Neural Computation Vol.9, No.8, Nov. 1997.

Minimax Entropy Principle and Its Application to Texture Modeling

Song Chun Zhu [1] Ying Nian Wu [2] David Mumford[1]

Abstract

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 certain set of feature statistics, a distribution can be built to bind these feature statistics together by maximizing the entropy over all distributions that reproduce these feature statistics. 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 model complexity is restricted because of the sample variation in the observed feature statistics. 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. Relationship between our theory and the mechanisms of neural computation is also discussed.

[1] Division of Applied Mathematics, Brown University, Providence, RI 02912
[2] Department of Statistics, Harvard University, Cambridge, MA 02138.

1 Introduction

This article proposes a general theory and methodology, called 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, 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 assumed that there exists a joint probability distribution (or density) $f(\mathbf{I})$ over the image space, $f(\mathbf{I})$ should concentrate on a subspace which 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 a 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 et al 1989), where the computation of the entropy and thus the choice of the optimal coding schemes depend on the estimation of the underlying probability distribution $f(\mathbf{I})$. 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 non-Gaussian 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). For example, in image restoration and surface reconstruction (Geman and Geman 1984, Blake and Zisserman 1987), a simple model of $f(\mathbf{I})$ should embodies the common features and statistics of natural looking images, for instance, in natural images adjacent pixels have similar intensity values, so that it will bias a vision algorithm against undesirable features such as noises and blurrings.

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 which have perceptually similar texture appearances. $p(\mathbf{I})$ is not only important for texture analysis such as texture segmentation and texture classification, but also plays a role in texture synthesis since texture images can be synthesized by drawing samples from $p(\mathbf{I})$. Furthermore, finding simple distributions to characterize textures helps us understand the mechanisms of human texture perception (Julesz 1995).

However, making inference about $f(\mathbf{I})$ is much more challenging than many of the learning problems in neural networks (Dayan et al., 1995, Xu 1995) for the following reasons.

Firstly, 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 the probability distribution 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 non-parametric inference methods, such as kernel methods, radial basis functions (see Ripley 1996) and mixture of Gaussian models (Jordan and Jacobs 1994).

Secondly, $f(\mathbf{I})$ is often far from being Gaussian, therefore some popular dimension reduction techniques, such as the principal component analysis (Jolliffe 1986),

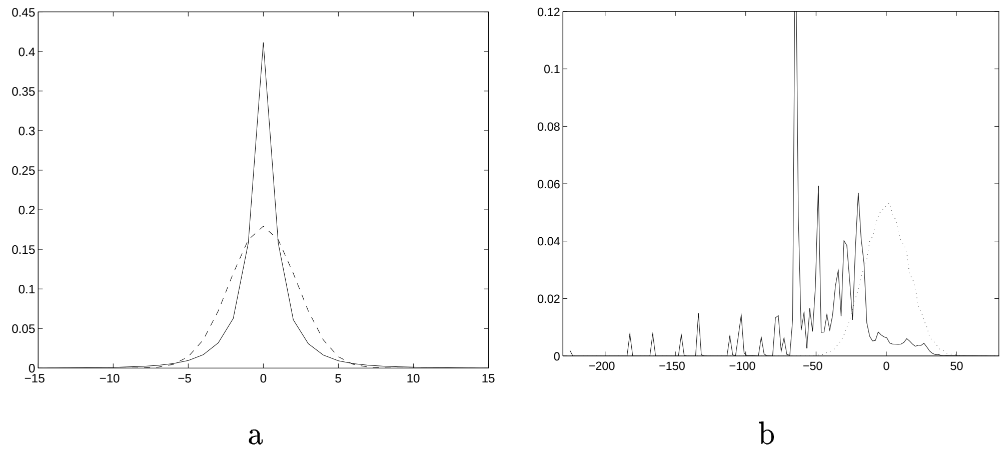

Figure 1: a. The histogram of intensity difference at adjacent pixels and Gaussian curve (dashed) of same mean and variance in domain $[-15, 15]$. b. Histogram of the filtered texton image (solid curve) and histogram of a filtered noise image (dotted curve).

and spectral analysis (Priestley 1981), do not appear to be directly applicable. As an illustration of the non-Gaussian property, figure (1.a) shows the empirical marginal distribution (or histogram) of the intensity differences of horizontally adjacent pixels of some natural images (Zhu and Mumford 1995). As a comparison, the Gaussian distribution with the same mean and variance is plotted as dashed curve in figure (1.a). Similar non-Gaussian properties are also observed in Field (1994). Another example is shown in figure (1.b), where the solid curve is the histogram of $F * \mathbf{I}$ with $\mathbf{I}$ being a texton image shown in figure (8.a), and F is a filter with the same texton (see section (3.5) for details). It is clear that the solid curve is far from being Gaussian, and as a comparison, the dotted curve in figure (1.b) is the histogram of $F * \mathbf{I}$ with $\mathbf{I}$ being a white noise image.

A key issue in building a statistical model is the balance between generality and simplicity — the model should include rich structures to adequately describe real world images and should be capable of modeling complexities due to high dimensionality and non-Gaussian property, and at the same time, it should be simple enough to be computationally feasible, and to give simple explanation to what we observe. To reduce complexity, it is often necessary to impose structures on the distribution.

1.2 Previous methods

In the past there have been mainly two methods 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 and Geman 1984, Mumford and Shah 1989), and the conditional auto-regression models in texture modeling (Besag 1973, Cross and 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 only afford 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 that the forms of the distributions should 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) designed as a model for cells in mammalian visual cortex, and some wavelet pyramids based on various coding criteria (Mallat 1989, Simoncelli and Adelson 1990, Coifman and Wickerhauser 1992, Donoho and 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 and 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 best characterize the images being modeled.

1.3 Our theory and methodology

In this paper, 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 features of an image can be expressed as $\phi^{(\alpha)}(\mathbf{I})$, where $\alpha = 1, 2, ..., K$ is the index of the features and $\phi^{(\alpha)}()$ can be vector valued functions of the image intensities. 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 can estimated by the sample mean of the feature computed from the training images. Then a model $p(\mathbf{I})$ is constructed such that it can reproduce the feature statistics as observed, i.e., $E_p[\phi^{(\alpha)}(\mathbf{I})] = E_f[\phi^{(\alpha)}(\mathbf{I})], \quad \text{for } \alpha = 1, 2, ..., K$. 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 ME distribution $p(\mathbf{I})$ constructed in (I) depends on the features that we selected, and the goodness of $p(\mathbf{I})$ is measured by the Kullback-Leibler divergence from $f(\mathbf{I})$ to $p(\mathbf{I})$ (Kullback and Leibler 1951). As we will show in the next section, this divergence 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 $p(\mathbf{I})$, which 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. In cases when the model complexity or the number of features K is fixed for computational reasons, the minimum entropy principle also provides a criterion for selecting the features which best characterize $f(\mathbf{I})$.

Computational procedures are proposed for parameter estimation and feature selection, and model complexity is studied in the presence of sample variations 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 because only a small set of filters is used when modeling a certain texture. Texture images are synthesized by drawing samples from the estimated models, and the correctness of estimated models are thus verified by checking whether the synthesized texture images have similar visual appearances to the observed images.

The rest of the paper is arranged as follows. Section (2) is devoted to a formal study of the minimax entropy principle, where a greedy algorithm for feature selection is proposed. Section (3) applies the minimax entropy principle to texture modeling. Section (3.5) consists of experiments of modeling a variety of textures. Finally section (4) concludes with a brief discussion.

2 The minimax entropy principle

To fix notation, let $\mathbf{I}$ be an image defined on a domain $\mathcal{D}$ (e.g., $\mathcal{D}$ can be a $N \times N$ lattice), where for each point $\vec{v} \in \mathcal{D}$, $\mathbf{I}(\vec{v}) \in \mathcal{L}$, which is an interval on the real line or a set of integers. It is assumed that the observed images $\{\mathbf{I}_i^{obs}, i = 1, ..., M\}$ are a random sample from a 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. The objective is to estimate $f(\mathbf{I})$ based on the observed images.

2.1 The maximum entropy principle

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

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

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

$$\Omega = \{p(\mathbf{I}) \ : \ E_p[\phi^{(\alpha)}(\mathbf{I})] = \mu_{obs}^{(\alpha)}, \ \ \alpha = 1, ..., K\}$$

be the set of distributions that reproduce the observed features, then we need to select a $p(\mathbf{I}) \in \Omega$ provided that $\Omega \neq \emptyset$.

As far as the observed feature statistics $\{\mu_{obs}^{(\alpha)}, \alpha = 1, ..., K\}$ are concerned, all the distributions in Ω explain them equally well, and they are not distinguishable from $f(\mathbf{I})$. The maximum entropy (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, i.e., $p(\mathbf{I})$ *should represent information no more than that is available* and in this sense, the ME principle is often called the minimum prejudice principle.

