(x, y \mid T, \theta) = const \cdot e^{\frac{1}{2T^2}(4(x\cos\theta + y\sin\theta)^2 + (-x\sin\theta + y\cos\theta)^2)} \\ \times e^{-i\frac{2\pi}{T}(x\cos\theta + y\sin\theta)}, \end{aligned} \tag{4.6}$$

   where T controls the scales and θ controls the orientations. We choose six scales $T = 2, 4, 6, 8, 10$, and 12 and six orientations $\theta = 0°, 30°, 60°, 90°, 120°$, and $150°$. Notice that these filters are not nearly orthogonal to each other, so there is overlap among the information captured by them. The sine and cosine components are denoted by $G\,sin(T, \theta)$ and $G\,cos(T, \theta)$, respectively.
4. The nonlinear Gabor filters, which are models for the complex cells, and responses from which are the powers of the responses from a pair of Gabor filters, $|\,Gabor(x, y \mid T, \theta) * I|^2$. This filter denoted by $SP(T, \theta)$ is, in fact, the local spectrum of $\mathbf{I}$ at (x, y) smoothed by a gaussian function.
5. Some specially designed filters for texton primitives. (See section 4.5.)

4.5 Experiments of Texture Modeling. This section describes the modeling of natural textures using the algorithm studied in sections 2 and 3. The first texture image is described in detail to illustrate the filter pursuit procedure.

Suppose we are modeling $f(\mathbf{I})$ where $\mathbf{I}$ is of 64×64 pixels. Figure 4a is an observed image of animal fur (128×128 pixels). We start from the filter set $S = \emptyset$ and $p(\mathbf{I}; \Lambda, S)$ a uniform distribution from which a uniform white noise image is sampled and is displayed in Figure 4b (128×128 pixels). The algorithm first computes $d'(\phi^{(1)})$ according to equations 3.7 and 3.8 for each

Table 1: The Entropy Decrease $d'(\phi^{(\beta)})$ for Filter Pursuit.

Filter	Size	$d'(\phi^{(1)})$	$d'(\phi^{(2)})$	$d'(\phi^{(3)})$	$d'(\phi^{(4)})$	$d'(\phi^{(5)})$	$d'(\phi^{(8)})$
δ	1×1	1018.2	42.2	50.8	20.0	26.4	*−1.8
$LG(\frac{\sqrt{2}}{2})$	3×3	4205.9	466.0	107.4	172.9	41.6	22.6
$LG(1)$	5×5	**4492.3**	—	—	—	—	*−1.4
$LG(2)$	9×9	20.2	465.7	159.3	24.5	6.3	18.5
$G\,cos(2, 0°)$	5×5	3140.8	188.3	140.4	137.0	**135.4**	*−3.2
$G\,cos(2, 30°)$	5×5	4240.3	668.0	307.6	**317.8**	—	*−1.9
$G\,cos(2, 60°)$	5×5	3548.8	124.6	25.1	21.9	14.2	7.5
$G\,cos(2, 90°)$	5×5	1063.3	62.1	38.1	90.3	40.7	1.1
$G\,cos(2, 120°)$	5×5	1910.7	26.2	2.0	2.5	47.6	16.4
$G\,cos(2, 150°)$	5×5	3717.2	220.7	189.2	161.7	9.3	−0.8
$G\,cos(4, 0°)$	7×7	958.2	25.7	17.9	5.3	8.2	6.4
$G\,cos(4, 30°)$	7×7	2205.8	125.5	61.0	75.2	35.0	0.9
$G\,cos(4, 60°)$	7×7	1199.5	32.7	35.4	12.2	10.9	6.9
$G\,cos(4, 90°)$	7×7	108.8	229.6	130.6	20.2	31.9	30.2
$G\,cos(4, 120°)$	7×7	19.2	**1146.4**	—	—	—	*−2.7
$G\,cos(4, 150°)$	7×7	157.5	247.1	10.4	101.9	56.0	3.9
$G\,cos(6, 0°)$	11×11	102.1	12.8	4.3	−1.2	19.0	1.8
$G\,cos(6, 30°)$	11×11	217.3	54.8	8.4	32.9	11.5	−1.7
$G\,cos(6, 60°)$	11×11	85.6	4.7	0.1	4.5	3.8	6.0
$G\,cos(6, 90°)$	11×11	13.6	134.8	192.4	−0.4	7.9	1.6
$G\,cos(6, 120°)$	11×11	321.7	706.8	**640.3**	—	—	*−2.8
$G\,cos(6, 150°)$	11×11	3.8	100.1	12.7	98.6	75.1	*−1.4
$G\,cos(8, 0°)$	15×15	−1.6	11.0	−0.2	4.6	9.7	14.3
$G\,cos(8, 30°)$	15×15	2.4	33.0	2.1	13.8	12.7	−0.1
$G\,cos(8, 60°)$	15×15	10.7	5.5	−1.2	4.1	6.8	1.3
$G\,cos(8, 90°)$	15×15	203.0	51.9	71.7	3.9	12.3	6.8
$G\,cos(8, 120°)$	15×15	586.8	276.6	361.8	58.2	58.2	3.7
$G\,cos(8, 150°)$	15×15	140.1	44.6	1.3	45.5	42.5	38.0

