IEEE Trans. on **Pattern Analysis and Machine Intelligence**, short version CVPR97.).

Learning Generic Prior Models for Visual Computation

Song Chun Zhu and David Mumford
Division of Applied Math
Box F, Brown University.
Providence, RI 02912.

Abstract

Many generic prior models have been widely used in computer vision ranging from image and surface reconstruction to motion analysis, and these models presume that surfaces of objects be smooth, and adjacent pixels in images have similar intensity values. However, there is little rigorous theory to guide the construction and selection of prior models for a given application. Furthermore, images are often observed at arbitrary scales, but none of the existing prior models are scale-invariant. Motivated by these problems, this article chooses general natural images as a domain of application, and proposes a theory for learning prior models from a set of observed natural images. Our theory is based on a maximum entropy principle, and the learned prior models are of Gibbs distributions. A novel information criterion is proposed for model selection by minimizing a Kullback-Leibler information distance. We also investigate scale invariance in the statistics of natural images and study a prior model which has scale invariant property. In this paper, in contrast with all existing prior models, negative potentials in Gibbs distribution are first reported. The learned prior models are verified in two ways. Firstly images are sampled from the prior distributions to demonstrate what typical images they stand for. Secondly they are compared with existing prior models in experiments of image restoration.

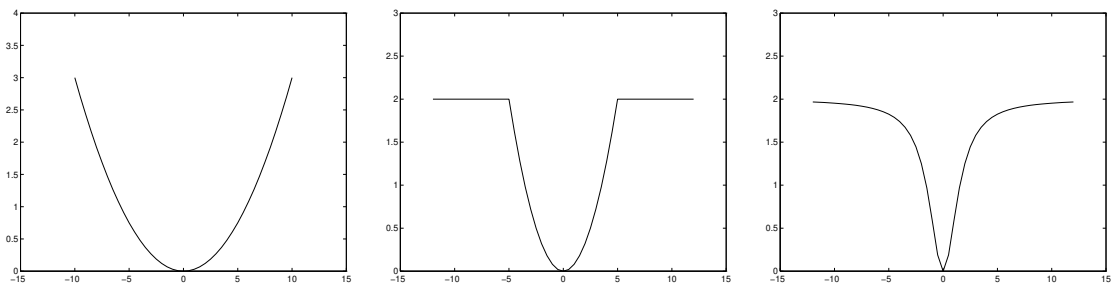
1 Introduction and motivation

In computer vision, many generic smoothness models have been taken for grant, and these models presume that surfaces of objects be smooth, and adjacent pixels in images have similar intensity values. Such models are often explained as prior probability distributions in terms of Bayesian statistics, and they play an important role in visual computation ranging from image restoration and segmentation, motion analysis, to 3D surface reconstruction.

For example, In image segmentation and restoration (Geman and Geman 1984, Blake and Zisserman 1987, Mumford and Shah 1989, Geman and McClure 1987), let $\mathbf{I}$ be an image defined over a lattice $\mathbf{S}$, then the generic prior models are expressed as the following joint probability distribution,

$$p(\mathbf{I}) = \frac{1}{Z} e^{-\sum_{(x,y)} \psi(\nabla_x \mathbf{I}(x,y)) + \psi(\nabla_y \mathbf{I}(x,y))} \tag{1}$$

where Z is a normalization factor, the summation is over all pixels $(x, y) \in \mathbf{S}$. ∇_x , ∇_y are differential operators, $\nabla_x \mathbf{I}(x, y) = \mathbf{I}(x+1, y) - \mathbf{I}(x, y)$, and $\nabla_y \mathbf{I}(x, y) = \mathbf{I}(x, y+1) - \mathbf{I}(x, y)$. Some typical forms of the potential function $\psi()$ are displayed in figure (1). The functions in figure (1.b) (1.c) have flat tails to preserve edges and object boundaries, and thus they are said to have advantages over the function of figure (1.a). Similar prior models are used in 3D surface reconstruction, and details of those models are referred to (Terzopoulos 1983, Belhumeur 1993).



a b c

Figure 1: Three existing forms for $\psi()$. a, Quadratic: $\psi(x) = ax^2$. b, Line process: $\psi(x) = ax^2$ if $\mid x \mid < \theta$, and $\psi(x) = a\theta^2$ otherwise. c, T-function: $\psi(x) = -\frac{1}{1+cx^2} + a$.

We observe that all existing generic prior models share three common properties.

1. They use differential operators to capture image smoothness. In discrete image lattice, these differential operators are linear filters.

2. They are expressed as Gibbs distributions, with potentials built on the responses of the above linear filters.

3. They are translation invariant (homogeneous) with respect to location (x, y).

Most of these prior models enjoy nice explanations in terms of regularization theory (Poggio, Torre, and Koch 1985),[1] physical modeling (Terzopoulos 1983),[2] Bayesian theory (Geman and Geman 1984, Blake and Zisserman 1987, Mumford and Shah 1989) and robust statistics (Black and Rangarajan 1994), and they do capture some of the basic intuitions of images. However, these prior models are essentially generalized from the traditional Ising model and Potts model for studying spin systems in physics (see Winkler 1995 for details), there is no obvious reason to believe that they will be appropriate for model real world images. Recently, *ad hoc* prior models are often chosen for mathematical convenience (Shah 1996).

In general, there is little rigorous theoretical or empirical justifications for applying these prior models to general images, and there is little theory to guide the construction and selection of prior models. To be concrete, for a given application, we shall ask why the differential operators are good choices in capturing image features, and what are the best forms for $p(\mathbf{I})$ and $\psi()$. Another interesting fact is that real world scenes are observed at arbitrary scales, thus a good prior model should remain the same for image features at multiple scales. However none of the above prior models has scale-invariant property, i.e, they are not renormalizable in terms of renormalization group theory (Wilson 1975).

Motivated by the above questions, this article shall study how to find the best generic prior model for visual computation, and we pose it as a statistical inference or a prior learning problem. For a given application domain, we assume that an image $\mathbf{I}$ is characterized by an underlying probability distribution $f(\mathbf{I})$. Suppose that we have access to a set of images in application, it is reasonable to assume that these observed images are typical samples from $f(\mathbf{I})$, and then the goal of learning prior model $p(\mathbf{I})$ is to make

[1] Where the smoothness term is explained as a stabilizer for solving "ill-posed" problems (Tikhonov 1977).

[2] If $\psi()$ is quadratic, then variational solutions minimizing the potential are splines, such as flexible membrane or thin plate models.

inference about $f(\mathbf{I})$. Our theory is based on a maximum entropy principle, proposed in a preceding paper for texture modeling (Zhu, Wu, and Mumford 1996). First, instead of being limited to a few differential operators, our theory examines whatever filters capture the structures of natural images, such as Gabor filters (Daugman 1985). Second, unlike previous prior models which subjectively assume some parametric forms for the potential function $\psi()$, our theory uses non-parametric form, and learns it from observed images. An information criterion is put forth for choosing the most informative filters (or features) through minimizing a Kullback-Leibler information distance between $p(\mathbf{I})$ and $f(\mathbf{I})$.

The learned prior models are verified in two ways. i) In the spirit of "analysis=synthesis" (Grenander 1976), images are sampled from the learned prior models, so we can see what typical images they stand for. ii) they are compared with existing prior models in experiments of image restoration and segmentation.

In this paper, we choose the domain of application to be general natural images.[3] We are interested in natural images for the following reasons. First, this is the most general set of images. It is obvious that the same strategy can be applied to any special applications, such as MRI images and 3D laser range data, where more specific prior models are expected. Second, in image coding theory and psychology, recently there are increasing interests in understanding the common features and their statistics of natural images (Ruderman and Bielek 1994). The statistics of natural images not only helps us understand why neurons in visual cortex behave as they do (Field 1994, Olshausen and Field 1995), but also sheds light on the choice of efficient coding schemes. A key problem in these work is to seek a probability model $p(\mathbf{I})$ which accounts for the common statistics of natural images. Third, most natural images are approximately scale invariant, and this motivates our interests in studying prior models which have scale invariant properties.

This paper is arranged as follows. Section (2) discusses the objective of learning prior models, and presents two extreme cases to archive such goal. Then in section (3) a general theory is proposed for learning prior models based on a maximum entropy principle. Section (4) presents a criterion for model selection, and section (5) demonstrates some experiments on the statistics of natural images and prior learning. Section (6) compares different prior models by simulated experiments on image restoration and segmentation.

[3]Here, natural images refer to any images of a real world scene, indoor or outdoor, and they shall be distinguished from non-natural ones such as noise-distorted or over-exposed images.

Finally section (7) concludes with a critic discussion.

2 Goal of prior learning and two extreme cases

Let image $\mathbf{I}$ be defined on an $N \times N$ lattice $\mathbf{S}$. For any pixel (x, y), $\mathbf{I}(x, y) \in \mathbf{L}$, and $\mathbf{L}$ is either an interval of $\mathbf{R}$ or $\mathbf{L} \subset \mathbf{Z}$. Thus we define a joint probability distribution $f(\mathbf{I})$ over the image space $\mathbf{L}^{N^2}$. $f(\mathbf{I})$ expresses how likely an image $\mathbf{I}$ is observed in a given application, and in this paper it is defined as the probability of human access to natural scenes. In other words, $f(\mathbf{I})$ should concentrate on a subspace of $\mathbf{L}^{N^2}$, which corresponds to natural images.

Let $NI^{obs} = \{\mathbf{I}_n^{obs}, \quad n = 1, 2.., M\}$ be a set of observed natural images, then we say *the objective of learning a generic prior model is to look for common features and their statistics from the observed natural images. Such features and their statistics are then incorporated into a probability distribution* $p(\mathbf{I})$ *so that* $p(\mathbf{I})$ *biases vision algorithms against image features which are not typical in natural images, such as noise distortions and blurring.*

Without loss of generality, we assume that all features are extracted by filters $F^{(\alpha)}$ with α being an index of filters. The window of $F^{(\alpha)}$ can be of any size from 1×1 to $N \times N$. $F^{(\alpha)}$ can be linear or nonlinear, and it can be a sophisticated function of image I. Given an image $\mathbf{I}$, under a certain boundary condition,[4] $\mathbf{I}^{(\alpha)}(x, y)$ denotes the filter response of $F^{(\alpha)}$ on $\mathbf{I}$ at (x, y). If $F^{(\alpha)}$ is a linear filter, then $\mathbf{I}^{(\alpha)}(x, y) = F^{(\alpha)} * \mathbf{I}(x, y)$

Definition 1 *We define the marginal distribution of* $f(\mathbf{I})$ *with respect to* $F^{(\alpha)}$ *at any* $(x, y) \in \mathbf{S}$ *as,*

$$f^{(\alpha)}(z, x, y) = \int \cdot \int_{\mathbf{I}^{(\alpha)}(x,y)=z} f(\mathbf{I}) d\mathbf{I} = E_f[\delta(z - I^{(\alpha)}(x, y))] \qquad \forall z \in \mathbf{R}$$

where $\delta(t)$ *is a Dirac function,* $\delta(t) = 1$*, if* $t = 0$ *and* $\delta(t) = 0$ *otherwise.*

To see that $f^{(\alpha)}(z, x, y)$ is a marginal distribution of $f(\mathbf{I})$, we show a simple case where the filter is a Dirac function located at (x_0, y_0), $F^{(0)}(x, y) = \delta(x - x_0)\delta(y - y_0)$. Thus the filter response $\mathbf{I}^{(0)}(x_0, y_0) = \mathbf{I}(x_0, y_0)$. We have,

$$f^{(\alpha)}(\mathbf{I}(x_0, y_0), x_0, y_0) = \int \cdot \int f(\mathbf{I}) \prod_{(x,y) \neq (x_0, y_0)} d\mathbf{I}(x, y).$$

[4] In this paper, we assume circulant boundary condition.