Thus the problem becomes the following constrained optimization problem,

$$\begin{aligned} & p(\mathbf{I}) = \arg\max\{-\int p(\mathbf{I}) \log p(\mathbf{I}) d\mathbf{I}\}, && (1) \\ \text{subject to} \quad & E_p[\phi^{(\alpha)}(\mathbf{I})] = \int \phi^{(\alpha)}(\mathbf{I}) p(\mathbf{I}) d\mathbf{I} = \mu_{obs}^{(\alpha)}, \quad \alpha = 1, ..., K, \\ \text{and} \quad & \int p(\mathbf{I}) d\mathbf{I} = 1. \end{aligned}$$

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

$$p(\mathbf{I}; \Lambda) = \frac{1}{Z(\Lambda)} \exp\{-\sum_{\alpha=1}^{K} < \lambda^{(\alpha)}, \phi^{(\alpha)}(\mathbf{I}) >\}, \tag{2}$$

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

$$Z(\Lambda) = \int \exp\{-\sum_{\alpha=1}^{K} < \lambda^{(\alpha)}, \phi^{(\alpha)}(\mathbf{I}) >\} d\mathbf{I}$$

is the partition function which normalizes $p(\mathbf{I}; \Lambda)$ into a probability distribution. Equation (2) specifies a simple parametric model, and the parameter Λ is solved at $\hat{\Lambda}$ which satisfies the constraints $p(\mathbf{I}; \hat{\Lambda}) \in \Omega$, i.e.,

$$E_{p(\mathbf{I};\hat{\Lambda})}[\phi^{(\alpha)}(\mathbf{I})] = \mu_{obs}^{(\alpha)}, \quad \alpha = 1, ..., K. \tag{3}$$

2.2 Estimation and computation

The $\hat{\Lambda}$ computed from equation (3) is actually the maximum likelihood estimate (MLE) of Λ. Let $L(\Lambda) = \frac{1}{M}\sum_{i=1}^{M} \log p(\mathbf{I}_i^{obs}; \Lambda)$ be the log-likelihood function, then it has the following properties.

- Property 1). $\frac{\partial L(\Lambda)}{\partial \lambda^{(\alpha)}} = -\frac{1}{Z}\frac{\partial Z}{\partial \lambda^{(\alpha)}} - \mu_{obs}^{(\alpha)} = E_{p(\mathbf{I};\Lambda)}[\phi^{(\alpha)}] - \mu_{obs}^{(\alpha)}, \quad \forall \alpha.$
- Property 2). $\frac{\partial^2 L(\Lambda)}{\partial \lambda^{(\alpha)} \lambda^{(\beta)'}} = E_{p(\mathbf{I};\Lambda)}[(\phi^{(\alpha)}(\mathbf{I}) - \mu_{obs}^{(\alpha)})(\phi^{(\beta)}(\mathbf{I}) - \mu_{obs}^{(\beta)})'], \quad \forall \alpha, \beta.$

By gradient ascent, maximizing the log-likelihood gives the following equation for solving Λ iteratively,

$$\frac{d\lambda^{(\alpha)}}{dt} = E_{p(\mathbf{I};\Lambda)}[\phi^{(\alpha)}(\mathbf{I})] - \mu_{obs}^{(\alpha)}, \ \ \alpha = 1, ..., K, \tag{4}$$

which converges to $\hat{\Lambda}$. Equation (4) follows from property 1) of $L(\Lambda)$. Property 2) means that the Hessian matrix of $L(\Lambda)$ is the covariance matrix $(\phi^{(1)}(\mathbf{I}), ..., \phi^{(K)}(\mathbf{I}))$ 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 $\alpha = 0, ..., K$, which is usually satisfied. So $L(\Lambda)$ is strictly concave with respect to Λ, and the solution for Λ uniquely exists.

At each step t of equation (4), the computation of $E_{p(\mathbf{I};\Lambda)}[\phi^{(\alpha)}(\mathbf{I})]$ is in general difficult, and we adopt the stochastic gradient method (Younes 1988) for approximation. For a fixed Λ, we synthesize some typical images $\{\mathbf{I}_i^{syn}, \ i = 1, .., M'\}$ by drawing samples from $p(\mathbf{I}; \Lambda)$ using the Gibbs sampler (Geman and Geman 1984) or other Markov chain Monte Carlo (MCMC) methods (see Winkler 1995), and approximate $E_{p(\mathbf{I};\Lambda)}[\phi^{(\alpha)}(\mathbf{I})]$ by the sample means, i.e.,

$$E_{p(\mathbf{I};\Lambda)}[\phi^{(\alpha)}(\mathbf{I})] \approx \frac{1}{M'}\sum_{i=1}^{M'} \phi^{(\alpha)}(\mathbf{I}_i^{syn}) = \mu_{syn}^{(\alpha)}(\Lambda), \ \ \alpha = 1, ..., K. \tag{5}$$

Therefore the iterative equation for computing Λ becomes

$$\frac{d\lambda^{(\alpha)}}{dt} = \Delta^{(\alpha)}(\Lambda) = \mu_{syn}^{(\alpha)}(\Lambda) - \mu_{obs}^{(\alpha)}, \ \ \alpha = 1, ..., K. \tag{6}$$

For the accuracy of the approximation in equation (5), 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 and Mumford 1996).

2.3 The minimum entropy principle

For now, let's suppose that the sample size M is large enough so that the expected feature statistics $\{E_f[\phi^{(\alpha)}(\mathbf{I}), \alpha = 1, ..., K\}$ can be estimated exactly by neglecting the estimation errors in the observed statistics $\{\mu^{(\alpha)}_{obs}, \alpha = 1, ..., K\}$. Then an ME distribution $p(\mathbf{I}; \Lambda^\star)$ is computed so that it reproduces the expected feature statistics, i.e.,

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

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 and Leibler 1951) divergence from $f(\mathbf{I})$ to $p(\mathbf{I}; \Lambda^\star)$,

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

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

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

See appendix for a proof.

In the above result, $entropy(f(\mathbf{I}))$ is fixed, and the entropy of $p(\mathbf{I}; \Lambda^\star)$ depends on the set of features $\{\phi^{(\alpha)}(\mathbf{I}), \ \alpha = 1, 2, ...\}$ included in the distribution $p(\mathbf{I}; \Lambda^\star)$. Thus minimizing $KL(f, p(\mathbf{I}; \Lambda^\star))$ is equivalent to minimizing the entropy of $p(\mathbf{I}; \Lambda^\star)$. We call this the minimum entropy principle, and it has the following intuitive interpretations. First, in information theory, $p(\mathbf{I}; \Lambda^\star)$ defines an optimal coding scheme with each $\mathbf{I}$ assigned a coding length $-\log p(\mathbf{I}; \Lambda^\star)$ (Shannon 1948), and $entropy(p(\mathbf{I}; \Lambda^\star)) = E_p[-\log p(\mathbf{I}; \Lambda^\star)]$ stands for the expected coding length. Therefore, a minimum entropy principle chooses the coding system with the shortest average coding length. Second, in statistics, $entropy(p(\mathbf{I}; \Lambda^\star))$ is the negative Kullback-Leibler divergence, up to a constant, from $p(\mathbf{I}; \Lambda^\star)$ to a uniform distribution, with the latter being a model for random noise images. To minimizing the entropy, $p(\mathbf{I}; \Lambda^\star)$ should be made as "orderly" or "regular" as possible. The philosophy of entropy minimization is that *we should make use of all the information* *or statistics observable to specify* $p(\mathbf{I};\Lambda)$. Unlike the maximum entropy principle which favors simplicity, the minimum entropy principle emphasizes generality of the model.

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, i.e.,

$$S^* = \arg\min_{|S|=K} entropy(p_S(\mathbf{I};\Lambda^\star)), \tag{7}$$

where $p_S(\mathbf{I};\Lambda^\star)$ denotes the fitted model using features in S. Let

$$\Omega_S = \{p(\mathbf{I}) \ : \ E_p[\phi^{(\alpha)}(\mathbf{I})] = E_f[\phi^{(\alpha)}(\mathbf{I})], \ \forall \phi^{(\alpha)} \in S\}$$

be the set of probability distributions which can reproduce the expected features statistics in S, then according to the maximum entropy principle,

$$p_S(\mathbf{I};\Lambda^\star) = \arg\max_{p\in\Omega_S} entropy(p). \tag{8}$$

Combining (7) and (8), we have

$$S^* = \arg\min_{|S|=K} \{\max_{p\in\Omega_S} entropy(p)\}. \tag{9}$$

We call equation (9) the *minimax entropy principle*, and we have demonstrated that this principle is consistent with the goal of modeling, i.e., finding the best estimate for the underlying distribution $f(\mathbf{I})$.

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

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

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 $p(\mathbf{I}; \Lambda^\star)$, and keep doing this until the entropy decrease is smaller than a certain value. To be precise, let $S = \{\phi^{(\alpha)},\ \ \alpha = 1, ..., K\}$ be the currently selected set of features, and let

$$p = p(\mathbf{I}; \Lambda) = \frac{1}{Z(\Lambda)} \exp\{-\sum_{\alpha=1}^{K} < \lambda^{(\alpha)}, \phi^{(\alpha)}(\mathbf{I}) >\} \tag{10}$$

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 becomes

$$p_+ = p(\mathbf{I}; \Lambda_+) = \frac{1}{Z(\Lambda_+)} \exp\{-\sum_{\alpha=1}^{K} < \lambda_+^{(\alpha)}, \phi^{(\alpha)}(\mathbf{I}) > - < \lambda_+^{(\beta)}, \phi^{(\beta)}(\mathbf{I}) >\}. \tag{11}$$

In general, $\lambda_+^{(\alpha)} \neq \lambda^{(\alpha)}$ for $\alpha = 1, ..., K$.

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

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

where

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

is the entropy decrease, which can be expressed in a quadratic form by the second-order Taylor expansion.