Notes: *This filter has been chosen. Value computed using feature $\phi^{(\beta)}$, not $\phi_{\perp}^{(\beta)}$. The boldface numbers are the largest in each column and are thus chosen in the algorithm.

filter, and $d'(\phi^{(1)})$ for some filters are listed in Table 1. Filter $LG(1)$ has the largest entropy decrease and thus is chosen as the first filter, $S = \{LG(1)\}$. Then a model $p(\mathbf{I}; \Lambda, S)$ is computed, and a synthesized image is shown in Figure 4c.

Comparing Figure 4c with Figure 4b, it is evident that this filter captures local smoothness features of the observed texture image. Continuing the algorithm, six more filters are sequentially added: (2) $G\,cos(4, 120°)$; (3) $G\,cos(6, 120°)$; (4) $G\,cos(2, 30°)$; (5) $G\,cos(2, 0°)$; (6) $G\,cos(6, 150°)$; and (7) intensity $\delta(\ )$. The texture images synthesized using 3, 4, and 7 filters are displayed in Figures 4d–f. Obviously, with more filters added, the synthesized texture image gets closer to the observed one. After choosing seven filters, the entropy decrease for all filters becomes very small; some are negative. Similar results are observed for those filters not listed in Table 1. This confirms our early assumption that the marginal distributions of a small

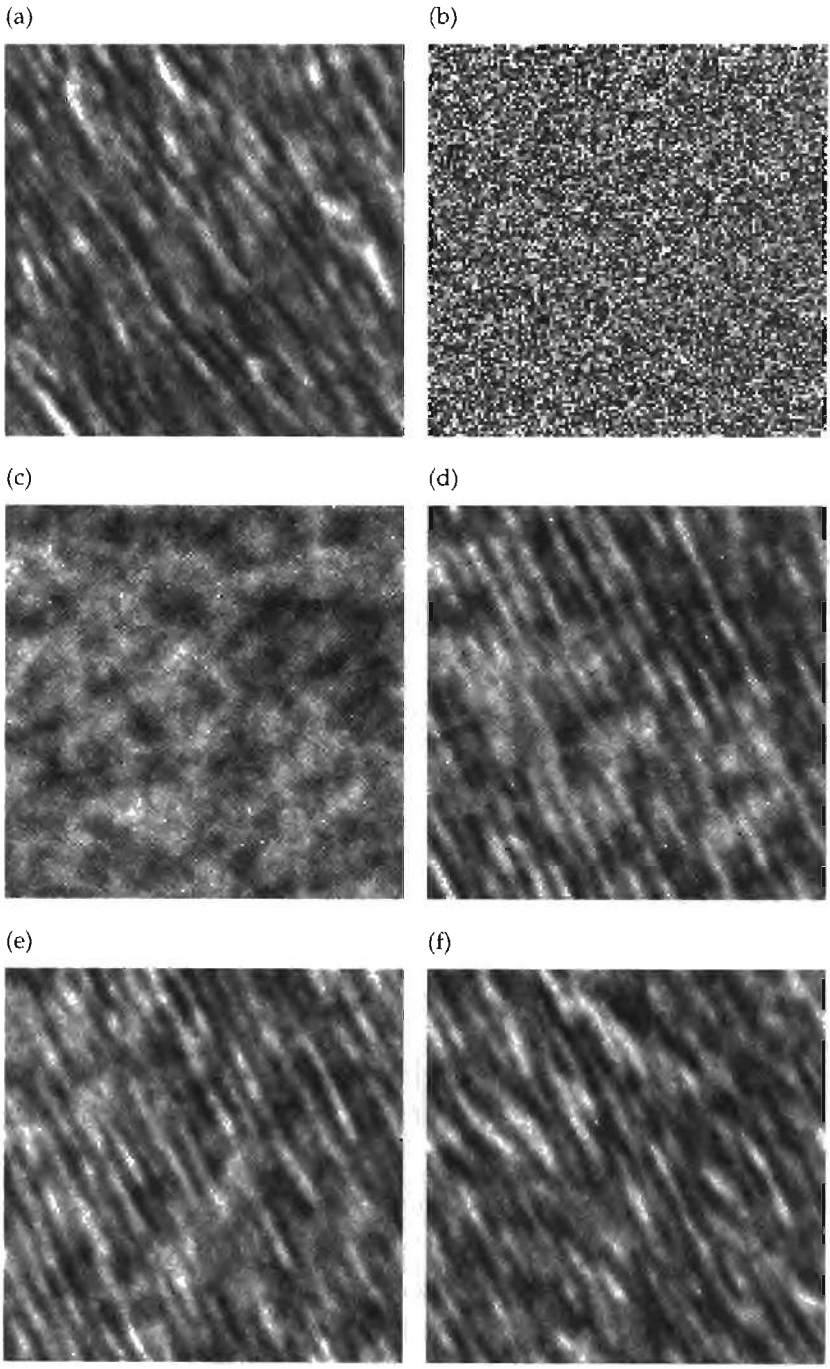

Figure 4: Synthesis of the fur texture. (a) The observed image. (b–f) The synthesized images using 0, 1, 3, 4, 7 filters, respectively.