For the purpose of learning a generic prior model, it is reasonable to assume that any image features have equal chance to occur at any location, so $f(\mathbf{I})$ is translation invariant, and $f^{(\alpha)}(z, x, y) = f^{(\alpha)}(z)$. We will discuss the limits of this assumption in section (7).

Definition 2 *Given a filter $F^{(\alpha)}$ and a specific image $\mathbf{I}$, we define the histogram of filtered image $\mathbf{I}^{(\alpha)}(x, y)$ as,*

$$H^{(\alpha)}(z) = \frac{1}{\mid \mathbf{S} \mid} \sum_{(x,y)} \delta(z - \mathbf{I}^{(\alpha)}(x, y)).$$

Similarly, given $F^{(\alpha)}$ and an observed image $\mathbf{I}^{obs}$, $H^{obs(\alpha)}(z)$ denotes the histogram of filtered image $\mathbf{I}^{obs(\alpha)}$, and

$$\bar{H}^{obs(\alpha)}(z) = \frac{1}{M} \sum_{\mathbf{I}_n^{obs} \in NI^{obs}} H^{obs(\alpha)}(z),$$

is the histogram averaged over all observed images in NI^{obs}. As a special case $\bar{H}^{obs(0)}(z)$ is the average intensity histogram of observed natural images. We note that $\bar{H}^{obs(\alpha)}(z)$ is an unbiased estimate for $f^{(\alpha)}(z)$, and as $M \to \infty$, $\bar{H}^{obs(\alpha)}(z)$ converges to $f^{(\alpha)}(z)$. This is stated in the following theorem.

Theorem 1 *$E_f[H^{(\alpha)}(z)] = f^{(\alpha)}(z)$, $E_f[\bar{H}^{obs(\alpha)}(z)] = f^{(\alpha)}(z)$, and $Var[\bar{H}^{obs(\alpha)}(z) - f^{(\alpha)}(z)] = \frac{1}{M} Var[H^{obs(\alpha)}(z) - f^{(\alpha)}(z)]$.*

[Proof]. The proof follows directly from the definitions and the ergodicity assumption.

Similarly, for any homogeneous probability distribution $p(\mathbf{I})$, $E_p[H^{(\alpha)}(z)]$ is the marginal distribution of $p(\mathbf{I})$ with respect to $F^{(\alpha)}$.

Now, to learn a prior probability model from a set of observed images $\{\mathbf{I}_n^{obs}, \quad n = 1, 2, ...M\}$, immediately we have two simple solutions. The first is,

$$p(\mathbf{I}) = \prod_{(x,y)} \bar{H}^{obs(0)}(\mathbf{I}(x, y)) \tag{2}$$

Taking $\psi_1(z) = -\log \bar{H}^{obs(0)}(z)$, we rewrite equation (2) as,

$$p(\mathbf{I}) = \frac{1}{Z} e^{-\sum_{(x,y)} \psi_1(\mathbf{I}(x,y))}. \tag{3}$$

The second solution is:

$$p(\mathbf{I}) = \frac{1}{M} \sum_{n=1}^{M} \delta(\mathbf{I} - \mathbf{I}_n^{obs}) \tag{4}$$

Suppose $\|\mathbf{I}_n^{obs}\|^2 = c_n$ for $n = 1, 2, ..., M$, then we write equation (4) as

$$p(\mathbf{I}) = \frac{1}{M} e^{-\sum_{n=1}^{M} \psi_2(<\mathbf{I}_n^{obs}, \mathbf{I}> - c_n)} \tag{5}$$

where $<,>$ is inner product, $\psi_2(0) = 0$, and $\psi_2(x) = \infty$ if $x \neq 0$. Here $\psi_2()$ is similar to Potts model (Winkler 1995).

These two solutions stand for two typical mechanisms for constructing probability models in the literature. The first is often used for image coding (Watson 1987), while the second is a special case of the learning scheme using radial basis functions (RBF) (Poggio and Girosi 1990).[5] Although the philosophies for learning these two prior models are very different, we observe that they share two common properties.

- **Property I**. *Like the prior model in equation (1), the potentials $\psi()$ and thus $p(\mathbf{I})$ are built on the responses of linear filters.* In the first case, the filter is the Dirac function $F^{(0)}$, and the window size of $F^{(0)}$ is 1×1. In the second case, each observed image $\mathbf{I}_n^{obs}$ is used as a linear filter of size $N \times N$, and we denote filter $F^{(obsn)}(x, y) = \mathbf{I}_n^{obs}$, for $n = 1, 2..., M$.

- **Property II**. *For each filter $F^{(\alpha)}$ chosen above, $E_p[H^{(\alpha)}(z)] = \bar{H}^{obs(\alpha)}(z)$, so as M increases, the marginal distribution of $p(\mathbf{I})$ converges to $f^{(\alpha)}$.* In the first case, as the random variables $\mathbf{I}(x, y)$ are identical and independent of each other for different (x, y) according to $p(\mathbf{I})$ in equation (2), it is straightforward to verify that $E_p[H^{(0)}(z)] = \bar{H}^{obs(0)}(z)$. In the second case, we have,

$$\begin{aligned} E_p[H^{(obsn)}] &= \int \cdot \int p(\mathbf{I}) \delta(z - < \mathbf{I}_n^{obs}, \mathbf{I} >) d\mathbf{I} \\ &= \int \cdot \int \frac{1}{M} \sum_{j=1}^{M} \delta(\mathbf{I} - \mathbf{I}_j^{obs}) \delta(z - < \mathbf{I}_n^{obs}, \mathbf{I} >) d\mathbf{I} \\ &= \frac{1}{M} \sum_{j=1}^{M} \delta(z - < \mathbf{I}_n^{obs}, \mathbf{I}_j^{obs} >) \\ &= \bar{H}^{obs(obsn)}. \end{aligned}$$

  Since it embodies the statistics of observed images, this second property is very important for $p(\mathbf{I})$, but unfortunately, it is in general not satisfied by existing prior models discussed in equation (1).

[5] In RBF, the basis functions are presumed to be smooth, such as a Gaussian function. Here, using $\delta()$ is more loyal to the observed data.

$p(\mathbf{I})$ in both cases meet our objective for learning a prior model from observed images. But intuitively, these two models are not good. Since in equation (3), filter $F^{(0)}$ does not capture spatial structures among adjacent pixels. While in equation (5), filters $F^{(obsn)}$ are too specific to be helpful in predicting features in those unobserved images.

In fact, the two kinds of filters used above lies in the two extremes of the spectrum of all linear filters. As discussed by Gabor (Gabor 1946), the δ filter is localized in space but is extended uniformly in frequency. In contrast, some other filters, like the sine waves, are well localized in frequency but are extended in space. Each filter $F^{(obsn)}$ includes a specific combination of all the components in both space and frequency, however such combination of features can hardly be common to general natural images. Thus the prior models will be inefficient when they are used in vision algorithm. A quantitative analysis of the goodness of these filters is given in table 1 at section (5.2).

To generalize these two extreme probability models, in the next section, we shall study a general theory for learning a prior probability model. First, this theory should be general enough to exploit whatever filters which are efficient in capturing the common structures of natural images, such as the Gabor filter which are proven to be optimally localized in both space and frequency (Daugman 1985). Second, to capture the statistics of the features extracted by filters, the prior model $p(\mathbf{I})$ should satisfy property II above.

3 Learning prior models by maximum entropy principle

Suppose that we choose an arbitrary set of filters $\mathbf{B}_k = \{F^{(\alpha)}, \alpha = 1, 2, ..., k\}$ to characterize the structures of natural images, and we denote $\mathbf{M}_k = \{\bar{H}^{obs(\alpha)}, \alpha = 1, 2, ..., k\}$. Then *the objective of prior learning is to compute a distribution* $p(\mathbf{I})$*, such that* $E_p[H^{(\alpha)}] = \bar{H}^{obs(\alpha)}$*, for* $\alpha = 1, 2, ..., k$. In other words, we let $p(\mathbf{I})$ and $f(\mathbf{I})$ have the same marginal distributions with respect to each chosen filter. The more filters we use, the more structures $p(\mathbf{I})$ bears, and the closer $p(\mathbf{I})$ is to $f(\mathbf{I})$.

In a preceding paper (Zhu, Wu, and Mumford 1996), it has been proven that **as the number of observed images** $M \to \infty$ **and as the filter number** $k \to \infty$**, with only linear filters used,** $p(\mathbf{I})$ **shall converge to the underlying distribution** $f(\mathbf{I})$**.** But for computational reason, it is often desirable to choose a small set of filters which most efficiently capture the image structures. We shall discuss how to choose filters in later

section.

Now, given $\mathbf{B}_k$ and $\mathbf{M}_k$, we denote,

$$\Omega_k = \{p(\mathbf{I}) \mid E_p[\delta(\mathbf{I}^{(\alpha)}(x,y) - z)] = \bar{H}^{obs(\alpha)}(z) \quad \forall z,\ \forall (x,y)\ \forall \alpha\}, \tag{6}$$

thus for our purpose, any $p(\mathbf{I}) \in \Omega_k$ will be a good prior model. From Ω_k we choose $p(\mathbf{I})$ that maximizes the entropy, i.e.,

$$p(\mathbf{I}) = \arg\max\{-\int p(\mathbf{I}) \log p(\mathbf{I}) d\mathbf{I}\}, \tag{7}$$

subject to $\quad E_p[H^{(\alpha)}] = E_p[\delta(\mathbf{I}^{(\alpha)}(x,y) - z)] = \bar{H}^{obs(\alpha)}(z) \qquad \forall z,\ \forall (x,y)\ \forall \alpha$

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

Since entropy is a measure of disorder and it is the negative Kullback-Leibler distance, up to a constant, from the uniform distribution (see Cover and Thomas 1985), the underlying philosophy of entropy maximization is the followings. While $p(\mathbf{I})$ satisfies the constraints along some dimensions, it is made as random as possible in other unconstrained dimensions. Therefore it minimizes the introduction of artificial information. The maximum entropy distribution $p(\mathbf{I})$ gives the least bias explanation for the constraints and thus the purest fusion of the extracted features and their statistics.