Proposition 1 *In the above notation,*

$$d(\phi^{(\beta)}) = \frac{1}{2}(E_p[\phi^{(\beta)}(\mathbf{I})] - E_f[\phi^{(\beta)}(\mathbf{I})])' V_{p'}^{-1} (E_p[\phi^{(\beta)}(\mathbf{I})] - E_f[\phi^{(\beta)}(\mathbf{I})]), \tag{12}$$

where $V_{p'}$ is the conditional variance of $\phi^{(\beta)}(\mathbf{I})$ given $\phi^{(\alpha)}(\mathbf{I}), \alpha = 1, 2, ..., K$ under a distribution p' whose expected feature statistics are between those of p and p_+.

See appendix for proof and discussion.

According to the above proposition, we can use (12) to drive the feature pursuit procedure, which has the following intuitive interpretation. Under the current model p, for any new feature $\phi^{(\beta)}$, $E_p[\phi^{(\beta)}(\mathbf{I})]$ is what we observe from the ensemble governed by p. If $E_p[\phi^{(\beta)}(\mathbf{I})]$ is close to $E_f[\phi^{(\beta)}(\mathbf{I})]$, 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 $E_f[\phi^{(\beta)}(\mathbf{I})]$ is very different from $E_p[\phi^{(\beta)}(\mathbf{I})]$ and including such $\phi^{(\beta)}$ makes the new model $p_+(\mathbf{I};\Lambda_+)$ a better approximation to $f(\mathbf{I})$. The saliency of the new feature is measured by $d(\phi^{(\beta)})$ which is the discrepancy between $E_p[\phi^{(\beta)}(\mathbf{I})]$ and $E_f[\phi^{(\beta)}(\mathbf{I})]$ 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.

Practically we can approximately compute $V_{p'}$ by replacing p' by the current model p. Furthermore, when the feature statistics $E_f[\phi^{(\beta)}(\mathbf{I})] \in \mathcal{B}$ all have the same scale, such as the histograms we will use for texture modeling in the next section, we may further simplify the measure of saliency by l_p-norm distance, i.e.,

$$d(\phi^{(\beta)}) \approx \|E_p[\phi^{(\beta)}(\mathbf{I})] - E_f[\phi^{(\beta)}(\mathbf{I})]\|_p.$$

In practice, this is estimated by

$$d(\phi^{(\beta)}) \approx \|\mu_{obs}^{(\beta)} - \mu_{syn}^{(\beta)}\|_p,$$

where $\mu_{obs}^{(\beta)}$ and $\mu_{syn}^{(\beta)}$ are respectively the sample statistics averaged over the observed images and the synthesized images as discussed in equation (6). This measure is used with $p = 1$ in the texture experiments in the next section.

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

2.5 Estimation error and model complexity

This subsection concerns corrections of the minimum entropy principle and feature pursuit procedure for the presence of the estimation error in the training images. The reader who is not interested in the technical details may skip this subsection.

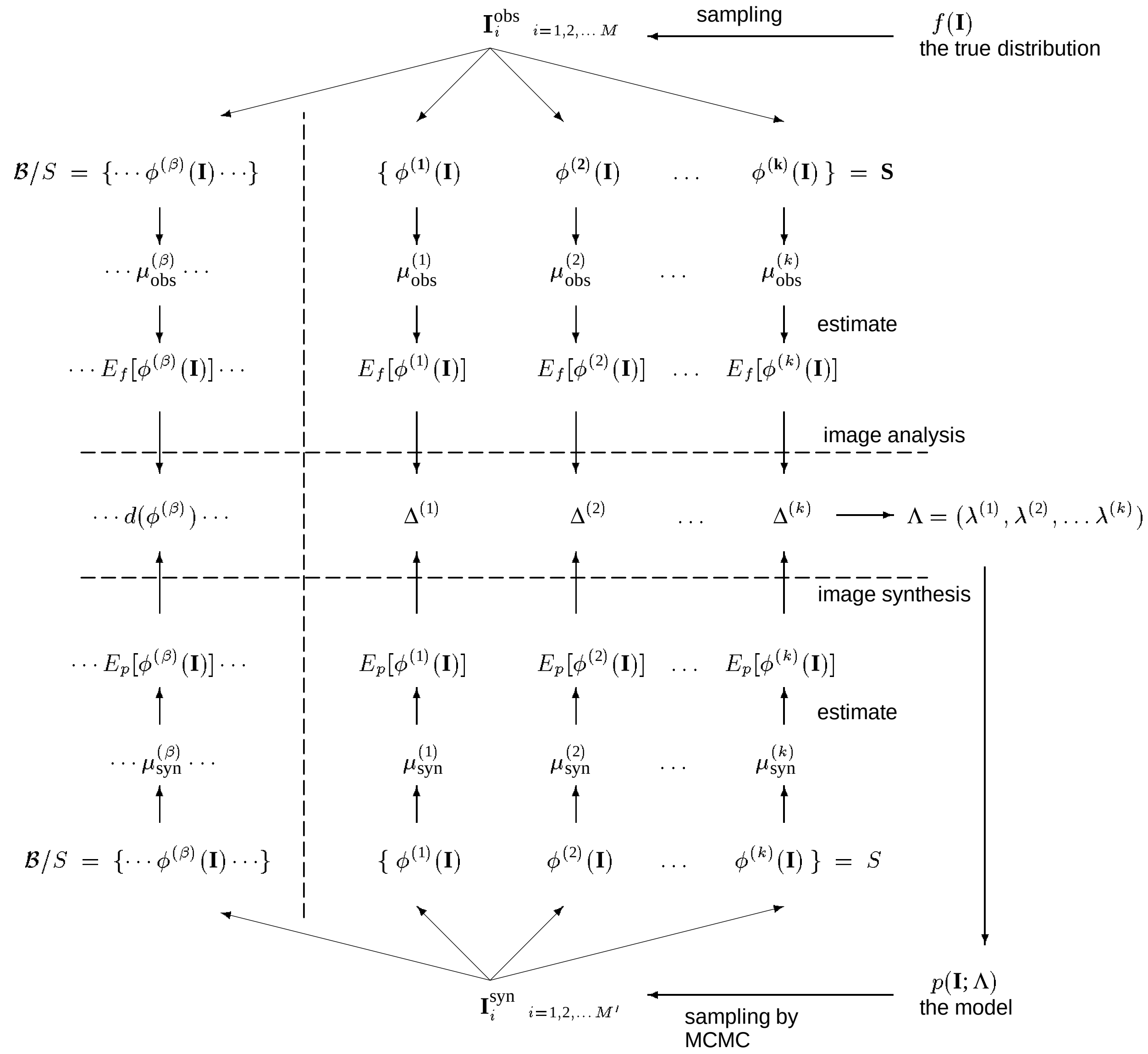
$\mathbf{I}_i^{\text{obs}}\ {}_{i=1,2,\dots M}$
sampling
$f(\mathbf{I})$
the true distribution
$\mathcal{B}/S = \{\cdots\phi^{(\beta)}(\mathbf{I})\cdots\}$
$\{\ \phi^{(\mathbf{1})}(\mathbf{I})\quad \phi^{(\mathbf{2})}(\mathbf{I})\quad \dots\quad \phi^{(\mathbf{k})}(\mathbf{I})\ \} = \mathbf{S}$
$\cdots\mu_{\text{obs}}^{(\beta)}\cdots$
$\mu_{\text{obs}}^{(1)}\quad \mu_{\text{obs}}^{(2)}\quad \cdots\quad \mu_{\text{obs}}^{(k)}$
estimate
$\cdots E_f[\phi^{(\beta)}(\mathbf{I})]\cdots$
$E_f[\phi^{(1)}(\mathbf{I})]\quad E_f[\phi^{(2)}(\mathbf{I})]\quad \dots\quad E_f[\phi^{(k)}(\mathbf{I})]$
image analysis
$\cdots d(\phi^{(\beta)})\cdots$
$\Delta^{(1)}\quad \Delta^{(2)}\quad \dots\quad \Delta^{(k)}$
$\Lambda = (\lambda^{(1)}, \lambda^{(2)}, \dots\lambda^{(k)})$
image synthesis
$\cdots E_p[\phi^{(\beta)}(\mathbf{I})]\cdots$
$E_p[\phi^{(1)}(\mathbf{I})]\quad E_p[\phi^{(2)}(\mathbf{I})]\quad \dots\quad E_p[\phi^{(k)}(\mathbf{I})]$
estimate
$\cdots\mu_{\text{syn}}^{(\beta)}\cdots$
$\mu_{\text{syn}}^{(1)}\quad \mu_{\text{syn}}^{(2)}\quad \cdots\quad \mu_{\text{syn}}^{(k)}$
$\mathcal{B}/S = \{\cdots\phi^{(\beta)}(\mathbf{I})\cdots\}$
$\{\ \phi^{(1)}(\mathbf{I})\quad \phi^{(2)}(\mathbf{I})\quad \dots\quad \phi^{(k)}(\mathbf{I})\ \} = S$
$p(\mathbf{I};\Lambda)$
the model
$\mathbf{I}_i^{\text{syn}}\ {}_{i=1,2,\dots M'}$
sampling by
MCMC

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

In previous subsections, for a set of features $\{\phi^{(\alpha)}, \alpha = 1, ..., K\}$, we have studied two ME distributions. One is $p(\mathbf{I}; \hat{\Lambda})$, which reproduces the observed feature statistics, i.e.,

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

and the other is $p(\mathbf{I}; \Lambda^\star)$, which reproduces the expected feature statistics, i.e.,

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

In the previous derivations, we assume that $\{E_f[\phi^{(\alpha)}(\mathbf{I})], \alpha = 1, ..., K\}$ can be estimated exactly by the observed statistics $\{\mu_{obs}^{(\alpha)}, \alpha = 1, ..., K\}$, 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's consider the minimum entropy principle, which relates the Kullback-Leibler divergence $KL(f, p(\mathbf{I}; \Lambda))$ to the entropy of the model $p(\mathbf{I}; \Lambda)$ for $\Lambda = \Lambda^\star$. Since in practice Λ is estimated at $\hat{\Lambda}$, the goodness of the model should be measured by $KL(f, p(\mathbf{I}; \hat{\Lambda}))$ instead of $KL(f, p(\mathbf{I}; \Lambda^\star))$, and it can be shown that

Proposition 2 *In the above notation,*

$$KL(f, p(\mathbf{I}; \hat{\Lambda})) = KL(f, p(\mathbf{I}; \Lambda^\star)) + KL(p(\mathbf{I}; \Lambda^\star), p(\mathbf{I}; \hat{\Lambda})). \tag{13}$$

See appendix for proof.