(a) (b)

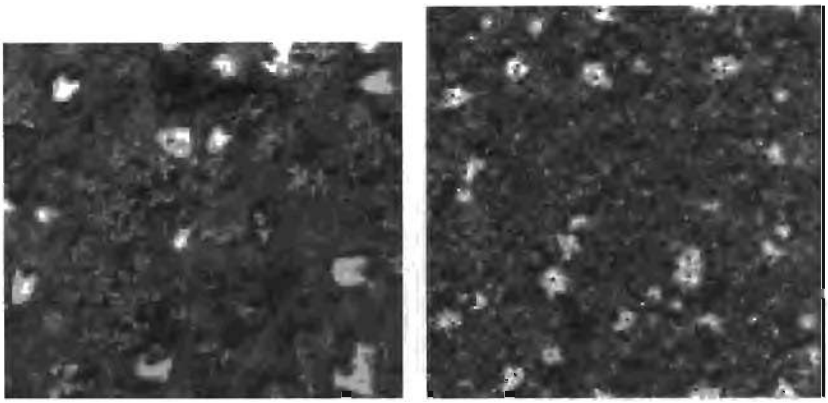

Figure 5: (a) The observed texture: mud. (b) The synthesized one using five filters.

(a) (b)

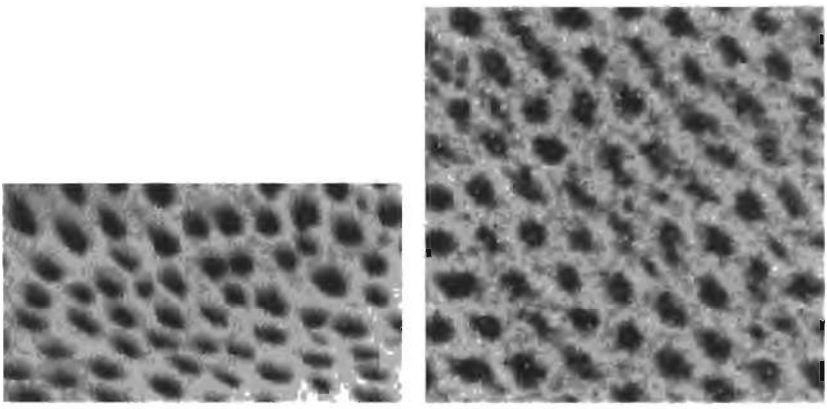

Figure 6: (a) The observed texture image: cheetah blob. (b) The synthesized one using six filters.

number of filtered images should be adequate for capturing the essential features of the underlying probability distribution $f(\mathbf{I})$.[3]

Figure 5a is the scene of mud ground with scattered animal footprints, which are filled with water and thus get brighter. This texture image shows sparse features. Figure 5b is the synthesized texture image using five filters.

Figure 6a is an image taken from the skin of a cheetah, and Figure 6b displays the synthesized texture using six filters. Notice that the original

[3] The synthetic fur texture in these figures is better than that in Zhu et al. (1996) since the L^1 criterion used here for filter pursuit has been replaced by the criterion of equation 3.7.