By Lagrange multiplier, solving the above optimization problem gives the following probability distribution,

$$p(\mathbf{I}) = \frac{1}{Z} e^{-\sum_{\alpha=1}^{k} \sum_{(x,y)} \{\int \lambda^{(\alpha)}(z) \delta(\mathbf{I}^{\alpha}(x,y) - z) dz\}} \tag{8}$$

$$= \frac{1}{Z} e^{-\sum_{\alpha=1}^{k} \sum_{(x,y)} \lambda^{(\alpha)}(\mathbf{I}^{(\alpha)}(x,y))}. \tag{9}$$

where $Z = \int \cdot \int e^{-\sum_{\alpha=1}^{k} \sum_{(x,y)} \lambda^{(\alpha)}(\mathbf{I}^{(\alpha)}(x,y))} d\mathbf{I}$ is a partition function. Note that in the above constrained optimization problem, z takes real values, i.e., there are infinite number of constraints. Therefore the Lagrange parameter λ takes the form as a continuous function of z in equation (8). In equation (9) $p(\mathbf{I})$ is a Gibbs distribution, and the potential is summation over functions of the filter responses $\mathbf{I}^{(\alpha)}(x,y)$. For each filter $F^{(\alpha)}$, we use the notation $\lambda^{(\alpha)}()$ to replace $\psi()$ in order to emphasize that $\lambda^{(\alpha)}()$ will be learned from data. In general, unlike the extreme case in equation (3), $\lambda^{(\alpha)}(z)$ is not equal to $\log \bar{H}^{obs(\alpha)}(z)$, and analytic solution for $\lambda^{(\alpha)}(z)$ is unavailable in general.

To proceed further, we derive a discrete form of equation (9). Let the filter responses $\mathbf{I}^{(\alpha)}(x,y)$ be quantitized into L discrete grey levels, so z takes values from set

$\{z_1^{(\alpha)}, z_1^{(\alpha)}, ..., z_L^{(\alpha)}\}$. As a result, $H^{(\alpha)}(z)$ is approximated by a piecewise-constant histogram of L bins. We denote it by a vector $H^{(\alpha)} = (H_1^{(\alpha)}, H_2^{(\alpha)}, ..., H_L^{(\alpha)})$ with $H_i^{(\alpha)} = \frac{1}{|\mathbf{S}|}\sum_{(x,y)} \delta(\mathbf{I}^{(\alpha)}(x,y) - z_i^{(\alpha)})$, $i = 1, ..., L$. In general, the width of these bins do not have to be equal, and the number of grey levels L for each filter response may vary. Since there are a finite number of constraints in the optimization problem, we write equation (8) as,

$$p(\mathbf{I}) = \frac{1}{Z} e^{-\sum_{\alpha=1}^{k}\sum_{(x,y)}\sum_{i=1}^{L}\{\lambda_i^{(\alpha)}\delta(\mathbf{I}^{(\alpha)}(x,y) - z_i^{(\alpha)})\}}.$$

Changing the order of summations, we have,

$$p(\mathbf{I}) = \frac{1}{Z} e^{-\sum_{\alpha=1}^{k}\sum_{i=1}^{L}\lambda_i^{(\alpha)} H_i^{(\alpha)}}, \tag{10}$$

where the potential function $\lambda^{(\alpha)}(z)$ is approximated by a piecewise-constant function, and we denote it by vector $\lambda^{(\alpha)} = (\lambda_1^{(\alpha)}, \lambda_2^{(\alpha)}, ..., \lambda_L^{(\alpha)})$. So equation (10) becomes

$$p(\mathbf{I}) = \frac{1}{Z} e^{-\sum_{\alpha=1}^{k} < \lambda^{(\alpha)}, H^{(\alpha)} >}. \tag{11}$$

The probability distribution $p(\mathbf{I})$ in the above equation has the following properties:

- **Property I**. $p(\mathbf{I})$ is specified by $\lambda = (\lambda^{(1)}, \lambda^{(2)}, ..., \lambda^{(k)})$, thus $p(\mathbf{I}) = p(\mathbf{I}; \lambda)$.
- **Property II**. Given an image $\mathbf{I}$, its histograms $H^{(1)}, H^{(2)}, ..., H^{(k)}$ are sufficient statistics, i.e. $p(\mathbf{I})$ is a function of $(H^{(1)}, H^{(2)}, ..., H^{(k)})$.

The partition function Z depends on λ, i.e., $Z = Z(\lambda^{(1)}, \lambda^{(2)}, ..., \lambda^{(K)})$, and it has the following nice properties $\forall \alpha \quad \beta$,

1) $$\frac{\partial \log Z}{\partial \lambda^{(\alpha)}} = \frac{1}{Z}\frac{\partial Z}{\partial \lambda^{(\alpha)}} = -E_p[H^{(\alpha)}],$$

2) $$\frac{\partial^2 \log Z}{\partial \lambda^{(\alpha)} \partial \lambda^{(\beta)}} = E_p[(H^{(\alpha)} - E_p[H^{(\alpha)}])(H^{(\beta)} - E_p[H^{(\beta)}]) = Cov_p[H^{(\alpha)}, H^{(\beta)}].$$

The second property tells us that the Hessian matrix of $\log Z$ is the covariance matrix of vector $(H^{(1)}, ..., H^{(k)})$ that is straight concave, so is $\log p(\mathbf{I})$. Therefore given a set of consistent constraints, the solution for $(\lambda^{(1)}, ..., \lambda^{(k)})$ is unique.

By gradient descent, solving the maximum entropy problem gives the following equations for computing $\lambda^{(\alpha)}$, $\alpha = 1, .., k$ iteratively,

$$\frac{d\lambda^{(\alpha)}}{dt} = E_{p(\mathbf{I};\lambda)}[H^{(\alpha)}] - \bar{H}^{obs(\alpha)} \quad \forall \alpha. \tag{12}$$

When equation (12) converges, the constraints in the optimization problem are satisfied.

It is worth mentioning that the above maximum entropy (ME) estimator is equivalent to the maximum likelihood estimator (ML),

$$\lambda = \arg\max_{\lambda} \prod_{n=1}^{M} p(\mathbf{I}_n^{obs}; \lambda) = \arg\max_{\lambda} \frac{1}{Z} e^{-\sum_{\alpha=1}^{k} <\lambda^{(\alpha)}, \bar{H}^{obs(\alpha)}>}.$$

This conclusion follows directly from property 1 of the partition function.

In equation (12), at each step t, given λ and hence $p(\mathbf{I};\lambda)$, the analytic form of $E_{p(\mathbf{I};\lambda)}(H^{(\alpha)})$ is not available. A numeric solution for $E_{p(\mathbf{I};\lambda)}(H^{(\alpha)})$ is to sample $p(\mathbf{I};\lambda)$, therefore synthesize an image $\mathbf{I}^{syn}$. Let $\mathbf{I}^{syn(\alpha)}$ be the filtered image, then we use $H^{syn(\alpha)}$, which is the histogram of $\mathbf{I}^{syn(\alpha)}$, to estimate $E_{p(\mathbf{I};\lambda)}(H^{(\alpha)})$.[6]

We use Gibbs sampler to sample from $p(\mathbf{I},\lambda)$ (Geman and Geman 1984). It simulates an inhomogeneous Markov chain in image space $\mathbf{L}^{|\mathbf{S}|}$. It starts from $\lambda^{(\alpha)} = 0$, $p(\mathbf{I};\lambda)$ an uniform distribution, and $\mathbf{I}^{syn}$ an uniform noise image. The Gibbs sampler randomly picks up a pixel (x,y), and flips $\mathbf{I}^{syn}(x,y)$ according to distribution $p(\mathbf{I};\lambda)$. Due to the Markovanity of $p(\mathbf{I};\lambda)$, the computation of the probability at each flipping step is only conditional on the neighborhood of (x,y), and the size of the neighborhood is equal to the biggest window size of all filters in $\mathbf{B}_k$. $N \times N$ times of flipping is called a *sweep*. At each fixed λ, we run the Gibbs sampler for 10 sweeps, and update λ by equation (12). Then the Gibbs sampler continues from image $\mathbf{I}^{syn}$ synthesized in the last step, and so on. The algorithm stops when $H^{syn(\alpha)}$ closely match $\bar{H}^{obs(\alpha)}$, for $\alpha = 1, 2, ..., k$, and the Markov chain becomes stationary.

Such inhomogeneous Markov chain is guaranteed to converge to the stationary process with distribution $p(\mathbf{I};\lambda)$ and λ the optimal solution, provided that we run the sampling process long enough (Younes 1988). But there is no good criterion to test whether the Markov chain has converged to its stationary process. Recently a criterion is proposed for testing convergence using coupled Markov chains (Propp and Wilson 1995), but it requires the probability model to be monotonic, and which is often not observed in these prior models except that $\lambda()$ is of a quadratic form. We experiment with a prior model of quadratic $\lambda()$, and the synthesized image is of the same size. The coupled Markov chains converge at 2350 sweeps according to the Propp and Wilson criterion. Therefore we assume that magnitude for sampling sweeps in our experiments. On the other aspect,

[6] To make the estimation accurate, the size of $\mathbf{I}^{syn}$ have to be big enough.

since histograms $H^{syn(\alpha)}$, $\alpha = 1, 2, ..., k$ are sufficient statistics of $\mathbf{I}^{syn}$ according to $p(\mathbf{I})$, by observing the change of $H^{syn(\alpha)}$ with time steps, we can monitor the convergence of the sampling process. Another point is important in the learning process. Since the two tails of the histogram usually represent statistical information for meaningful structures, however, their values are relatively very low. Thus during the learning process, we not only demand $\|\bar{H}^{obs(\alpha)} - H^{syn(\alpha)}\| \leq \epsilon$ but also requires $\|\log \bar{H}^{obs(\alpha)} - \log H^{syn(\alpha)}\| \leq \rho$ for $\alpha = 1, 2, ..., k$.

4 Information criterion for model selection

Last section presents a general theory for learning a probability model based on an arbitrary set of filters, this section proposes an information criterion for selecting the most important filters which capture common features of natural images.

Suppose we have a general filter bank $\mathbf{B}$, given $\mathbf{B}_k \subset \mathbf{B}$ and $\mathbf{M}_k$, a maximum entropy probability model $p_k(\mathbf{I})$ is learned as in last section. Index k stands for the number of filters used in the prior model and it also tells model complexity. The goodness of $p_k(\mathbf{I})$ depends on the filters in $\mathbf{B}_k$, and is measured by the Kullback-Leibler information distance between $f(\mathbf{I})$ and $p_k(\mathbf{I})$,

$$I(f, p_k) = \int \cdot \int f(\mathbf{I}) \log \frac{f(\mathbf{I})}{p_k(\mathbf{I})} d\mathbf{I} = E_f[\log f(\mathbf{I})] - E_f[\log p_k(\mathbf{I})].$$

We notice that the statistics of natural images, i.e., $H^{obs(\alpha)}$ $\alpha = 1, ..., k$, vary from image to image. Here with slight abuse of notation, $H^{obs(\alpha)}$ depends on $\mathbf{I}_n^{obs(\alpha)}$. For each specific image $\mathbf{I}_n^{obs}$ or a group of images in a specific domain, it is often desirable to estimate a specific underlying distribution $f_n(\mathbf{I})$. The true marginal distribution of $f_n(\mathbf{I})$,i.e., $E_{f_n}[H^{(\alpha)}]$, is more accurately estimated by $H^{obs(\alpha)}$ than by $\bar{H}^{obs(\alpha)}$. Let $E_{p_k^*}[H^{(\alpha)}] = H^{obs(\alpha)}$ for $\alpha = 1, 2, ..., k$, we obtain a ME prior model $p_k^*(\mathbf{I})$ for this specific domain. Indeed, the more specific model $p_k^*(\mathbf{I})$ is, the more useful it will be in a vision algorithm when it is applied to the right set of images. Thus the goodness of $p_k(\mathbf{I})$ with respect to $\mathbf{I}_n^{obs}$ is measured by $I(f_n, p_k)$.