That is, because of the estimation error, $p(\mathbf{I}; \hat{\Lambda})$ does not come as close to $f(\mathbf{I})$ as $p(\mathbf{I}; \Lambda^\star)$ does, and the extra noise is measured by $KL(p(\mathbf{I}; \Lambda^\star), p(\mathbf{I}; \hat{\Lambda}))$, which increases with model complexity. In fact, $\hat{\Lambda}$ in model $p(\mathbf{I}; \hat{\Lambda})$ is a random variable depending on the random sample $\{\mathbf{I}_i^{obs}, i = 1, ..., M\}$, so is $KL(f, p(\mathbf{I}; \hat{\Lambda}))$. Let E_{obs} stands for the expectation with respect to the training images, applying E_{obs} to both sides of equation (13), we have,

$$E_{obs}[KL(f, p(\mathbf{I}; \hat{\Lambda}))]$$

$$
\begin{aligned}
&= KL(f, p(\mathbf{I};\Lambda^\star)) + E_{obs}[KL(p(\mathbf{I};\Lambda^\star), p(\mathbf{I};\hat{\Lambda}))] \\
&= entropy(p(\mathbf{I};\Lambda^\star)) - entropy(f) + E_{obs}[KL(p(\mathbf{I};\Lambda^\star), p(\mathbf{I};\hat{\Lambda}))]. \qquad (14)
\end{aligned}
$$

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

Proposition 3 *In the above notation,*

$$entropy(p(\mathbf{I};\Lambda^\star)) = E_{obs}[entropy(p(\mathbf{I};\hat{\Lambda}))] + E_{obs}[KL(p(\mathbf{I};\hat{\Lambda}), p(\mathbf{I};\Lambda^\star))]. \qquad (15)$$

See appendix for proof.

According to Proposition 3, the entropy of $p(\mathbf{I};\hat{\Lambda})$ is on average smaller than the entropy of $p(\mathbf{I};\Lambda^\star)$, this is because $\hat{\Lambda}$ is estimated from each specific training data, and hence $p(\mathbf{I};\hat{\Lambda})$ does better job than $p(\mathbf{I};\Lambda^\star)$ in fitting the training data.

Combining equation (14) and equation (15), we have

$$E_{obs}[KL(f, p(\mathbf{I};\hat{\Lambda}))] = E_{obs}[entropy(p(\mathbf{I};\hat{\Lambda}))] - entropy(f) + C_1 + C_2 \qquad (16)$$

where the two correction terms are

$$C_1 = E_{obs}[KL(p(\mathbf{I};\Lambda^\star), p(\mathbf{I};\hat{\Lambda}))], \qquad C_2 = E_{obs}[KL(p(\mathbf{I};\hat{\Lambda}), p(\mathbf{I};\Lambda^\star))].$$

Following Ripley (1996, Section 2.2), we have

$$C_1 = C_2 = \frac{1}{2M} trace[V_f V_{p^*}^{-1}] + O(M^{-3/2})$$

where $V_f = Var_f[(\phi^{(1)}(\mathbf{I}), ..., \phi^{(K)}(\mathbf{I}))]$, and $V_{p^*} = Var_{p(\mathbf{I};\Lambda^\star)}[(\phi^{(1)}(\mathbf{I}), ..., \phi^{(K)}(\mathbf{I}))]$. V_f and V_{p^*} can be estimated from the observed images and synthesized images respectively. If $V_f \approx V_{p^*}$, then C_1, C_2 are approximately the number of free parameters in the model, i.e., the model complexity, divided by $2M$. Therefore, we have the following form of the Akaike information criterion (Akaike 1977),

$$E_{obs}[KL(f, p(\mathbf{I};\hat{\Lambda}))] \approx E_{obs}[entropy(p(\mathbf{I};\hat{\Lambda}))] - entropy(f) + \frac{1}{M} trace(V_f V_{p^*}^{-1}),$$

where we drop the higher order term $O(M^{-3/2})$. Equation (17) leads to the following correction of the minimum entropy principle.

$$S^* = \arg\min_S \{ \ entropy(p_S(\mathbf{I};\hat{\Lambda})) + \frac{1}{M} trace(V_f V_{p^*}^{-1}) \ \}, \qquad (17)$$

which chooses the optimal set of features over all possible S, where the decrease in entropy by including more features is balanced by the second term which measures the model complexity. 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).

This corrected version of minimum entropy principle leads to a corrected version of the feature pursuit procedure, where at each step, the decrease in $entropy(p(\mathbf{I}; \hat{\Lambda}))$ by introducing a new feature is penalized by the increase in model complexity. The entropy decrease can still be approximated by the quadratic form or even more simply the l_p-norm distance, and the increase in model complexity can be roughly measured by the number of free parameters brought by the new feature. The feature pursuit procedure stops as soon as the entropy decrease does not compensate for the increase in model complexity.

When $\mathbf{I}$ is a homogeneous image, the features $\phi^{(\alpha)}$ themselves can often be expressed as the average of local features $\psi^{(\alpha)}$ over all pixels, where $\psi^{(\alpha)}$ is a function defined on the local windows W centered at $\vec{v} \in \mathcal{D}$, i.e.,

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

Therefore, even if M is small, large sample effects can still take place when the image is large, and asymptotic studies can also be conducted for this situation. However, this is often complicated by phase transition, and we shall not pursue it in this article.

3 Application to texture modeling

This section applies the minimax entropy principle to texture modeling.

3.1 The general problem

Texture is an important characteristic of surface property in visual scenes and is a power cue in visual perception. A general model for textures has long been sought for 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.

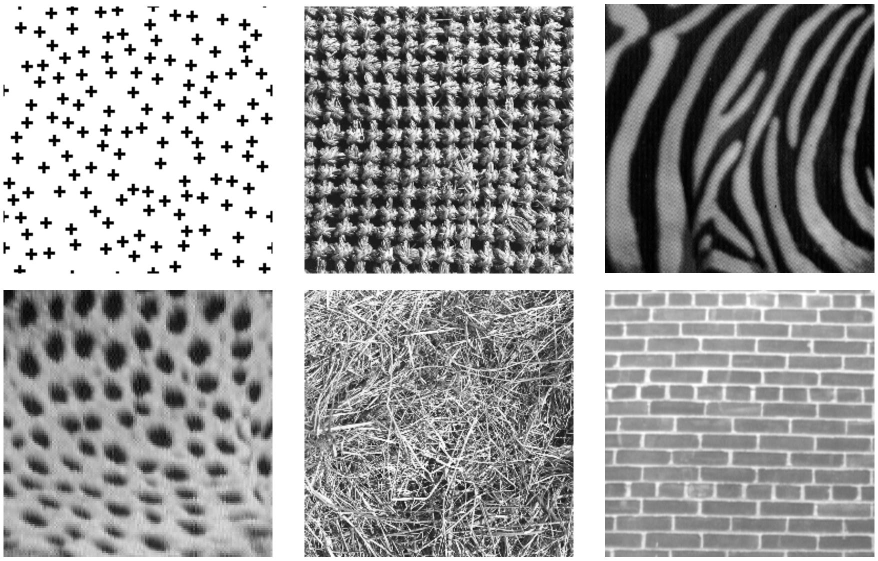

Figure 3: Some typical texture images.

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 and Kass 1991, Picard 1996). 2) The kth-order statistics model for texture perception, i.e. the famous Julesz's conjecture (Julesz 1962). 3) Markov random field models. For a discussion of previous models and methods, the reader is referred to (Zhu, Wu, and Mumford 1996).

In our method, a texture is considered as an ensemble of images of similar texture appearances, and this texture ensemble is governed by a probability distribution

$f(\mathbf{I})$, where $\mathbf{I}$ is defined on a random field $\mathcal{D}$ as discussed in section (2). The objective of texture modeling is to estimate $f(\mathbf{I})$ by building a model $p(\mathbf{I};\Lambda)$ from a set of observed images. Model $p(\mathbf{I};\Lambda)$ should be consistent with human texture perception in the sense that if $p(\mathbf{I};\Lambda)$ estimates $f(\mathbf{I})$ closely, then the sample images from $p(\mathbf{I};\Lambda)$ should be perceptually similar to the training images.

3.2 Choosing features and their statistics

To apply the minimax entropy principle to texture modeling, the first step is to choose features and their statistics, i.e. $\phi^{(\alpha)}(\mathbf{I})$ and $\mu_{obs}^{(\alpha)}$ $\alpha = 1, 2, ..., K$.