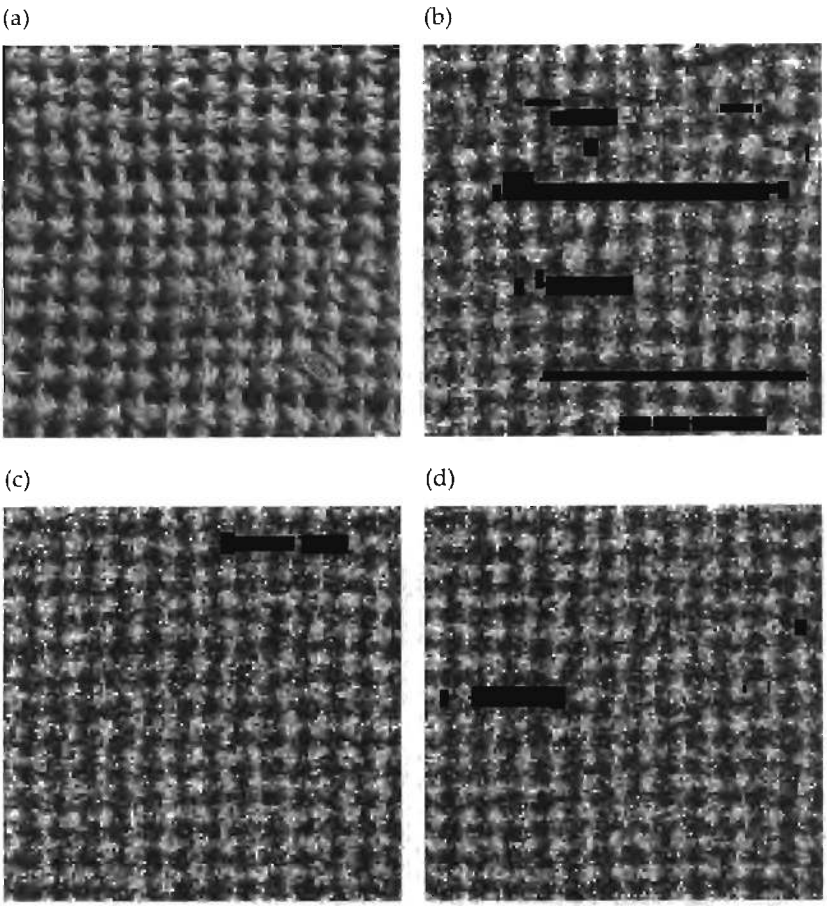

Figure 7: (a) The input image of fabric. (b) The synthesized image with two pairs of Gabor filters plus the Laplacian of gaussian filter. (c, d) Two more images sampled at different steps of the Gibbs sampler.

observed texture image is not homogeneous, since the shapes of the blobs vary systematically with spatial locations, and the left upper corner is darker than the right lower one. The synthesized texture, shown in Figure 6b, also has elongated blobs introduced by different filters, but the bright pixels seem to spread uniformly across the image due to the effect of entropy maximization.

Figure 7a shows a texture of fabric that has clear periods along both horizontal and vertical directions. We choose two nonlinear filters: spectrum analyzers $SP(17, 0°)$ and $SP(17, 90°)$, with their periods T tuned to the periods of the texture, and the window sizes of the filters are 33×33 pixels. We also use the intensity filter $\delta(\ )$ and filter $LG(\sqrt{2}/2)$ to take care of the

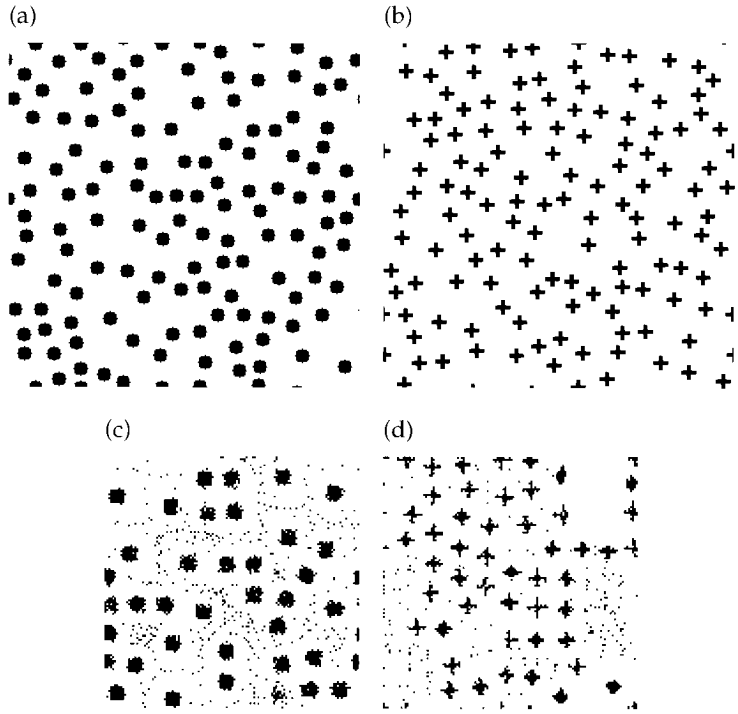

Figure 8: Two typical texton images of 256 × 256 pixels: (a) circle and (b) cross. (c, d) The two synthesized images of 128 × 128 pixels.

intensity histogram and the smoothness features. Three synthesized texture images are displayed in Figures 7b–d at different sampling steps. This experiment shows that once the Markov chain becomes stationary or gets close to stationary, the sampled images from $p(\mathbf{I})$ will always have perceptually similar appearances but with different details.

Figures 8a and 8b show two special binary texture images formed from identical textons (circles and crosses), which are studied extensively by psychologists for the purpose of understanding human texture perception. Our interest here is to see whether this class of textures can still be modeled by FRAME. We use the linear filter whose impulse response function is a mask with the corresponding texton at the center. With this filter selected, Figure 1b plots the histograms of the filtered image $F * \mathbf{I}$, with $\mathbf{I}$ being the texton image observed in Figure 8a (solid curve) and a uniform noise image (dotted curve). Observe that there are many isolated peaks in the observed histogram, which stand for important image features. The computation of the model is complicated by the nature of such isolated peaks, and we proposed an annealing approach for computing Λ (for details see Zhu et al.,

1996). Figures 8c and 8d show two synthesized images.