To compute $I(f_n, p_k)$, we observe the following theorem.

Theorem 2 *Given k filters, let $p_k^*(\mathbf{I})$ and $p_k(\mathbf{I})$ be two maximum entropy distributions with k filters, Let $p_k^*(\mathbf{I})$ and $f_n(\mathbf{I})$ have the same marginal distributions with respect to the k filters. Then $I(f_n, p_k) = I(f_n, p_k^*) + I(p_k^*, p_k)$*

[proof] Since $E_{f_n}[H^{(\alpha)}] = E_{p_k^*}[H^{(\alpha)}]$, $\alpha = 1, 2, ..., k$, thus

$$\begin{aligned} E_{f_n}[\log p_k^*(\mathbf{I})] &= E_{f_n}[-\log Z - \sum_{\alpha=1}^{k} < \lambda^{(\alpha)}, H^{(\alpha)} >] \\ &= -\log Z - \sum_{\alpha=1}^{k} < \lambda^{(\alpha)}, E_{f_n}[H^{(\alpha)}] > \\ &= -\log Z - \sum_{\alpha=1}^{k} < \lambda^{(\alpha)}, E_{p_k^*}[H^{(\alpha)}] > \\ &= E_{p_k^*}[\log p_k^*(\mathbf{I})] \end{aligned}$$

Similarly, we have $E_{f_n}[\log p_k(\mathbf{I})] = E_{p_k^*}[\log p_k(\mathbf{I})]$. Therefore,

$$\begin{aligned} I(f_n, p_k) &= E_{f_n}[\log f_n(\mathbf{I})] - E_{f_n}[\log p_k^*(\mathbf{I})] + E_{f_n}[\log p_k^*(\mathbf{I})] - E_{f_n}[\log p_k(\mathbf{I})] \\ &= I(f_n, p_k^*) + I(p_k^*, p_k) \end{aligned}$$ □

Fixing model complexity k, we need to choose the most efficient k filters so that $I(f_n, p_k)$ is minimized. We adopt a stepwise greedy algorithm for minimizing $I(f_n, p_k)$. It starts from $i = 0$ and $p_0(\mathbf{I})$ an uniform distribution, then it sequentially introduces one filter at a time. At each step i $(0 \leq i \leq k)$, we have chosen $\mathbf{B}_{i-1}$ and obtained probability distribution $p_{i-1}(\mathbf{I})$, then the i-th filter $F^{(i)}$ is chosen to minimize $I(f_n, p_i) - I(f_n, p_{i-1})$, i.e.,

$$F^{(i)} = \arg \min_{F^{(i)} \in \mathbf{B} - \mathbf{B}_{i-1}} \{I(f_n, p_i) - I(f_n, p_{i-1})\}$$

By theorem 2, we obtain,

$$I(f_n, p_i) - I(f_n, p_{i-1}) = -I(p_i^*, p_{i-1}^*) + I(p_i^*, p_i) - I(p_{i-1}^*, p_{i-1}). \tag{13}$$

To compute $I(f_n, p_i) - I(f_n, p_{i-1})$, we introduce the following theorem.

Theorem 3 *Let $p_0(\mathbf{I})$ and $p(\mathbf{I})$ be two maximum entropy distributions, $E_{p_0}[H^{(\alpha)}] = h_0^{(\alpha)}$ and $E_p[H^{(\alpha)}] = h^{(\alpha)}$ for $\alpha = 1, 2, ..., k$. Denote $h_0 = (h_0^{(1)}, h_0^{(2)}, ..., h_0^{(k)})$ and $h = (h^{(1)}, h^{(2)}, ..., h^{(k)})$. Fixing h_0, the Kullback-Leibler distance $I(p_0, p)$ is a function of h, i.e., $I(p_0, p) = I(h)$, then $I(h) = (h - h_0)Var^{-1}(h_0)(h - h_0)^T + O(\|h - h_0\|^3)$, where $Var(h_0)$ is a variance matrix.*

[Proof] See appendix for proof.

The virtue of theorem 3 is to measure the distance between $p_0(\mathbf{I})$ and $p(\mathbf{I})$ in terms of the distances of their marginal distributions. However, the variance matrix is computationally intractable, instead, we use the following L_1 norm distance to approximate $I(p_0, p)$,

$$I(p_0, p) \simeq \frac{1}{2} \sum_{\alpha=1}^{k} \|h^{(\alpha)} - h_0^{(\alpha)}\|$$

In summary, if we use the L_1 norm distance above, then together with equation (13), we approximate $I(f_n, p_i) - I(f_n, p_{i-1})$ by,

$$\Delta_n^{(i)} = \frac{1}{2}\|H^{obs(i)} - \bar{H}^{obs(i)}\| - \frac{1}{2}\|H^{obs(i)} - E_{p^*_{i-1}}[H^{(i)}]\| \tag{14}$$

The first term is an increase of distance introduced by the error of replacing $H^{obs(i)}$ with $\bar{H}^{obs(i)}$ when computing $p_k(\mathbf{I})$, while the second term reduces the distance because $p_k^*(\mathbf{I})$ improves $p_{k-1}^*(\mathbf{I})$ by incorporating the information extracted by the i-th filter.

$\Delta_n^{(i)}$ depends on $\mathbf{I}_n^{obs}$, therefore the overall effect of choosing a new filter should be averaged over all observed natural images,

$$\bar{\Delta}^{(i)} = \frac{1}{2M} \sum_{\mathbf{I}_n^{obs} \in NI^{obs}} (\|H^{obs(i)} - \bar{H}^{obs(i)}\| - \|H^{obs(i)} - E_{p^*_{i-1}}[H^{(i)}]\|).$$

The smaller $\bar{\Delta}^{(i)}$ is, the more common information filter $F^{(i)}$ captures, thus better it is. In practice, to estimate $E_{p^*_{i-1}}[H^{(i)}]$, we need to compute $p_{i-1}(\mathbf{I})$ and synthesize an image for each $\mathbf{I}_n^{obs}$, which is computationally expansive. Instead we approximate $E_{p^*_{i-1}}[H^{(i)}]$ by $E_{p_{i-1}}[H^{(i)}]$.

Definition 3 *An* information criterion *(IC) for each filter $F^{(\alpha)}$ at step $1 \leq i \leq k$ is defined by,*

$$IC = \frac{1}{2M} \sum_{\mathbf{I}_n^{obs} \in NI^{obs}} \|H^{obs(\alpha)} - E_{p_{i-1}}[H^{(\alpha)}]\| - \frac{1}{2M} \sum_{\mathbf{I}_n^{obs} \in NI^{obs}} \|H^{obs(\alpha)} - \bar{H}^{obs(\alpha)}\|$$

we call the first term average information gain *(AIG) by choosing $F^{(\alpha)}$ as the i-th filters, and the second term* average information fluctuation *(AIF).*

Intuitively, AIG measures the average error between a marginal distribution of $f_n(\mathbf{I})$ and a marginal distribution of $p_{i-1}(\mathbf{I})$ obtained in the previous step. For a filter $F^{(\alpha)}$, the bigger AIG is, the more information $F^{(\alpha)}$ captures. AIF is a measure of disagreement between observed images. The bigger AIF is, the less common $F^{(\alpha)}$ is shared by all images.

5 Experiments

5.1 Statistics of natural images

We collect a set of 44 natural images. These images are from various sources, some are digitized by a laser scanner from personal album, postcards, and some are from a Corel image database. These images include both indoor and outdoor pictures, country and urban scenes, and all images are normalized to have intensity between 0 and 31. Six images are shown in Figure (2).

It is well known that natural images are very different from random noise ones (Ruderman and Bialek 1994, Field 1994). And we know that the marginal distributions of $\mathbf{I}^{(\alpha)}$ with linear filters $F^{(\alpha)}$ alone are capable of characterizing $f(\mathbf{I})$ – the complete statistics of natural images (Zhu, Wu and Mumford 1996). In this paper, we shall only study the histograms of linearly filtered image.

First, for some features, the statistics of natural images vary largely from image to image. As an example, we choose filter $F^{(0)}$, like the first extreme case in section (2). The average intensity histogram of the 44 images $\bar{H}^{obs(0)}(z)$ is plotted in figure (3.a), while figure (3.b) is the intensity histogram of an individual image (the temple image in figure (2)). In contrast, figure (3.c) is the intensity histogram of an uniform noise image. It appears that $\bar{H}^{obs(0)}(z)$ is close to an uniform distribution, while the difference between figure (3.a) and figure (3.b) is very big. Thus for filter $F^{(0)}$, IC is small (see table 1).

Second, if we choose other filters, the histograms of filter responses are amazingly consistent across all 44 natural images, and they are very different from the histograms of noise images. For example, we study filter ∇_x in figure (4). Figure (4.a) is the average histogram of 44 filtered natural images, figure (4.b) is the histogram of an individual filtered image (the same image as in figure (3.b)), and figure (4.c) is the histogram of a filtered uniform noise image.

The average histogram in figure (4.a) is very different from a Gaussian distribution. To see this, figure (5.a) plots it against a Gaussian curve (dashed one) of the same mean and same variance. The histogram of natural images has higher kurtosis and heavier tails. Similar results are reported in (Field 1994). To see the difference of the tails, figure (5.b) plots the logarithm of the two curves. Figure (5) suggests that the potential functions $\lambda()$ in a prior model should be different from quadratic, however the latter is widely adopted