Without loss of generality, features of texture images can be extracted by "filters" $F^{(\alpha)}$, 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 $\vec{v} \in \mathcal{D}$, i.e., $\mathbf{I}^{(\alpha)}(\vec{v}) = F^{(\alpha)}(\mathbf{I}_{W+\vec{v}})$ is a function depending on the intensities inside window W centered at $\vec{v}$, 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 replaced by

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

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

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

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

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

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

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

There are two motivations for us to choose the histograms of filtered images as feature statistics. The first comes from the following mathematical result.

Theorem 2 *Let $f(\mathbf{I})$ be the continuous probability distribution of an $N \times N$ texture image, Then $f(\mathbf{I})$ is determined by the marginal distributions $f^{(\alpha)}(z)$.*

See appendix for a proof.

Therefore if we choose $\mu_{obs}^{(\alpha)}(z) \approx f^{(\alpha)}(z) = E_f[H^{(\alpha)}(\mathbf{I}, z)]$ as the observed statistics, and $p(\mathbf{I}; \Lambda)$ is an ME distribution so that $E_{p(\mathbf{I};\Lambda)}[H^{(\alpha)}(\mathbf{I})] = E_f[H^{(\alpha)}(\mathbf{I})]$ for all possible α, then $p(\mathbf{I}; \Lambda) = f(\mathbf{I})$. But this will involve uncountable number of filters $F^{(\alpha)}$ and each filter is as big as the image $\mathbf{I}$.

However, recent psychophysical research on human texture perception suggests that two homogeneous textures are often difficult to discriminate when they produce similar *marginal distributions* of responses from *a bank of filters* (Bergen and Adelson 1991, Chubb and Landy 1991). This means that it is plausible to ignore some statistical properties of $f(\mathbf{I})$ which are not important for human texture discrimination.

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). We limit our model to homogeneous textures, thus $f(\mathbf{I})$ is stationary with respect to location $\vec{v}$. 2). All features which 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. 3). Only a finite set of filters are used. Although we often have access to only one or a few training texture images, assumption 1) and 2) enable ergodicity takes effects, so that the observed histograms of the filter images make reasonable estimates for the marginal distributions of $f(\mathbf{I})$.

Substituting $H^{(\alpha)}(\mathbf{I})$ for $\phi^{(\alpha)}(\mathbf{I})$ in equation (2), we obtain

$$p(\mathbf{I}; \Lambda) = \frac{1}{Z(\Lambda)} \exp\{-\sum_{\alpha=1}^{K} < \lambda^{(\alpha)}, H^{(\alpha)}(\mathbf{I}) >\}, \tag{18}$$

which we call the FRAME (Filter, Random field, And Minimax Entropy) model. Here the angle brackets indicates that we are taking a sum over bin z: i.e., $< \lambda^{(\alpha)}, H^{(\alpha)}(\mathbf{I}) >= \sum_z \lambda_z^{(\alpha)} H^{(\alpha)}(\mathbf{I}, z)$.

The computation of the parameters Λ and the selection of filters $F^{(\alpha)}$ proceed as described in the last section. Detailed analysis of the texture modeling algorithm is referred to (Zhu, Wu, Mumford 1996).

3.3 FRAME: a new class of MRF models

In this section, we derive a continuous form for the FRAME model in equation (18), and compare it with existing MRF models.

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

$$E_{p(\mathbf{I};\Lambda)}[\frac{1}{|\mathcal{D}|} \sum_{\vec{v} \in \mathcal{D}} \delta(z - \mathbf{I}^{(\alpha)}(\vec{v}))] = \mu_{obs}^{(\alpha)}(z), \quad \forall z \in \mathbf{R},\ \forall \vec{v} \in \mathcal{D},\ \forall \alpha. \tag{19}$$

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

$$\begin{aligned} p(\mathbf{I}; \Lambda) &= \frac{1}{Z(\Lambda)} \exp\{-\sum_{\alpha=1}^{K} \sum_{\vec{v} \in \mathcal{D}} \int \lambda^{(\alpha)}(z) \frac{1}{|\mathcal{D}|} \sum_{\vec{v} \in \mathcal{D}} \delta(z - \mathbf{I}^{(\alpha)}(\vec{v})) dz\} \\ &= \frac{1}{Z(\Lambda)} \exp\{-\sum_{\alpha=1}^{K} \sum_{\vec{v} \in \mathcal{D}} \lambda^{(\alpha)}(\mathbf{I}^{(\alpha)}(\vec{v}))\}. \end{aligned} \tag{20}$$

Since z is a continuous variable, there are infinite number of constraints, therefore the Lagrange multipliers $\Lambda = (\lambda^{(\alpha)}(), \alpha = 1, ..., K)$ 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 (20) as,

$$p(\mathbf{I}; \Lambda) = \frac{1}{Z(\Lambda)} \exp\{-\sum_{\alpha=1}^{K} \sum_{\vec{v}} \lambda^{(\alpha)}(F^{(\alpha)} * \mathbf{I}(\vec{v}))\} \tag{21}$$

Clearly, equation (20) and (21) are Markov Random Field (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 learnt from the training images, so they can incorporate high order statistics and thus model non-Gaussian properties of images. The FRAME model has much stronger expressive power than existing MRF models which are based on pair cliques, and at the same time, the complexity of the model is under check since 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 (3.5)). 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 are referred to (Zhu, Wu, Mumford 1996).

3.4 Designing a filter bank

To describe a wide variety of textures, we need to specify a general filter bank, from which filters can be selected when describing a certain texture. This filter bank serves as the "vocabulary", and the selected filters can be considered as "words", 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 et al. 1989).

1) The intensity filter, i.e., $\delta()$, 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

$$LG(x, y \mid T) = const \cdot (x^2 + y^2 - T^2)e^{-\frac{x^2+y^2}{T^2}}, \tag{22}$$

where $T = \sqrt{2}\sigma$ controls the scales of the filters. We choose eight scales with $T = \sqrt{2}, 1, 2, 3, 4, 5, 6$. 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

$$Gabor(x, y \mid T, \theta) = const \cdot e^{\frac{1}{2T^2}(4(xcos\theta+ysin\theta)^2+(-xsin\theta+ycos\theta)^2)}e^{-i\frac{2\pi}{T}(xcos\theta+ysin\theta)}, \tag{23}$$

where T controls the scales and θ controls the orientations. We choose 6 scales $T = 2, 4, 6, 8, 10, 12$ and 6 orientations $\theta = 0^o, 30^0, 60^o, 90^o, 120^o, 150^o$. One can 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 $Gsin(T, \theta)$ and $Gcos(T, \theta)$ respectively.

4) The non-linear 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, $\mid Gabor(x, y \mid T, \theta) * I \mid^2$, which, in fact, is the local spectrum of $\mathbf{I}$ at (x, y) smoothed by a Gaussian function, and therefore such filters serve as local spectrum analyzers.

5) Some specially designed filters for texton primitives, see subsection (3.5).

3.5 Experiments of texture modeling

This section describes the modeling of natural textures using the algorithm studied in section (2), and the first texture image is described in details in order to illustrate how the filter pursuit algorithm works.

Figure (4.a) is an observed image of animal fur. We starts from the uniform white noise image, which is displayed in figure (4.b). Then the algorithm picks up the first filter, which is a 5×5 pixels Laplacian of Gaussian filter with scale $T = 1.0$, and which has the largest entropy decrease ($d(\phi^{(\beta)}) = 0.611$) among all the filters in the filters bank. Then a texture image is synthesized by matching the histogram of the filter response and is shown in figure (4.c).

Comparing figure (4.c) with figure (4.b), it is evident that this filter captures local smoothness features of the observed texture image. Continuing the algorithm, 5 more filters are sequentially added, which are, respectively, 2) $Gcos(6.0, 120^o)$, 3) $Gcos(2.0, 30^o)$, 4) $Gcos(12, 60^o)$, 5) $Gcos(10.0, 120^o)$, and 6) intensity $\delta()$, each of which captures features at various scales and orientations. The $d(\phi^{(\beta)})$, i.e., the measure of entropy decrease for these filters are respectively $0.424, 0.207, 0.132, 0.157, 0.059$ and the texture images synthesized using $2, 3, 6$ filters are displayed in figure (4.d, 4.e, 4.f). Obviously, with more filters added, the synthesized texture image gets closer to the observed one. It appears that the filters chosen in later steps make less contributions to $p(\mathbf{I})$, and thus confirms our early assumption that the marginal distributions of a small number of filtered images should be adequate for capturing the essential features of the underlying probability distribution $f(\mathbf{I})$.

Figure (5.a) 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 (5.b) is the synthesized texture image using 5 filters.