5 Discussion

This article proposes a minimax entropy principle for building probability models in a variety of applications. Our theory answers two major questions. The first is feature binding or feature fusion: how to integrate image features and their statistics into a single joint probability distribution without limiting the forms of the features. The second is feature selection: how to choose a set of features to characterize best the observed images. Algorithms are proposed for parameter estimation and stochastic simulation. A greedy algorithm is developed for feature pursuit, and the minimax entropy principle is corrected for the presence of sample variations.

As an example of applications, we apply the minimax entropy principle to modeling textures. There are various artificial categories for textures with respect to various attributes, such as Fourier and non-Fourier, deterministic and stochastic, and macro- and microtextures. FRAME erases these artificial boundaries and characterizes them in a unified model with different filters and parameter values. It has been well recognized that the traditional MRF models, as special cases of FRAME, can be used to model stochastic, non-Fourier microtextures. From the textures we synthesized, it is evident that FRAME is also capable of modeling periodic and deterministic textures (fabric), textures with large-scale elements (fur and cheetah blob), and textures with distinguishable textons (circles and cross bars).

Our method for texture modeling was inspired by and bears some similarities to the recent work by Heeger and Bergen (1995) on texture synthesis, where many natural-looking texture images are successfully synthesized by matching the histograms of filter responses organized in the form of a pyramid. Compared with Heeger and Bergen's algorithm, the FRAME model is distinctive in the following aspects. First, we obtain a probability model $p(\mathbf{I}; \Lambda, S)$ instead of merely synthesizing texture images. Second, the Monte Carlo Markov chain for model estimation and texture sampling is guaranteed to converge to a stationary process that follows the estimated distribution $p(\mathbf{I}; \Lambda, S)$ (Geman & Geman, 1984), and the observed histograms can be matched closely. However, the FRAME model is computationally expensive, and approaches for further facilitating the computation are yet to be developed. For more discussion on this aspect, see Zhu et al. (1996).

Many textures seem still difficult to model, such as the two human synthesized cloth textures shown in Figure 9. It appears that synthesizing such textures requires far more sophisticated features than those we have used in the texture modeling experiments, and these features may correspond to a high-level visual process, such as the geometrical properties of object shape. In this article, we choose filters from a fixed set of filters, but in general it is not understood how to design such set of features or structures for an arbitrary applications.

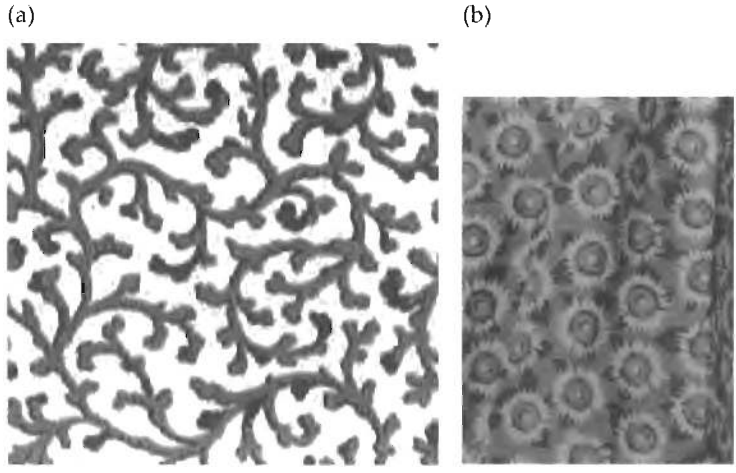

Figure 9: Two challenging texture images.

An important issue is whether the minimax entropy principle for model inference is "biologically plausible" and might be considered a model for the method used by natural intelligences in constructing models of classes of images. From a computational standpoint, the maximum entropy phase of the algorithm consists mainly of approximating the values of the Lagrange multipliers, which we have done by hill climbing with respect to log likelihood. Specifically, we have used Monte Carlo methods to sample our distributions and plugged the sampled statistics into the gradient of log likelihood. One of the authors has conjectured that feedback pathways in the cortex may serve the function of forming mental images on the basis of learned models of the distribution on images (Mumford, 1992). Such a mechanism might well sample by Monte Carlo as in the algorithm in this article. That theory further postulated that the cortex seeks out the "residuals," the features of the observed image different from those of the mental image. The algorithm shows how such residuals can be used to drive a learning process in which the Lagrange multipliers are gradually improved to increase the log likelihood. We would conjecture that these Lagrange multipliers are stored as suitable synaptic weights in the higher visual area or in the top-down pathway. Given the massively parallel architecture, the apparent stochastic component in neural firing, and the huge amount of observed images processed every day, the computational load of our algorithm may not be excessive for cortical implementation.

The minimum entropy phase of our algorithm has some direct experimental evidence in its favor. There has been extensive psychophysical experimentation on the phenomenon of preattentive texture discrimination.

We propose that textures that can be preattentively discriminated are exactly those for which suitable filters have been incorporated into a minimum entropy cortical model and that the process by which subjects can train themselves to discriminate new sets of textures preattentively is exactly that of incorporating a new filter feature into the model. Evidence that texture pairs that are not preattentively segmentable by naive subjects become segmentable after practice has been reported by many groups, most notably by Karni and Sagi (1991). The remarkable specificity of the reported texture discrimination learning suggests that very specific new filters are incorporated into the cortical texture model, as in our theory.