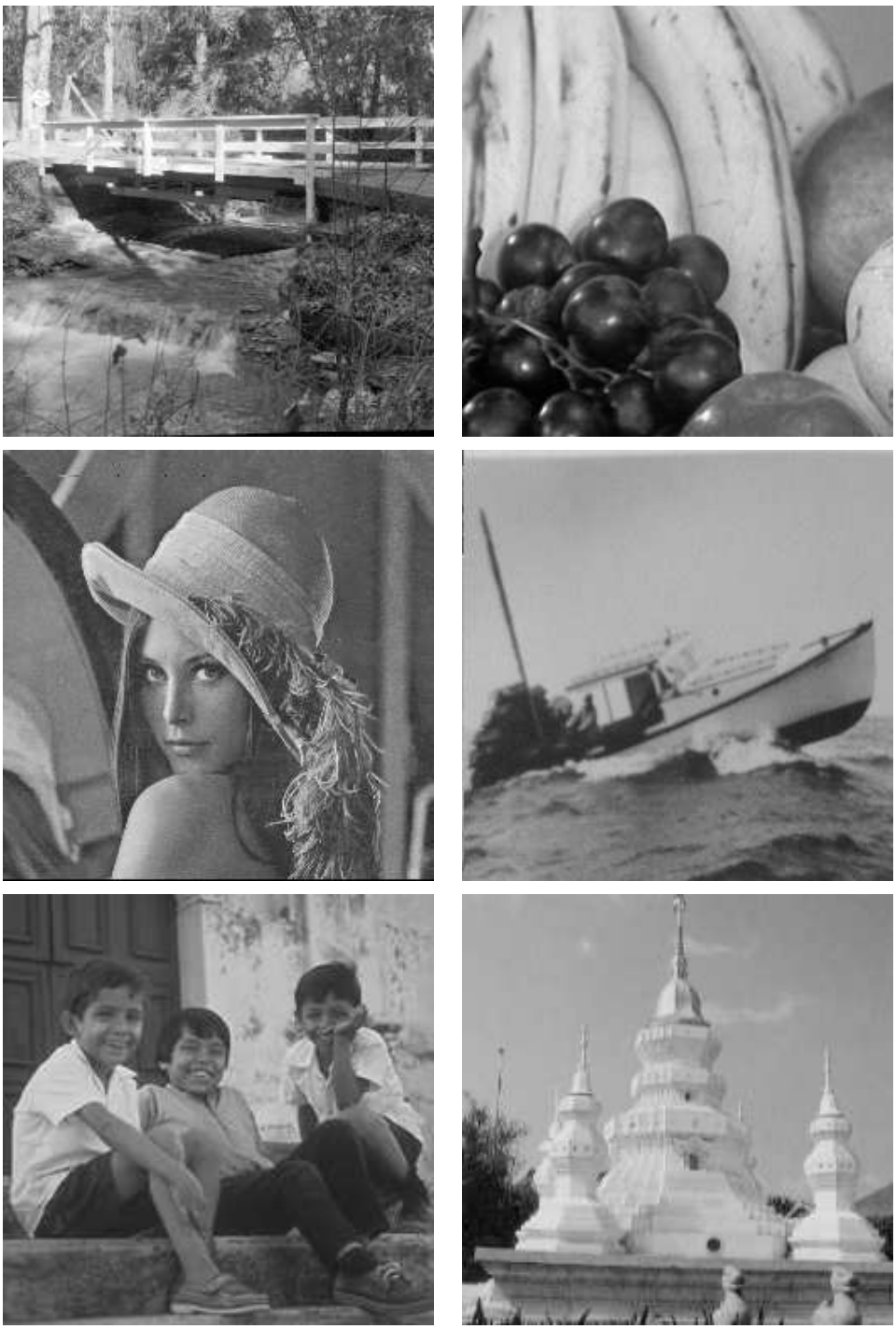
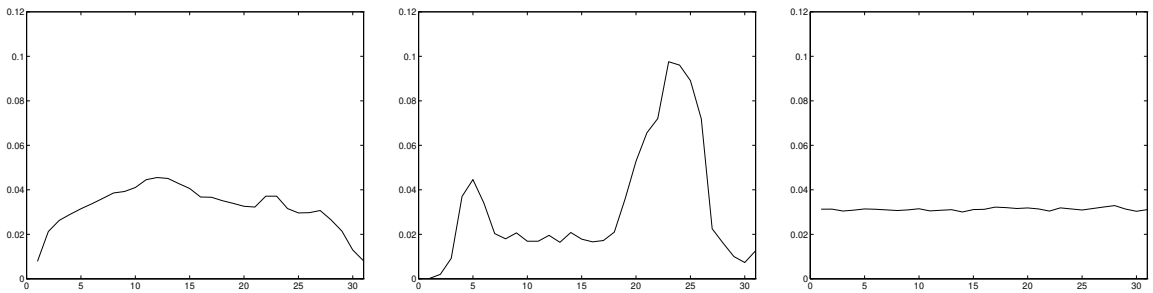
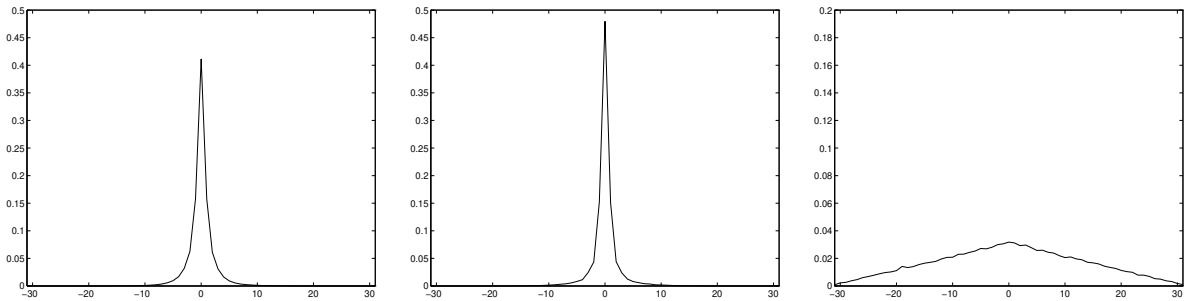


Figure 2: 6 out of the 44 collected natural images.



a b c

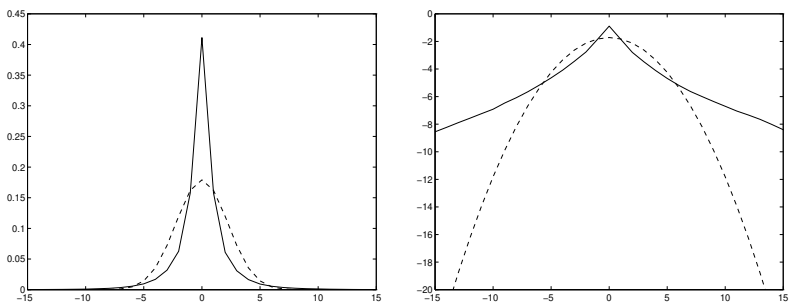
Figure 3: The intensity histograms in domain $[0, 31]$, a, averaged over 44 natural images, b, an individual natural image, c, an uniform noise image.



a b c

Figure 4: The histograms of $\nabla_x \mathbf{I}$ plotted in domain [-15, 15], a. averaged over 44 natural images, b, an individual natural image, c, an uniform noise image.

in computer vision.



a b

Figure 5: a. The histogram of $\nabla_x \mathbf{I}$ plotted against Gaussian curve (dashed) of same mean and variance in domain $[-15, 15]$. b, The logarithm of the two curves in a.

Third, the statistics of natural images are scale invariant with respect to some features. As an example, we study filters ∇_x and ∇_y again. For each image $\mathbf{I}_n^{obs} \in NI^{obs}$, we build a pyramid PD_n with $\mathbf{I}_n^{(s)}$ being the image at s-th layer. We set $\mathbf{I}_n^{(0)} = \mathbf{I}_n^{obs}$, and let

$$\mathbf{I}_n^{(s+1)}(x,y) = \mathbf{I}_n^{(s)}(2x,2y) + \mathbf{I}_n^{(s)}(2x,2y+1) + \mathbf{I}_n^{(s)}(2x+1,2y) + \mathbf{I}_n^{(s)}(2x+1,2y+1).$$

The size of $\mathbf{I}_n^{(s)}$ is $N/2^s \times N/2^s$.

For filter ∇_x, let $\bar{H}_x^{(s)}$ be the average histogram of $\nabla_x \mathbf{I}_n^{(s)}$, over $n = 1, 2, ..., 44$. Figure (6.a) plots $\bar{H}_x^{(s)}$, for $s = 0, 1, 2$, and they are almost identical. To see the tails clearer, we display $\log \bar{H}_x^{(s)}$, $s = 0, 1, 2$ in figure (6).c. The difference between them are still small. Similar results are observed for $\bar{H}_y^{(s)}$ $s = 0, 1, 2$, the latter are average histograms of $\nabla_y \mathbf{I}_n^{obs}$. In contrast, figure (6.b) plots the histograms of $\nabla_x \mathbf{I}^{(s)}$ with $\mathbf{I}^{(s)}$ being an uniform noise image at scales $s = 0, 1, 2$.

Combining the second and third aspects above, we conclude that the histograms of $\nabla_x \mathbf{I}_n^{(s)}, \nabla_y \mathbf{I}_n^{(s)}$ are very consistent across all observed natural images for $n = 1, 2, ..., 44$ and across scales $s = 0, 1, 2$. Moreover, let

$$\bar{H} = \frac{1}{6}(\bar{H}_x^{(0)} + \bar{H}_x^{(1)} + \bar{H}_x^{(2)} + \bar{H}_y^{(0)} + \bar{H}_y^{(1)} + \bar{H}_y^{(2)}), \tag{15}$$

we notice that $\log \bar{H}$ can be well fit to a function of the following form.

$$\Psi(z) = \frac{a}{(1 + \frac{|z|}{c})^{\beta}}$$

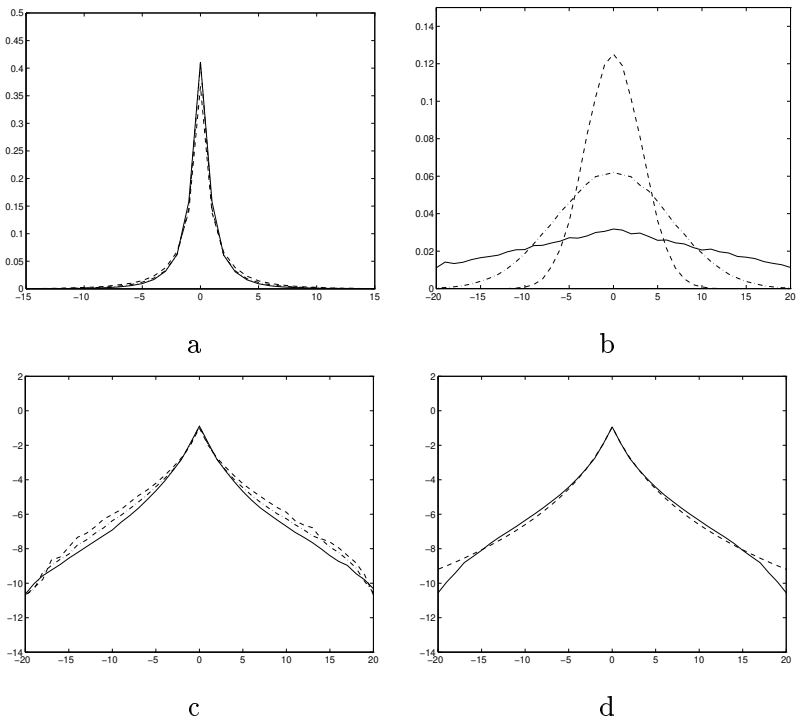


Figure 6: a. $\bar{H}_x^{(s)}$ $s = 0, 1, 2$. b. histograms of a filtered uniform noise image at scales: $s = 0$ (solid curve), $s = 1$ (dash-dotted curve), and $s = 2$ (dashed curve). c. $\log \bar{H}_x^{(s)}$, $s = 0$ (solid), $s = 1$ (dash-dotted) and $s = 2$ (dashed). d. $\log \bar{H}$ (solid), and $\Psi(\nabla \mathbf{I})$ (dashed).

where a, c, β are constants. In figure (6.d), the solid curve is $\log \bar{H}$, and the dashed curve is function $\Psi(z)$ with $a = 4.75, \beta = 5.0$.

In the experiments of next subsection, we will give quantitative answers to the efficiency of filters and the exact forms of the potential functions $\lambda()$. The scale invariant property of natural images is largely caused by the following aspects. 1). Natural images contains objects of various sizes. 2). A natural scene is often taken into image at arbitrary distances. In next subsection we shall study a probability model which has scale-invariant property.