Figure (6.a) is an image taken from the skin of a cheetah, and figure (6.b) displays the synthesized texture using 6 filters. One may notice that the original 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 (6.b), 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 (7.a) shows a texture of fabric which has clear periods along both horizontal and vertical directions. We want to use this texture to test the use of non-linear filters, so we choose 2 spectrum analyzers, one in the horizontal direction, the other

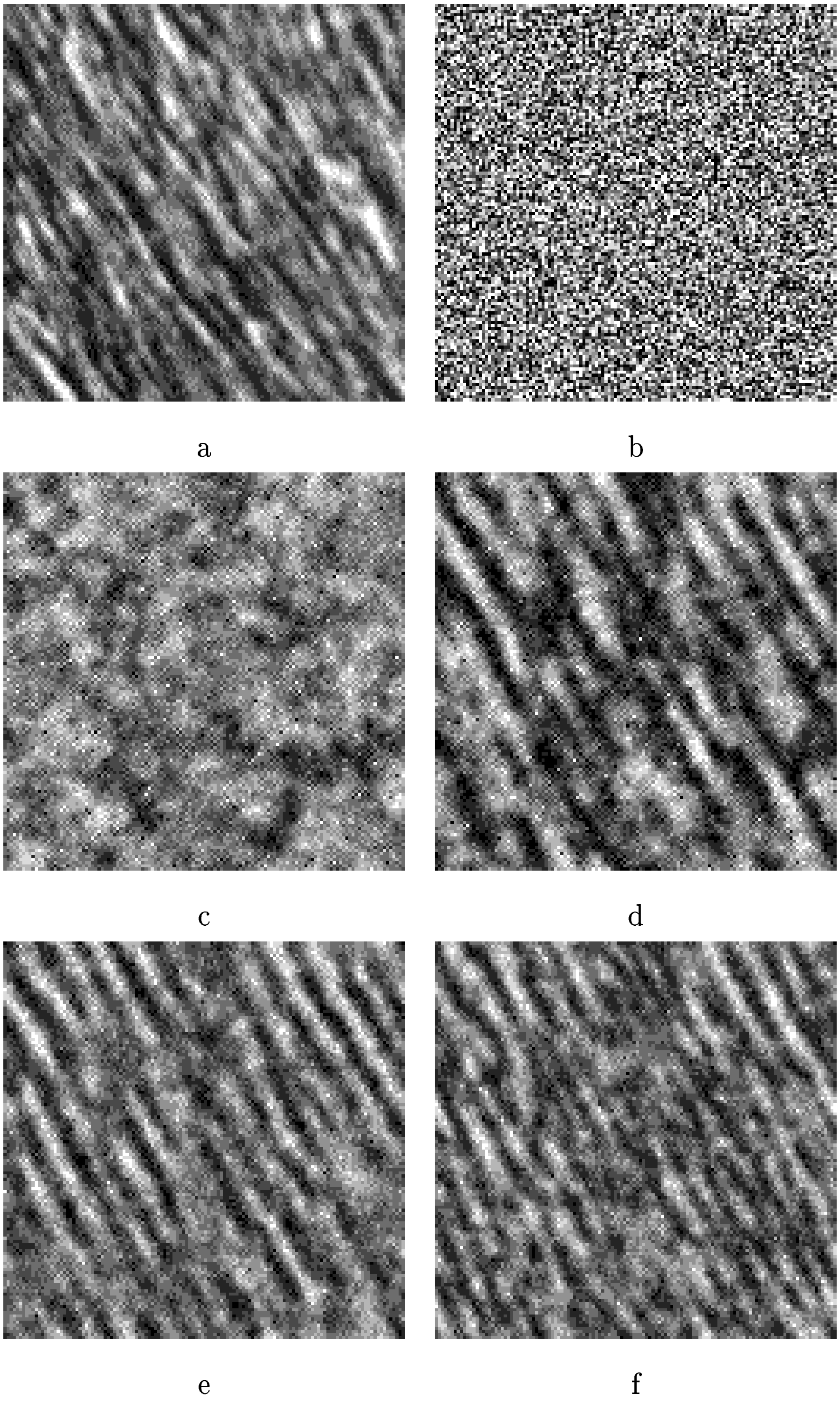

Figure 4: Synthesis of the fur texture. a. the observed image, b),c),d),e),f) are the synthesized images using $0, 1, 2, 3, 6$ filters respectively.

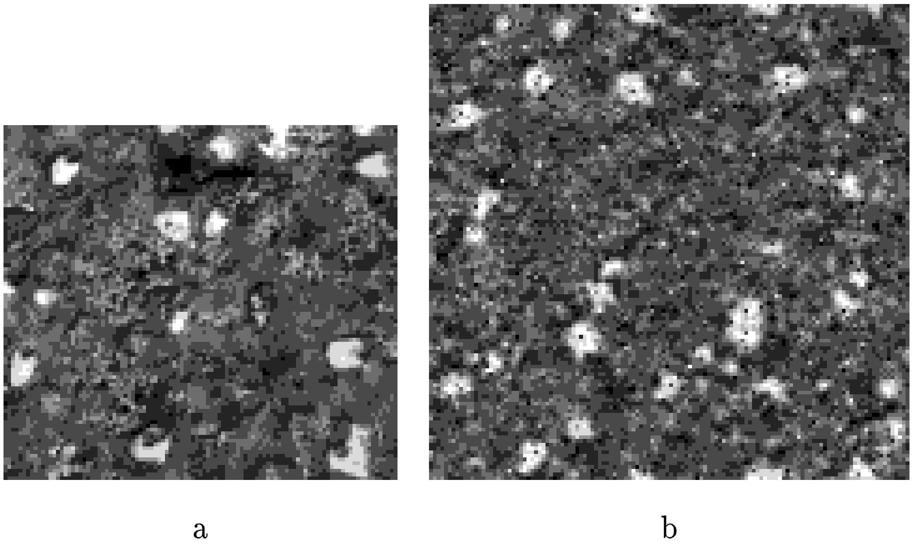

Figure 5: a. the observed texture–mud, b, the synthesized one using 5 filters

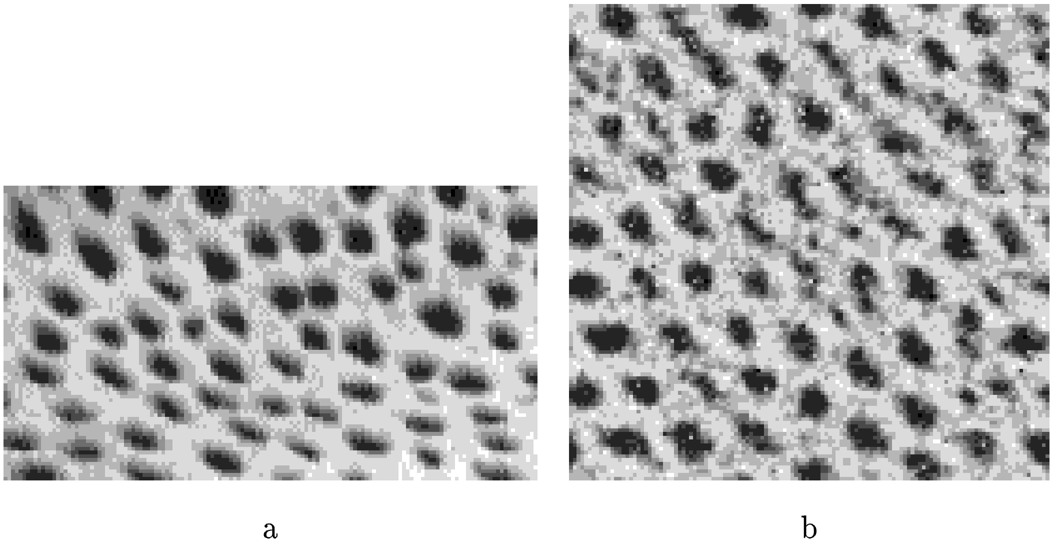

Figure 6: a) the observed texture image – cheetah blob. b) the synthesized one using 6 filters

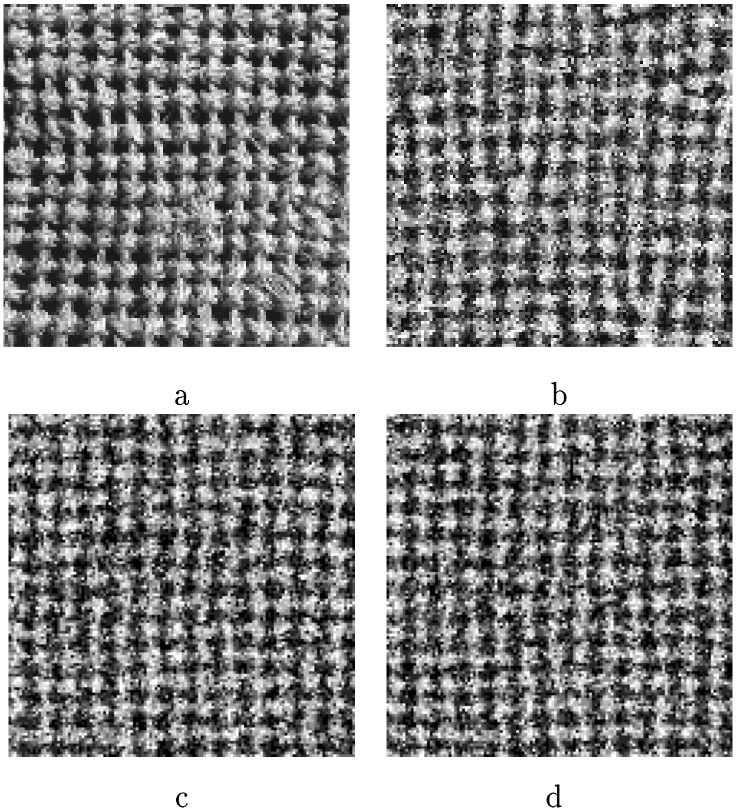

Figure 7: a. the input image of fabric, b. the synthesized image with 2 pairs of Gabor filters plus the Laplacian of Gaussian filter. c,d two more images sampled at different steps of the Gibbs sampler.

in the vertical direction with their periods 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 a Laplacian of Gaussian filter $LG(\sqrt{2}/2)$ with window size 3 × 3 pixels, to take care of the intensity histogram and the smoothness features. Three synthesized texture images are displayed in figure (7.b, 7.c, 7.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.