Appendix: Mathematical Details

Proof of Theorem 1. Let $\Lambda^* = (\lambda^{*(1)}, \lambda^{*(2)}, \ldots, \lambda^{*(K)})$ be the parameter. By definition we have $E_{p(\mathbf{I};\Lambda^*,S)}[\phi^{(\alpha)}(\mathbf{I})] = E_f[\phi^{(\alpha)}(\mathbf{I})], \quad \alpha = 1, \ldots, K.$

$$
\begin{aligned}
E_f[\log p(\mathbf{I};\Lambda^*,S)] &= -E_f[\log Z(\Lambda^*)] - \sum_{\alpha=1}^{K} E_f[\langle \lambda^{*(\alpha)}, \phi^{(\alpha)}(\mathbf{I})\rangle], \\
&= -\log Z(\Lambda^*) - \sum_{\alpha=1}^{K} \langle \lambda^{*(\alpha)}, E_f[\phi^{(\alpha)}(\mathbf{I})]\rangle, \\
&= -\log Z(\Lambda^*) - \sum_{\alpha=1}^{K} \langle \lambda^{*(\alpha)}, E_{p(\mathbf{I};\Lambda^*,S)}[\phi^{(\alpha)}(\mathbf{I})]\rangle, \\
&= E_{p(\mathbf{I};\Lambda^*,S)}[\log p(\mathbf{I};\Lambda^*,S)] = -\mathit{entropy}(p(\mathbf{I};\Lambda^*,S)),
\end{aligned}
$$

and the result follows.

Proof of Proposition 3. Let $\Phi = (\phi^{(1)}(\mathbf{I}), \ldots, \phi^{(K)}(\mathbf{I})), \Phi_+ = (\Phi(\mathbf{I}), \phi^{(\beta)}(\mathbf{I}))$. We have the entropy decrease

$$
\begin{aligned}
d(\phi^{(\beta)}) &= KL(p_+; p) \\
&= \frac{1}{2}(E_p[\Phi_+(\mathbf{I})] - E_{p_+}[\Phi_+(\mathbf{I})])' Var_{p'}[\Phi_+(\mathbf{I})]^{-1} \\
&\quad \times (E_p[\Phi_+(\mathbf{I})] - E_{p_+}[\Phi_+(\mathbf{I})]) \qquad \text{(A.1)} \\
&= \frac{1}{2}(E_p[\phi^{(\beta)}(\mathbf{I})] - E_f[\phi^{(\beta)}(\mathbf{I})]) V_{p'}^{-1} \\
&\quad \times (E_p[\phi^{(\beta)}(\mathbf{I})] - E_f[\phi^{(\beta)}(\mathbf{I})]). \qquad \text{(A.2)}
\end{aligned}
$$

Equation A.1 follows a second-order Taylor expansion argument (corollary 4.4 of Kullback, 1959, p. 48), where p' is a distribution whose expected feature statistics are between those of p and p_+, and

$$
Var_{p'}[\Phi_+(\mathbf{I})] = \begin{pmatrix} Var_{p'}[\Phi(\mathbf{I})] & Cov_{p'}[\Phi(\mathbf{I}), \phi^{(\beta)}(\mathbf{I})] \\ Cov_{p'}[\phi^{(\beta)}(\mathbf{I}), \Phi(\mathbf{I})] & Var_{p'}[\phi^{(\beta)}(\mathbf{I})] \end{pmatrix}
$$

$$= \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}.$$

Equation A.2 results from the fact that $E_{p_1}[\Phi(\mathbf{I})] = E_p[\Phi(\mathbf{I})]$, and by the Schur formula, it is well known that $V_{p'} = V_{22} - V_{21}V_{11}^{-1}V_{12}$.

Proof of Proposition 4. From the proof of Theorem 1, we know $E_f[\log p(\mathbf{I}; \Lambda^\star, S)] = E_{p(\mathbf{I};\Lambda^\star,S)}[\log p(\mathbf{I}; \Lambda^\star, S)]$, and by similar derivation we have $E_f[\log p(\mathbf{I}; \Lambda, S)] = E_{p(\mathbf{I};\Lambda^\star,S)}[\log p(\mathbf{I}; \Lambda, S)]$ for any Λ.

$$\begin{aligned}
&KL(f, p(\mathbf{I}; \Lambda, S)) \\
&\quad = E_f[\log f(\mathbf{I})] - E_f[\log p(\mathbf{I}; \Lambda, S)] \\
&\quad = E_f[\log f(\mathbf{I})] - E_{p(\mathbf{I};\Lambda^\star,S)}[\log p(\mathbf{I}; \Lambda, S)] \\
&\quad = E_f[\log f(\mathbf{I})] - E_f[\log p(\mathbf{I}; \Lambda^\star, S)] + E_{p(\mathbf{I};\Lambda^\star,S)}[\log p(\mathbf{I}; \Lambda^\star, S)] \\
&\qquad - E_{p(\mathbf{I};\Lambda^\star,S)}[\log p(\mathbf{I}; \Lambda, S)] \\
&\quad = KL(f, p(\mathbf{I}; \Lambda^\star, S)) + KL(p(\mathbf{I}; \Lambda^\star, S), p(\mathbf{I}; \Lambda, S)).
\end{aligned}$$

The result follows by setting $\Lambda = \hat{\Lambda}$.