5.2 Simulations

We study the following linear filters in the general filter bank $\mathbf{B}$, and compare them in terms of the information criterion derived in previous section.

1. An intensity filter $\delta()$ or $F^{(0)}$.
2. Isotropic center-surround filters, i.e., the Laplacian of Gaussian filters.

$$LG(x, y, s) = const \cdot (x^2 + y^2 - s^2) e^{-\frac{x^2+y^2}{s^2}}, \tag{16}$$

where $s = \sqrt{2}\sigma$ stands for the scale of the filter. We denote these filters by $LG(s)$. A special filter is $LG(\frac{\sqrt{2}}{2})$, which has a 3×3 window $[0, \frac{1}{4}, 0; \frac{1}{4}, -1, \frac{1}{4}; 0, \frac{1}{4}, 0]$, and we denote it by Δ.

3. Gabor filters with both sine and cosine components, which are models for the frequency and orientation sensitive simple cells.

$$G(x, y \mid s, \theta) = const \circ Rot(\theta) \circ e^{\frac{1}{2s^2}(4x^2+y^2)} e^{-i\frac{2\pi}{s}x}. \tag{17}$$

It is sine wave at frequency $\frac{2\pi}{s}$ modulated by an elongated Gaussian function, and rotated at θ angle. We denote the real and image parts of $G(x, y \mid s, \theta)$ by $Gcos(s, \theta)$ and $Gsin(s, \theta)$. These filters are not even approximately orthogonal to each other. Any other filters can be added to the filter bank, and it is beyond this paper to discuss the definition of an optimal filter bank.

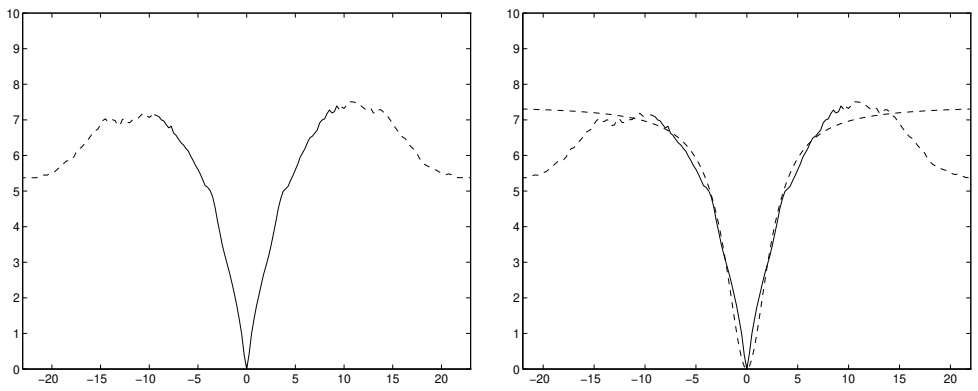
Experiment I.

We compute the AIF (average information fluctuation) with respect to all filters in our filter bank. Due to space limit, we only list the results for a small number of filters in the third column of table 1. We start from filter number $i = 0$, $p_0(\mathbf{I})$ an uniform

distribution, and an uniform noise image $\mathbf{I}^{syn0}$ is synthesized. For each filter $F^{(\alpha)}$, the marginal distribution $E_{p_0}[H^{(\alpha)}(z)]$ is estimated by the histogram of $F^{(\alpha)} * \mathbf{I}^{syn0}$. Then the average information gain (AIG) and the information criterion (IC) are list in table 1 under column $p_0(\mathbf{I})$. According table 1, filter Δ has the biggest IC ($= 0.642$), therefore it is the first filter to choose. We also compare the two extreme cases discussed in section (2). For the $\delta()$ filter, AIF is very big, and AIG is only slightly bigger than AIF. For filter $I^{obs(i)}$, $AIF = \frac{M-1}{M}$ i.e. the biggest among all filters, and $AIG \to 1$. In both cases, ICs are the two smallest.

Choosing Δ as $F^{(1)}$, a maximum entropy distribution $p_1(\mathbf{I})$ is learned,

$$p_1(\mathbf{I}) = \frac{1}{Z} e^{-\sum_{(x,y)} \lambda^{(1)}(\Delta \mathbf{I}(x,y))}.$$



a b

Figure 7: a. The potential function $\lambda^{(1)}()$ plotted in domain $[-25, 25]$. b. Fitting $\lambda^{(1)}()$ with function $\psi(x) = a(1 - 1/(1 + (\mid x \mid /c)^2))$, with $a = 7.4, c = 10$

The potential function $\lambda^{(1)}(z)$ is discretized as a vector of 101 bins for $z \in [-25, 25]$, and it is plotted in figure (7) by linear interpolation. It has one degree of freedom, i.e., we can add an arbitrary constant c to $\lambda^{(1)}()$ without changing the probability, for c is absorbed by Z. In figure (7), we notice that $\lambda^{(1)}(\Delta \mathbf{I})$ has the smallest values around 0, thus encourages smooth image structures. But it decreases for $\mid \Delta \mathbf{I} \mid \geq 9.5$, see the two tails displayed by dashed lines. This is caused by the fact that $\bar{H}^{obs(1)}(\Delta \mathbf{I}) = 0$ [7] if

[7] In fact, we treat $\bar{H}^{obs(1)}(\Delta \mathbf{I}) = 0$ if $\bar{H}^{obs(1)}(\Delta \mathbf{I}) < \frac{1}{N^2}$, with $N \times N$ being the size of synthesized image.

Filter	Filter Size	AIF	$p_0(\boldsymbol{\mathcal{I}})$		$p_1(\boldsymbol{\mathcal{I}})$		$p_3(\boldsymbol{\mathcal{I}})$		$p_s(\boldsymbol{\mathcal{I}})$	
			AIG	IC	AIG	IC	AIG	IC	AIG	IC
δ	1x1	0.278	0.317	0.039	0.317	0.039	0.317	0.039	0.317	0.039
Δ	3x3	0.157	0.799	0.642	0.158	0.001	0.161	0.004	0.198	0.041
$LG(1)$	5x5	0.142	0.727	0.586	0.189	0.047	0.156	0.014	0.155	0.014
$LG(2)$	9x9	0.131	0.418	0.288	0.283	0.152	0.214	0.084	0.148	0.017
$LG(4)$	13x13	0.125	0.267	0.142	0.322	0.197	0.246	0.120	0.155	0.030
∇_x	1x2	0.147	0.716	0.568	0.247	0.100	0.148	0.001	0.155	0.008
∇_y	2x1	0.133	0.732	0.600	0.254	0.121	0.134	0.001	0.148	0.015
$Gcos(2, 0^o)$	5x5	0.119	0.716	0.597	0.239	0.119	0.181	0.062	0.197	0.078
$Gcos(2, 90^o)$	5x5	0.155	0.673	0.518	0.238	0.083	0.188	0.034	0.189	0.035
$Gsin(2, 0^o)$	5x5	0.125	0.573	0.447	0.344	0.219	0.170	0.045	0.141	0.015
$Gsin(2, 90^o)$	5x5	0.126	0.666	0.540	0.241	0.115	0.156	0.030	0.154	0.028
$Gcos(4, 0^o)$	7x7	0.133	0.569	0.436	0.321	0.187	0.203	0.070	0.164	0.031
$Gcos(4, 90^o)$	7x7	0.144	0.545	0.401	0.304	0.160	0.214	0.070	0.183	0.039
$Gsin(4, 0^o)$	7x7	0.124	0.555	0.431	0.334	0.209	0.186	0.063	0.149	0.025
$Gsin(4, 90^o)$	7x7	0.131	0.535	0.405	0.322	0.191	0.196	0.065	0.163	0.033
$Gcos(6, 0^o)$	11x11	0.125	0.384	0.259	0.340	0.216	0.220	0.095	0.145	0.020
$Gcos(6, 90^o)$	11x11	0.129	0.398	0.269	0.340	0.211	0.229	0.100	0.165	0.035
$Gsin(6, 0^o)$	11x11	0.123	0.398	0.275	0.360	0.237	0.211	0.089	0.139	0.016
$Gsin(6, 90^o)$	11x11	0.125	0.397	0.272	0.351	0.226	0.219	0.094	0.155	0.030
I^{obsi}	NxN	$\frac{M-1}{M}$	1	$\frac{1}{M}$	1	$\frac{1}{M}$	1	$\frac{1}{M}$	1	$\frac{1}{M}$

Table 1: The information criterion for filter selection.

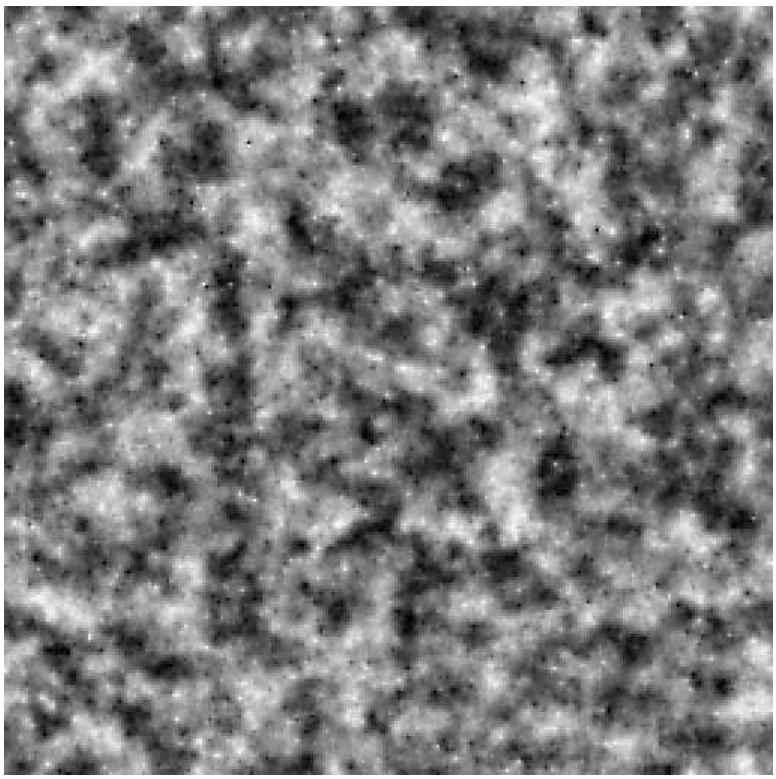


Figure 8: A typical sample of $p_1(\mathbf{I})$ (256×256 pixels).

$\mid \Delta \mathbf{I} \mid \geq 9.5$. As a result, the calculation of $\lambda^{(1)}(z)$ is not precise at the two tails, and we may choose any function $\tilde{\lambda}(z)$ as long as $\tilde{\lambda}(z) = \lambda^{(1)}(z)$ for $\mid z \mid < 9.5$, and $\tilde{\lambda}(z) \geq \lambda^{(1)}(z)$ for $\mid z \mid \geq 9.5$. In figure (7.b), a fitting curve $\psi(x) = 7.4(1 - 1/(1 + (\mid x \mid /10)^2))$ is plotted against $\lambda^{(1)}()$. It fits the solid part of $\lambda^{(1)}()$ very well, and it is bigger than most of the dashed part of $\lambda^{(1)}()$.