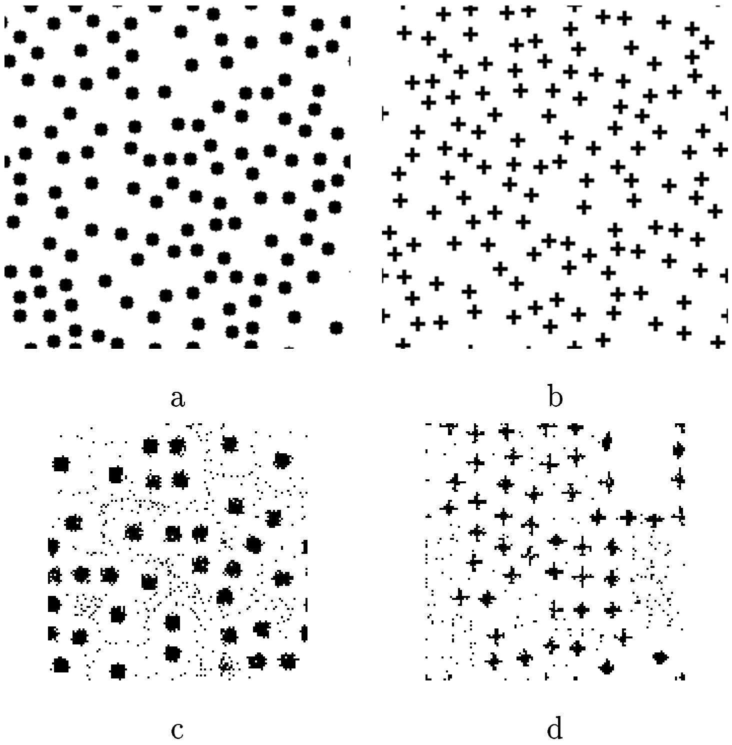

Figure 8: Two typical texton images of 256 × 256 pixels. a) circle, b) cross, c). and d) are the two synthesized images of 128 × 128 pixels.

Figure (8.a) and (8.b) 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 15×15 pixels mask with the corresponding primitive at the center. With this filter selected, the FRAME algorithm starts from a uniform white noise image, and gradually matches the histogram of the filter responses from the simulated image to the histogram obtained from the observed image. However, with an examination of figure (1.b), which plots the histograms obtained from the observed image (solid curve) and that from a uniform noise image (dotted curve), one can observe that there are many isolated peaks in observed histogram, which set up “potential wells”, so that it is very unlikely to change a filter response from one peak to another during the FRAME algorithm which flips one pixel at a time, and therefore it will take a long time for the algorithm to match the histograms. In these experiments, we proposed an annealing approach which smooth the histograms and matching the smoothed histogram first and gradually the target histogram becomes less and less smoothed. Some details of this heuristics is referred to (Zhu,Wu,Mumford 1996). Figure (8.c), and (8.d) are two synthesized images.

3.6 More on texture modeling

There are various artificial categories for textures with respect to various attributes, such as Fourier and non-Fourier, deterministic and stochastic, and macro- and micro-textures. 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, micro-textures. 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), thus FRAME realizes the full potential of MRF models.

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 has the following advantages. Firstly we obtain a probability model $p(\mathbf{I}; \Lambda)$ instead of merely synthesizing texture images. Secondly the Monte Carlo Markov chain for model estimation and texture sampling is guaranteed to converge to a stationary process which follows the estimated distribution $p(\mathbf{I}; \Lambda)$ (Geman and Geman 1984), and the observed histograms can be matched closely. Thirdly a theoretical proof is provided to show that if the marginal distributions of filter responses for all linear filters under $f(\mathbf{I})$ are matched, then we eventually obtain the underlying model, i.e., $p(\mathbf{I}; \Lambda) = f(\mathbf{I})$. But the FRAME model is computationally very expensive. Approaches for further facilitating the computation are yet to be developed, for more discussion in this aspect, the reader is referred to (Zhu, Wu, and Mumford 1996).

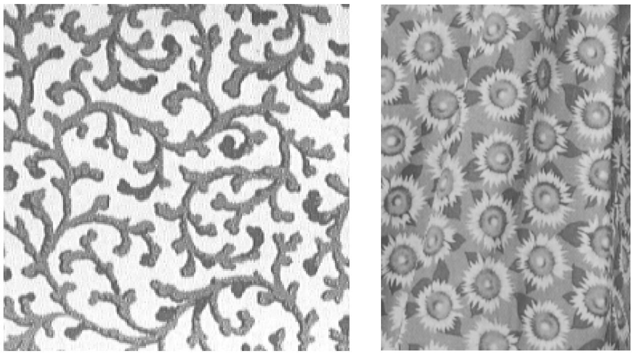

Figure 9: Two challenging texture images.

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 or high-level features than those we have used in this work, and these high-level features may correspond to high-level visual process such as the geometrical properties of object shape. In this paper, we choose filters from a fixed set of filters, but in general it is not understood that how to design such set of features or structures for an arbitrary applications.

4 Discussion

This paper 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 best characterize the observed images. Algorithms are proposed for parameter estimation and stochastic simulation. A greedy algorithm is developed for feature pursuit, and the model complexity is studied in the presence of sample variations.

The minimax entropy principle is applied to texture modeling where the feature extracted from images are the empirical marginal distributions (or histograms) of filtered images. A new MRF model– FRAME is derived, and the experiments described in section (3.5) demonstrate that our method is capable of modeling a wide variety of textures which are previously considered as belonging to different categories. The results of texture modeling support the psychological experiments which suggest that two texture images are often difficult to discriminate if they produce similar empirical marginal distributions for filter responses from a bank of filters (Bergen and Adelson 1991, Chubb and Landy 1991).

Our theory and methodology also contributes to possible image representing strategy in our brain after a biologically-plausible Gabor filter analysis. 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. The maximum entropy phase of the algorithm, from a computational standpoint, consists mainly in 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 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 paper. That theory further postulated that cortex seeks out the 'residuals', the features of the observed image which are different from those of the mental image. The present 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 a) the massively parallel architecture, b) the apparent stochastic component in neural firing and c) 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 'pre-attentive' texture discrimination. We propose that textures that can be pre-attentively 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 pre-attentively discriminate new sets of textures is exactly that of incorporating a new filter feature into the model. Evidence that texture pairs which are not pre-attentively 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.

Recently, the minimax entropy principle has been applied to model general natural images (Zhu and Mumford 1996). It is our hope that this work will simulate future research efforts in this direction.

Appendix: mathematical details

1) Proof of Theorem 1: Since $E_{p(\mathbf{I};\Lambda^\star)}[\phi^{(\alpha)}(\mathbf{I})] = E_f[\phi^{(\alpha)}(\mathbf{I})], \;\; \alpha = 1, ..., K.$

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

and the result follows. □

2) Proof of Proposition 1: Let $\Phi(\mathbf{I}) = (\phi^{(1)}(\mathbf{I}), ..., \phi^{(K)}(\mathbf{I}))$. Clearly, $E_p[\Phi(\mathbf{I})] = E_{p_+}[\Phi(\mathbf{I})] = E_f[\Phi(\mathbf{I})]$. Let $\Phi_+ = (\Phi(\mathbf{I}), \phi^{(\beta)}(\mathbf{I}))$. By a Taylor expansion argument (Corollary 4.4 of Kullback 1959, page 48), the entropy decrease is

$$\begin{aligned} d(\phi^{(\beta)}) &= D(p_+;p) \\ &= \frac{1}{2}(E_{p_+}[\Phi_+(\mathbf{I})] - E_p[\Phi_+(\mathbf{I})])' Var_{p'}[\Phi_+(\mathbf{I})]^{-1}(E_{p_+}[\Phi_+(\mathbf{I})] - E_p[\Phi_+(\mathbf{I})]) \\ &= \frac{1}{2}(E_f[\phi^{(\beta)}(\mathbf{I})] - E_p[\phi^{(\beta)}(\mathbf{I})]) V_{p'}^{-1}(E_f[\phi^{(\beta)}(\mathbf{I})] - E_p[\phi^{(\beta)}(\mathbf{I})]), \end{aligned}$$

where p' is a distribution whose expected feature statistics are between those of p and p_+, $V_{11} = Var_{p'}[\Phi(\mathbf{I})]$, $V_{22} = Var_{p'}[\phi^{(\beta)}(\mathbf{I})]$, $V_{12} = Cov_{p'}[\Phi(\mathbf{I}), \phi^{(\beta)}(\mathbf{I})]$, and $V_{p'} = V_{22} - V_{12}' V_{11}^{-1} V_{12}$ is the conditional variance. Hence the result follows. □

The conditional variance $V_{p'}$ can also be interpreted as follows. Let $C = -V_{11}^{-1}V_{12}$, and let $\phi_\perp^{(\beta)}(\mathbf{I}) = \phi^{(\beta)}(\mathbf{I}) + C'\Phi(\mathbf{I})$ being 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 $\Phi(\mathbf{I})$, i.e., $\phi_\perp^{(\beta)}(\mathbf{I})$ is an orthogonalization of $\phi^{(\beta)}(\mathbf{I})$ with respect to $\Phi(\mathbf{I})$. Then we have $V_{p'} = Var_{p'}[\phi_{\perp}^{(\beta)}(\mathbf{I})]$, i.e., $V_{p'}$ is the variance of $\phi^{(\beta)}(\mathbf{I})$ with its dependence on $\Phi(\mathbf{I})$ being eliminated.