Proof of Proposition 5. By Proposition 2 we have $L(\hat{\Lambda}, S) = -entropy(p(\mathbf{I}; \hat{\Lambda}, S))$. By similar derivation, we can prove that $E_{\text{obs}}[L(\Lambda^\star, S)] = -entropy(p(\mathbf{I}; \Lambda^\star, S))$ and

$$L(\hat{\Lambda}, S) - L(\Lambda^\star, S) = KL(p(\mathbf{I}; \hat{\Lambda}, S), p(\mathbf{I}; \Lambda^\star, S)). \tag{A.3}$$

Applying E_{obs} to both sides of equation A.3, we have

$$\begin{aligned}
&-E_{\text{obs}}[entropy(p(\mathbf{I}; \hat{\Lambda}, S))] + entropy(p(\mathbf{I}; \Lambda^\star, S)) \\
&\quad = E_{\text{obs}}[KL(p(\mathbf{I}; \hat{\Lambda}, S), p(\mathbf{I}; \Lambda^\star, S))],
\end{aligned}$$

and the result follows.

Proof of Equation 3.6. To simplify the notation, we denote

$$\begin{aligned}
Var_{p_+^*}[\Phi_+(\mathbf{I})] &= \begin{pmatrix} Var_{p_+^*}[\Phi(\mathbf{I})] & Cov_{p_+^*}[\Phi(\mathbf{I}), \phi^{(\beta)}(\mathbf{I})] \\ Cov_{p_+^*}[\phi^{(\beta)}(\mathbf{I}), \Phi(\mathbf{I})] & Var_{p_+^*}[\phi^{(\beta)}(\mathbf{I})] \end{pmatrix} \\
&= \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix},
\end{aligned}$$

$$A = \begin{pmatrix} I_1 & 0 \\ -X_{21}X_{11}^{-1} & I_2 \end{pmatrix},$$

and

$$B = \begin{pmatrix} Var_{p_+^*}[\Phi(\mathbf{I})] & 0 \\ 0 & Var_{p_+^*}[\phi_\perp^{(\beta)}(\mathbf{I})] \end{pmatrix},$$

where I_1, I_2 are identity matrices, and $\phi_\perp^{(\beta)}(\mathbf{I})$ is uncorrelated with $\Phi(\mathbf{I})$ under p_+^*. So we have $A\Phi_+(\mathbf{I}) = (\Phi(\mathbf{I}), \phi_\perp^{(\beta)}(\mathbf{I}))'$ and $AVar_{p_+^*}[\Phi_+(\mathbf{I})]A' = B$.

Thus $Var_{p_+^*}^{-1}[\Phi_+(\mathbf{I})] = A'B^{-1}A$, and since $Var_{p_+^*}[\Phi(\mathbf{I})] = Var_{p^*}[\Phi(\mathbf{I})]$, therefore

$$\begin{aligned} tr(Var_f[\Phi_+(\mathbf{I})]Var_{p_+^*}^{-1}[\Phi_+(\mathbf{I})]) &= tr(Var_f[\Phi_+(\mathbf{I})](A')B^{-1}A) \\ &= tr((AVar_f[\Phi_+(\mathbf{I})]A')B^{-1}) \\ &= tr(Var_f[\Phi(\mathbf{I})]Var_{p^*}^{-1}[\Phi(\mathbf{I})]) \\ &\quad + tr(Var_f[\phi_\perp^{(\beta)}(\mathbf{I})]Var_{p_+^*}^{-1}[\phi_\perp^{(\beta)}(\mathbf{I})]) \end{aligned}$$

and equation 3.6 follows.

Acknowledgments

We are very grateful to two anonymous referees, whose insightful comments greatly improved the presentation of this article. This work is supported by the NSF grant DMS-91-21266 to David Mumford. The second author is supported by a grant to Donald B. Rubin.

References

Akaike, H. (1977). On entropy maximization principle. In P. R. Krishnaiah (Ed.), *Applications of Statistics* (pp. 27–42). Amsterdam: North-Holland.

Barlow, H. B., Kaushal, T. P., & Mitchison, G. J. (1989). Finding minimum entropy codes. *Neural Computation, 1*, 412–423.

Bergen, J. R., & Adelson, E. H. (1991). Theories of visual texture perception. In D. Regan (Ed.), *Spatial Vision*, Boca Raton, FL: CRC Press, 1991.

Besag, J. (1973). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Royal Stat. Soc., B, 36*, 192–236.