A typical sample image of $p_1(\mathbf{I})$, called $\mathbf{I}^{syn1}$, is shown in figure (8). To compute $p_1(\mathbf{I})$ and to sample from $p_1(\mathbf{I})$, a Gibbs sampler is run for $6,000$ sweeps, i.e., each pixel is flipped about $6,000$ times.

Experiment II.

At the second step, for each filter $F^{(\alpha)}$, we estimate $E_{p_1}[H^{(\alpha)}]$ by histogram of $F^{(\alpha)} * \mathbf{I}^{syn1}$, and the information criterion for each filter is list under column $p_1(\mathbf{I})$ in table 1. We notice that the IC for filter Δ drop to near 0, and IC also drops for other filters because these filters are in general not independent of Δ. Some small filters like $LG(1), LG(2)$ have smaller ICs than others, due to higher correlations between them and Δ.

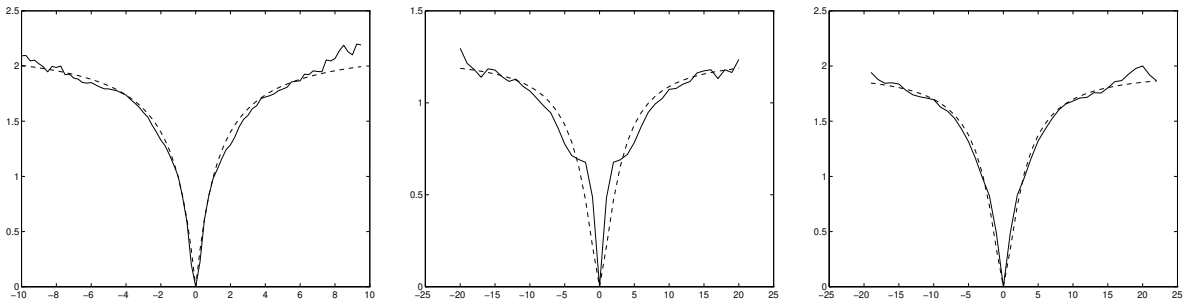
Although big filters have larger IC, computationally it is much expensive to incorporate them into the prior model. As a compromise, we choose both ∇_x and ∇_y as the second and third filter for symmetry. Therefore a prior model $p_3(\mathbf{I})$ is learned,

$$p_3(\mathbf{I}) = \frac{1}{Z} e^{-\sum_{(x,y)} \lambda^{(1)}(\Delta \mathbf{I}) + \lambda^{(2)}(\nabla_x \mathbf{I}) + \lambda^{(3)}(\nabla_y \mathbf{I})}.$$

The potential functions $\lambda^{(\alpha)}(), \alpha = 1, 2, 3$ are plotted in figure (9). Since $\bar{H}^{obs(1)}(z) = 0$ if $\mid z \mid \geq 9.5$, and $\bar{H}^{obs(\alpha)}(z) = 0$ if $\mid z \mid \geq 22$ for $\alpha = 2, 3$, we only plot $\lambda^{(1)}(z)$ for $z \in [-9.5, 9.5]$ and plot $\lambda^{(2)}(z), \lambda^{(3)}(z)$ for $z \in [-22, 22]$. These three curves are fitted with the functions $\psi_1(x) = 2.1(1 - 1/(1 + (\mid x \mid /4.8)^{1.32})$, $\psi_2(x) = 1.25(1 - 1/(1 + (\mid x \mid /2.8)^{1.5})$, $\psi_3(x) = 1.95(1 - 1/(1 + (\mid x \mid /2.8)^{1.5})$ respectively. A synthesized image sampled from $p_3(\mathbf{I})$ is displayed in figure (10).

So far, we have used three filters to extract the structures and statistics of natural images, the synthesized image in figure (10) is still far from natural ones. Especially, even though the learned potential functions $\lambda^{(\alpha)}(z),\ \alpha = 1, 2, 3$ all have flat tails to encourage intensity breaks, but it only generates small speckles instead of big regions and long edges as we previously expected. Based on this synthesized image, we compute the AIG and IC for all filters, and the results are list in table 1 in column $p_3(\mathbf{I})$.

Experiment III.



a b c

Figure 9: The three potential functions for filters a. Δ, b. ∇_x, and c. ∇_y. Dashed curves are the fitting functions: a. $\psi_1(x) = 2.1(1 - 1/(1 + (\mid x \mid /4.8)^{1.32})$, b. $\psi_2(x) = 1.25(1 - 1/(1 + (\mid x \mid /2.8)^{1.5})$, and c. $\psi_3(x) = 1.95(1 - 1/(1 + (\mid x \mid /2.8)^{1.5})$

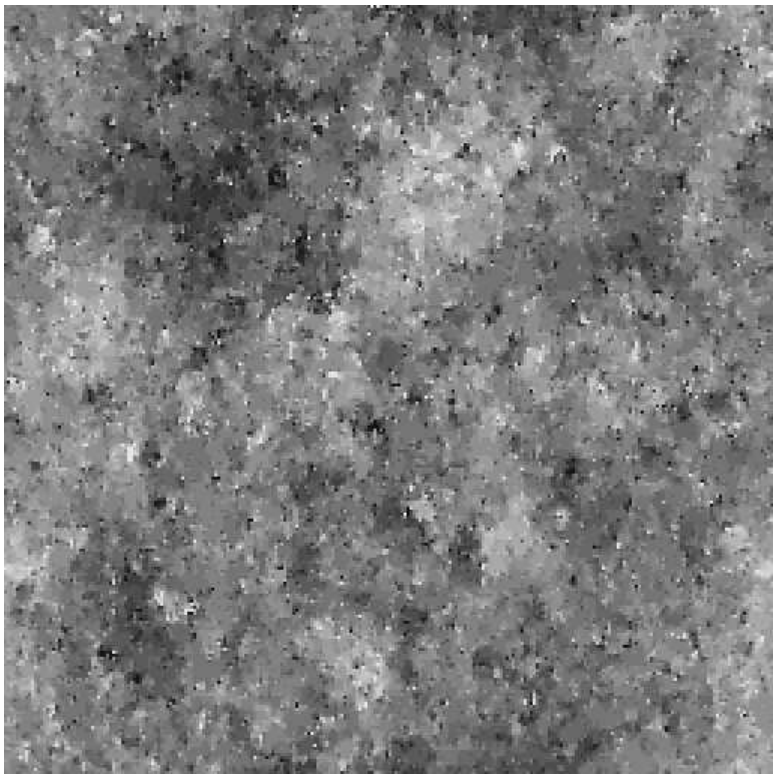


Figure 10: A typical sample of $p_3(\mathbf{I})$ (256×256 pixels).

In this experiment, we shall study a probability distribution which has the scale invariant property with respect to filters ∇_x and ∇_y as we find in natural images. Given an image $\mathbf{I}$ defined on an $N \times N$ lattice $\mathbf{S}$. We build a pyramid in the same way as in section (5.1). Let $\mathbf{I}^{(s)}, s = 0, 1, 2, 3$ be four layers of the pyramid. We set $\mathbf{I}^{(0)} = \mathbf{I}$, and $\mathbf{I}^{(s)}$ is defined on lattice $\mathbf{S}^s$, which is of size $N/2^s \times N/2^s$. Let $H_x^{(s)}(z)$ denotes the histogram of $\nabla_x \mathbf{I}^{(s)}(x, y)$ and $H_y^{(s)}(z)$ denotes the histogram of $\nabla_y \mathbf{I}^{(s)}(x, y)$ for $s = 0, 1, 2, 3$.

We ask for a probability model $p(\mathbf{I})$ which satisfies,

$$E_p[H_x^{(s)}(z, x, y)] = \bar{H}(z), \qquad \forall z \quad \forall (x, y) \in \mathbf{S}^s, \quad s = 0, 1, 2, 3.$$
$$E_p[H_y^{(s)}(z, x, y)] = \bar{H}(z), \qquad \forall z \quad \forall (x, y) \in \mathbf{S}^s, \quad s = 0, 1, 2, 3.$$

where $\bar{H}(z)$ is defined in equation (15). This results in a maximum entropy distribution of the following form,

$$p_s(\mathbf{I}) = \frac{1}{Z} e^{-\sum_{s=0}^{3} \sum_{(x,y) \in \mathbf{S}^s} [\lambda_x^{(s)}(\nabla_x \mathbf{I}^{(s)}) + \lambda_y^{(s)}(\nabla_y \mathbf{I}^{(s)})]}.$$

We learn this probability model, and figure (11) displays $\lambda_x^{(s)}$, $s = 0, 1, 2, 3$. At the beginning of the learning process, all $\lambda_x^{(s)}$, $s = 0, 1, 2, 3$ are of the forms displayed in figure (9) with low values around zero to encourage smoothness. As the learning proceeds, gradually $\lambda_x^{(3)}$ turns "upside down" with lower values at the two tails. Then $\lambda_x^{(2)}$ and $\lambda_x^{(1)}$ turn upside down one by one. Similar results are observed for $\lambda_y^{(s)}$, $s = 0, 1, 2, 3$. Figure (13) is a typical sample image from $p_s(\mathbf{I})$. To demonstrate it has scale invariant properties, in figure (12) we show the histograms $H_x^{(s)}$ and $\log H_x^{(s)}$ of the synthesized image for $s = 0, 1, 2, 3$.

The learning process iterates more than $10,000$ sweeps. To verify the learned λs, we restart a homogeneous Markov chain from noise image using the learned model, after 6000 sweeps, the Markov chain goes to the same point.

Remark 1. In figure (11), we notice that $\lambda_x^{(s)}$ are negative [8] for $s = 1, 2, 3$, as we know, this is contrary to all existing prior models in computer vision. First of all, as the image intensity has finite range $[0, 31]$, $\nabla_x \mathbf{I}^{(s)}$ is defined in $[-31, 31]$. Therefore we may define $\lambda^{(s)}(z) = 0$ for $| z | > 31$, so $p_s(\mathbf{I})$ is legitimately defined. Second, such potentials have significant meanings in visual computation. In image restoration, when a high intensity

[8] Although the potentials are still positive if a positive constant is added to $\lambda_x^{(s)}$, here we use the word "negative" to distinguish the upside-down form from existing models.

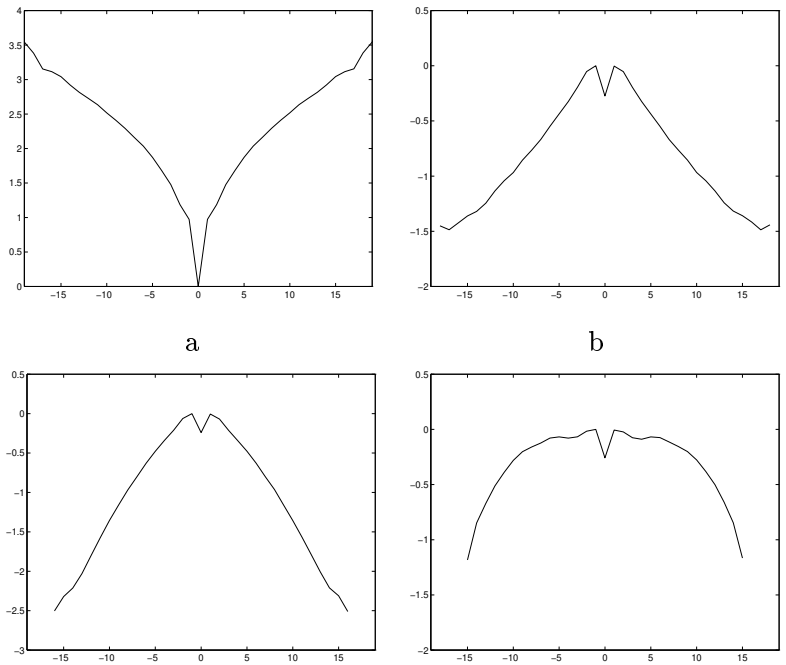


Figure 11: Learned $\lambda_x()$s. a. $\lambda_x^{(0)}$, b. $\lambda_x^{(1)}$, c. $\lambda_x^{(2)}$, d. $\lambda_x^{(3)}$.