3) Proof of Proposition 2: From the proof of Theorem 1, we know $E_f[\log p(\mathbf{I};\Lambda^\star)] = E_{p(\mathbf{I};\Lambda^\star)}[\log p(\mathbf{I};\Lambda^\star)]$, and by similar derivation we have $E_f[\log p(\mathbf{I};\Lambda)] = E_{p(\mathbf{I};\Lambda^\star)}[\log p(\mathbf{I};\Lambda)]$ for any Λ. Thus

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

So the result follows by setting $\Lambda = \hat{\Lambda}$. The above derivation also shows that $p(\mathbf{I};\Lambda^\star)$ best approximates $f(\mathbf{I})$ among all possible $p(\mathbf{I};\Lambda)$. □

4) Proof of Proposition 3 As in subsection 2.2, let

$$L_{obs}(\Lambda) = \frac{1}{M}\sum_{i=1}^{M} \log p(\mathbf{I}_i^{obs};\Lambda)$$

be the log-likelihood function, where we use the subscript *obs* to emphasize the fact that $L_{obs}(\Lambda)$ is a random variable depending on the observed images. First, we show

$$\begin{aligned}
L_{obs}(\hat{\Lambda}) &= \frac{1}{M}\sum_{i=1}^{M}\{-\log Z(\hat{\Lambda}) - \sum_{\alpha=1}^{K} < \hat{\lambda}^{(\alpha)}, \phi^{(\alpha)}(\mathbf{I}_i^{obs}) >\} \\
&= -\log Z(\hat{\Lambda}) - \sum_{\alpha=1}^{K} < \hat{\lambda}^{(\alpha)}, \mu_{obs}^{(\alpha)} > \\
&= -\log Z(\hat{\Lambda}) - \sum_{\alpha=1}^{K} < \hat{\lambda}^{(\alpha)}, E_{p(\mathbf{I};\hat{\Lambda})}[\phi^{(\alpha)}(\mathbf{I})] > \\
&= -entropy(p(\mathbf{I};\hat{\Lambda})).
\end{aligned}$$

By similar derivation, we can prove that $E_{obs}[L_{obs}(\Lambda^\star)] = -entropy(p(\mathbf{I};\Lambda^\star))$ and

$$L_{obs}(\hat{\Lambda}) - L_{obs}(\Lambda^\star) = KL(p(\mathbf{I};\hat{\Lambda}), p(\mathbf{I};\Lambda^\star)). \qquad (24)$$

Applying E_{obs} to both sides of equation (24), we have

$$-E_{obs}[entropy(p(\mathbf{I};\hat{\Lambda}))] + entropy(p(\mathbf{I};\Lambda^\star)) = E_{obs}[KL(p(\mathbf{I};\hat{\Lambda}), p(\mathbf{I};\Lambda^\star))],$$

and the result follows. □.

5) Proof of Theorem 2: $f(\mathbf{I})$ can be connected to $f^{(\alpha)}(z)$ via the Fourier transform. First, an application of the inverse Fourier transform gives

$$f(\mathbf{I}) = \frac{1}{(2\pi)^{N^2}} \int \cdot \int e^{2\pi i <I,\ F>} \hat{f}(F) dF$$

where F is a vector of the same size as $\mathbf{I}$, and $\hat{f}(F)$ is the characteristic function of $f(\mathbf{I})$,

$$\begin{aligned}
\hat{f}(F) &= \int \cdot \int e^{-2\pi\, i <F,\, \mathbf{I}>} f(\mathbf{I}) d\mathbf{I} \\
&= \int e^{-2\pi\, iz} dz \int \cdot \int_{<F,\, \mathbf{I}>=z} f(\mathbf{I}) d\mathbf{I} \\
&= \int e^{-2\pi\, iz} dz \int \cdot \int \delta(< F, I > -z) f(\mathbf{I}) d\mathbf{I} \\
&= \int e^{-2\pi\, iz}\ f^{(\alpha)}(z) dz
\end{aligned}$$

where $< \cdot, \cdot >$ is the inner product, $f^{(\alpha)}(z) = \int \cdot \int \delta(< F^{(\alpha)}, \mathbf{I} > -z) f(\mathbf{I}) d\mathbf{I}$ is the marginal distribution of $< F^{(\alpha)}, \mathbf{I} >$, with $F^{(\alpha)}$ being a specific linear filter, and α the index of filters. So the result 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

[1] Akaike, H, "On entropy maximization principle", In *Applications of Statistics*, ed. Krishnaiah, P.R., p27-42, Amsterdam: North-Holland, 1977.

[2] Barlow, H.B., Kaushal, T.P. and Mitchison, G.J., "Finding minimum entropy codes." *Neural Computation.* Vol. 1, pp412-423, 1989.

[3] Bergen, J. R. and Adelson, E. H. "Theories of visual texture perception." *Spatial Vision* D.Regan (eds.), CRC press, 1991.

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

[5] Blake, A, and Zisserman, A. *Visual reconstruction.* MIT press, 1987.

[6] Chubb, C. and Landy, M. S. "Orthogonal distribution analysis: a new approach to the study of texture perception." *Comp. Models of Visual Proc.* Landy. etc (ed.) MIT press 1991.

[7] Coifman, R.R. and Wickerhauser, M.V.. "Entropy based algorithms for best basis selection." *IEEE Trans. on Information Theory.*, Vol.38,pp713-718, 1992.

[8] Cross, G. R. and Jain, A. K. "Markov random field texture models." *IEEE, PAMI*, **5**, 25-39. 1983.

[9] Daugman, J. "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of Optical Soc. Amer.* Vol.2, No.7, 1985.

[10] Dayan, P., Hinton, G.E., Neal, R.N., and Zemel, R.S. "The Helmholtz machine", *Neural Computation.* 1995.

[11] Donoho, D.L. and Johnstone, I.M. "Ideal de-noising in an orthonormal basis chosen from a library of bases. " *Acad.Sci.Paris, Ser I*, Vol.319, pp1317-1322,1994.

[12] Field, D. "What is the goal of sensory coding?", *Neural Computation.* 6, 559-601, 1994.

[13] Gabor, D. "Theory of communication." IEE Proc.vol 93, no26, 1946.

[14] Geman, S. and Geman, D. "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images." *IEEE Trans. PAMI,* 6, 721-741. 1984.

[15] Heeger, D. J. and Bergen, J. R. "Pyramid-based texture analysis/synthesis." *Computer Graphics*, in press. 1995.

[16] Jaynes, E.T. "Information theory and statistical mechanics". *Physical Review.* **106**, pp620-630, 1957.

[17] Jolliffe, I. T. *Principle Components Analysis.* New York: Springer, 1986.

[18] Jordan, M.I. and Jacobs, R.A., "Hierarchical mixtures of experts and the EM algorithm", *Neural Computation.* 6,181-214, 1994.

[19] Julesz, B. "Visual pattern discrimination." *IRE Transactions of Information Theory* IT-8, pp84-92,1962.

[20] Julesz, B. *Dialogues on Perception*, 1995.

[21] Karni, A and Sagi, D, " Where practice makes perfect in texture discrimination –evidence for primary visual cortex plasticity", *Proc. Nat. Acad. Sci. US,* vol. 88, 4966-4970, 1991.

[22] Kullback, S. and Leibler, R. A. "On information and sufficiency", *Annual Math. Stat.* vol.22, pp79-86, 1951.

[23] Kullback, S. *Information Theory and Statistics.* John and Wiley and Sons, Inc. New York, 1959.

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

[25] Mumford, D.B. and Shah, J. "Optimal Approximations by Piecewise Smooth Functions and Associated Variational Problems." *Comm. Pure Appl. Math.,* 42, pp 577-684. 1989.

[26] Mumford, D.B. " On the Computational Architecture of the Neocortex II: The role of cortico-cortical loops", *Biological Cybernetics*, **66**, pp. 241-251, 1992.

[27] Picard, R. W. 1996. "A society of models for video and image libraries", MIT media lab. technical report No.360.

[28] Priestley, M. B. (1981) *Spectral Analysis and Time Series,* Academic Press.

[29] Rissanen, J. *Stochastic Complexity in Statistical Inquiry.* Singapore, World Scientific, 1989.

[30] Silverman, M. S., Grosof, D. H., De Valois, R. L., and Elfar, S. D. "Spatial-frequency organization in primate striate cortex." *Proc. Natl. Acad. Sci. U.S.A.*, **86**. 1989.

[31] Simoncelli, E. P. Freeman,W.T, Adelson, E. H, Heeger, D.J. "Shiftable multi-scale transforms. *IEEE Trans. on information theory.* Vol. 38, pp587-607, 1992.

[32] Winkler, G. *Image Analysis, Random Fields and dynamic Monte Carlo Methods*, Springer-Verlag 1995.

[33] Witkin, A. and Kass, M. "Reaction-diffusion textures." *Computer Graphics*, **25**, 299-308. 1991.

[34] Xu, L. "Ying-Yang machine: a Bayesian-Kullback scheme for unified learnings and new results on vector quantization", *Proc. Int'l Conf. on Neural Info. Proc.* 1995.

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

[36] Zhu, S.C. and Mumford, D.B. "Learning generic prior models for visual computation." *Harvard Robotics Lab. Technique Report. 95-03* 1995.

[37] Zhu, S.C. *Statistical and Computatinal Theories for Image Segmentation, Texture Modeling and Object Recognition.* Ph.D Thesis, Harvard University, 1996.

[38] Zhu, S.C., Wu, Y.N. and Mumford,D.B. " FRAME: Filters, Random fields And Maximum Entropy—to a unified theory for texture modeling ", *Harvard Robotics Lab. Technique Report.* 1996.