Brown, L. D. (1986). *Fundamentals of statistical exponential families: With applications in statistical decision theory*. Hayward, CA: Institute of Mathematical Statistics.

Chubb, C., & Landy, M. S. (1991). Orthogonal distribution analysis: A new approach to the study of texture perception. In M. S. Landy & J. A. Movshon (Eds.), *Computational models of visual processing*. Cambridge, MA: MIT Press.

Coifman, R. R., & Wickerhauser, M. V. (1992). Entropy based algorithms for best basis selection. *IEEE Trans. on Information Theory, 38*, 713–718.

Cross, G. R., & Jain, A. K. (1983). Markov random field texture models. *IEEE, PAMI, 5*, 25–39.

Daugman, J. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of Optical Soc. Amer.*, 2(7).

Dayan, P., Hinton, G. E., Neal, R. N., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation*, 7, 889–905.

Donoho, D. L., & Johnstone, I. M. (1994). Ideal de-noising in an orthonormal basis chosen from a library of bases. *Acad. Sci. Paris, Ser. I*, 319, 1317–1322.

Field, D. (1994). What is the goal of sensory coding? *Neural Computation*, 6, 559–601.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Trans. PAMI*, 6, 721–741.

Heeger, D. J., & Bergen, J. R. (1995). Pyramid-based texture analysis/synthesis. In *Computer Graphics Proceedings* (pp. 229–238).

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review, 106*, 620–630.

Jolliffe, I. T. (1986). *Principle components analysis.* New York: Springer-Verlag.

Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation, 6*, 181–214.

Julesz, B. (1962). Visual pattern discrimination. *IRE Transactions of Information Theory, IT-8*, 84–92.

Julesz, B. (1995). *Dialogues on perception.* Cambridge, MA: MIT press.

Karni, A., & Sagi, D. (1991). Where practice makes perfect in texture discrimination—evidence for primary visual cortex plasticity. *Proc. Nat. Acad. Sci. U.S.A., 88*, 4966–4970.

Kullback, S. (1959). *Information theory and statistics.* New York: Wiley.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annual Math. Stat., 22*, 79–86.

Mallat, S. (1989). Multi-resolution approximations and wavelet orthonormal bases of $L^2(R)$. *Trans. Amer. Math. Soc., 315*, 69–87.

Mumford, D. B. (1992). On the computational architecture of the neocortex II: The role of cortico-cortical loops. *Biological Cybernetics, 66*, 241–251.

Mumford, D. B., & Shah, J. (1989). Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. Pure Appl. Math., 42*, 577–684.

Picard, R. W. (1996). *A society of models for video and image libraries* (Technical Rep. No. 360). Cambridge, MA: MIT Media Lab.

Priestley, M. B. (1981). *Spectral analysis and time series.* San Diego: Academic Press.

Ripley, B. (1996). *Pattern recognition and neural networks.* Cambridge: Cambridge University Press.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry.* Singapore: World Scientific.

Sherman, M. (1996). Variance estimation for statistics computed from spatial lattice data. *J. R. Statistics Soc., B, 58*, 509–523.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27.

Silverman, M. S., Grosof, D. H., De Valois, R. L., & Elfar, S. D. (1989). Spatial-frequency organization in primate striate cortex. *Proc. Natl. Acad. Sci. U.S.A., 86*.

Simoncelli, E. P., Freeman, W. T., Adelson, E. H., & Heeger, D. J. (1992). Shiftable multi-scale transforms. *IEEE Trans. on Information Theory, 38*, 587–607.

Watson, A. (1987). Efficiency of a model human image code. *J. Opt. Soc. Am. A, 4(12)*, 2401–2417.

Winkler, G. (1995). *Image analysis, random fields and dynamic Monte Carlo methods.* Berlin: Springer-Verlag.

Witkin, A., & Kass, M. (1991). Reaction-diffusion textures. *Computer Graphics, 25*, 299–308.

Xu, L. (1995). Ying-Yang machine: A Bayesian-Kullback scheme for unified learnings and new results on vector quantization. *Proc. Int'l Conf. on Neural Info. Proc.* Hong Kong.

Younes, L. (1988). Estimation and annealing for Gibbsian fields (STMA V30 1845). *Annales de l'Institut Henri Poincaré, Section B, Calcul des Probabilities et Statistique, 24*, 269–294.

Zhu, S. C. (1996). *Statistical and computatinal theories for image segmentation, texture modeling and object recognition.* Unpublished Ph.D. dissertation, Harvard University.

Zhu, S. C., & Mumford, D. B. (1997). Learning generic prior models for visual computation. *Proc. of Int'l Conf. on Computer Vision and Pattern Recognition.* Puerto Rico.

Zhu, S. C., Wu, Y. N., & Mumford, D. B. (1996). FRAME: Filters, random fields and maximum entropy—to a unified theory for texture modeling. In *Proc. of Int'l Conf. on Computer Vision and Pattern Recognition.* San Francisco.

Received April 10, 1996; accepted March 27, 1997.

Address correspondence to zhu@dam.brown.edu.