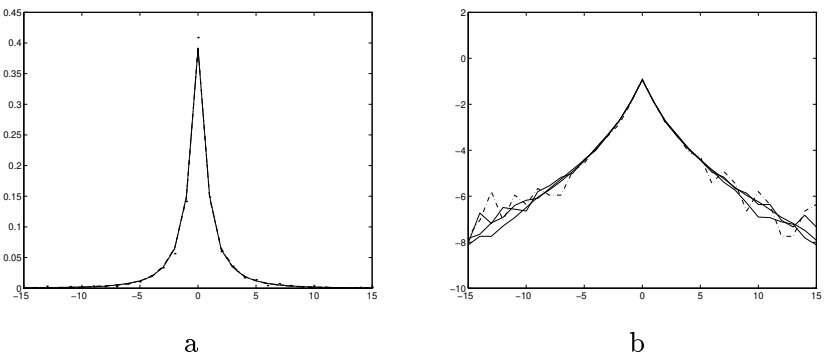


Figure 12: a. the histograms of synthesize image at 4 scales. b. the logarithm of histograms in a.

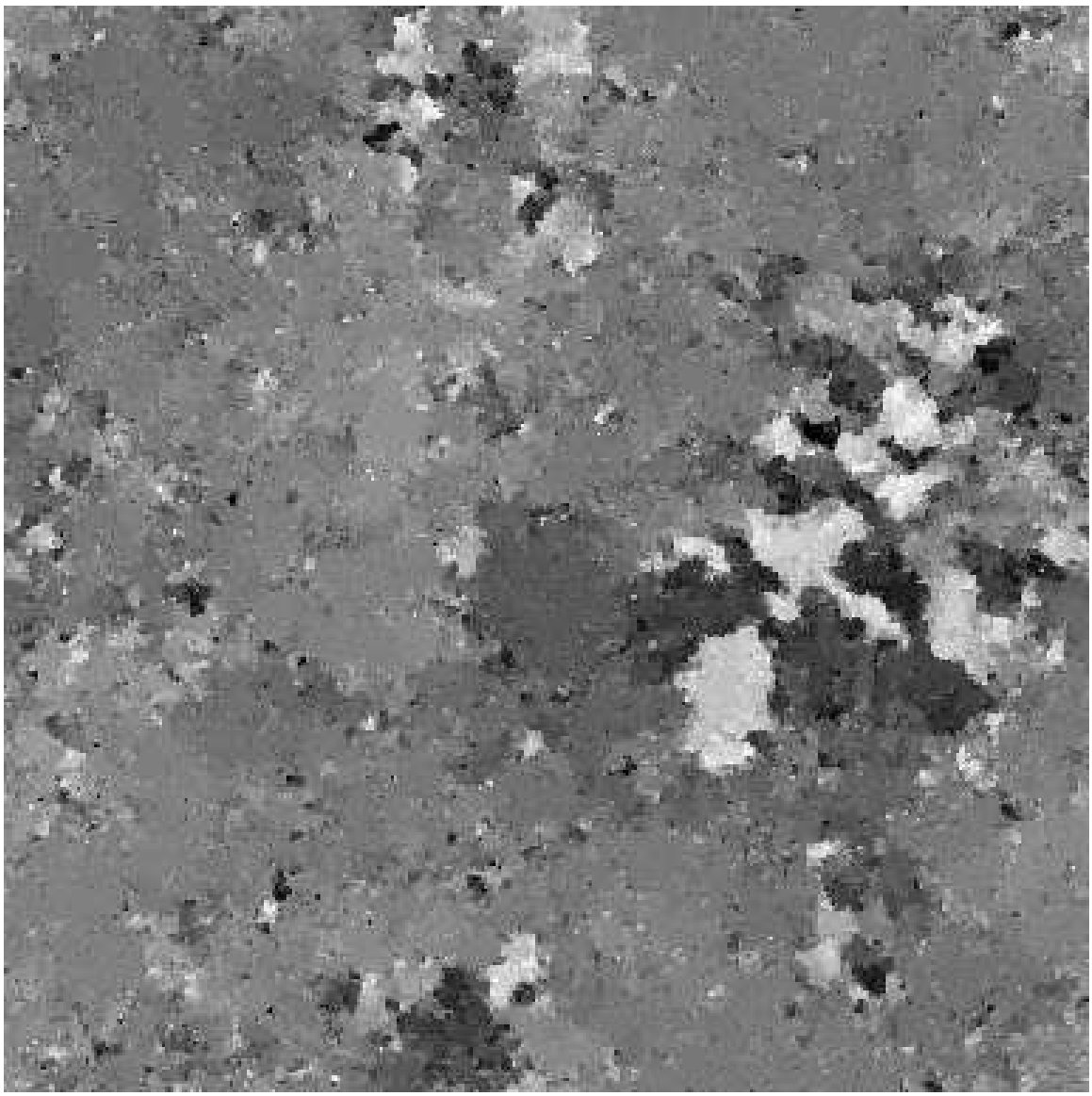


Figure 13: A typical sample of $p_s(\mathbf{I})$ (384×384 pixels).

difference $\nabla_x \mathbf{I}^{(s)}(x,y)$ presents, it is very likely to be noise if $s = 0$. However this is not true for $s = 1, 2, 3$. Additive noises can hardly pass to the high layers of the pyramid because at each layer the 2×2 averaging operator reduce the variance of noise by 4 times. When $\nabla_x \mathbf{I}^{(s)}(x,y)$ is large for $s = 1, 2, 3$, it is more likely to be a true edge and object boundary. So in $p_s(\mathbf{I})$, $\lambda^{(0)}$ suppresses noise at the first layer, while $\lambda^{(s)}, s = 1, 2, 3$ encourage sharp edges to form, and thus enhance blurred boundaries. We notice that figure (13) shows regions of various scales, and the intensity contrasts are also higher at the boundary. These are missing in figure (8) and figure (10). Third, based on the image in figure (13), we compute IC and AIG for all filters and list them in table 1 under column $p_s(\mathbf{I})$. We note that the IC for all filters drops below 0.08, which suggests that $p_s(\mathbf{I})$ is close to $f(\mathbf{I})$, and simply adding more filters may not improve this generic prior model very much.

Remark 2. As images are often presented at unknown scales, a good prior model should be scale invariant. This leads to an interesting but difficult problem in physics–the renormalization group theory (Wilson 1975), but existing techniques cannot provide an analytical solution to this question. Instead we only study $p_s(\mathbf{I})$ which is scale invariant at a few scales.

Remark 3. All the prior models that we learned as well as the existing prior models discussed in section 1 have no preference about the image intensity domain. Therefore the image intensity has uniform distribution, but we limit it inside $[0, 31]$, thus the first row of table 1 has the same value for IC and AIG.

6 Comparison of prior models

This section compares the performance of $p_s(\mathbf{I})$ in image restoration and segmentation with two previously used models, 1) the line process model denoted by $p_l(\mathbf{I})$, 2) the T-function prior denoted by $p_t(\mathbf{I})$. $p_l(\mathbf{I})$ and $p_t(\mathbf{I})$ are displayed in figure (1.b), (1.c) respectively.

In the experiments, given an image $\mathbf{I}$, we distort it to $\mathbf{I}^d$ by smoothing and/or adding i.i.d. Gaussian noise from distribution $N(\mu, \sigma^2)$. The data model $p(\mathbf{I}^d \mid \mathbf{I})$ is known as a Gaussian. Then choosing a prior model $p(\mathbf{I})$, by Bayesian rule we restore an image $\mathbf{I}$ by maximizing *a posteriori* probability (MAP) $p(\mathbf{I}^d \mid \mathbf{I})p(\mathbf{I})$. Following (Geman and Geman

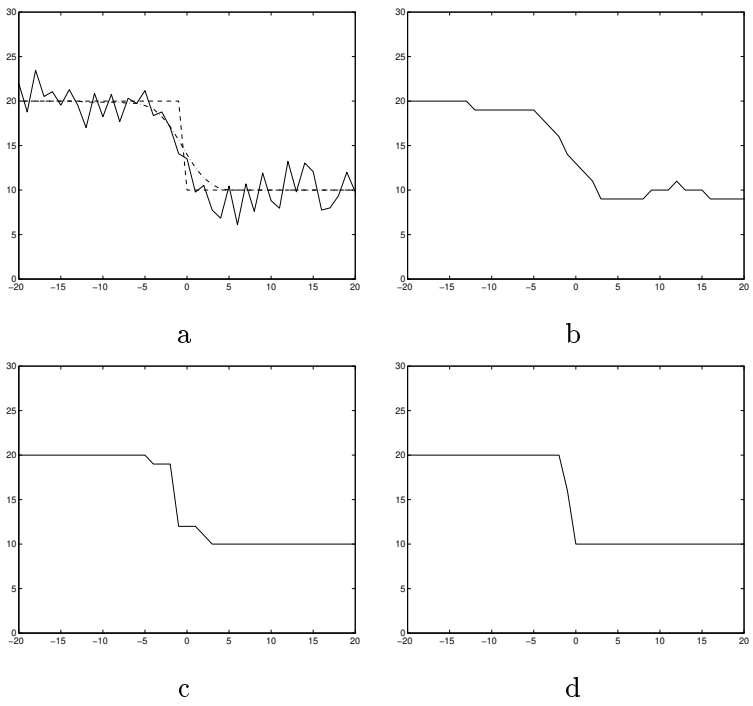


Figure 14: a. A step edge (dashed curve), the smoothed edge (dash-dotted curve), then added with Gaussian noise (solid); b,c,d are respectively the restored edges using $p_l(\mathbf{I})$, $p_t(\mathbf{I})$, and $p_s(\mathbf{I})$.

1984), we use simulated annealing to compute the MAP-estimate.

In figure (14.a), a step edge is first smoothed and Gaussian noises are added. The restored step edges by prior models $p_l(\mathbf{I})$, $p_t(\mathbf{I})$ and $p_s(\mathbf{I})$ are respectively shown in figure (14.b), (14.c), and (14.d). The parameters for $p_l(\mathbf{I})$ and $p_t(\mathbf{I})$ are adjusted so that the best effects are presented. Obviously $p_s(\mathbf{I})$ is the strongest in enhancing the blurred boundary. The potential function in the line process model is quadratic and flat around zero, thus it allows small fluctuation between adjacent locations.

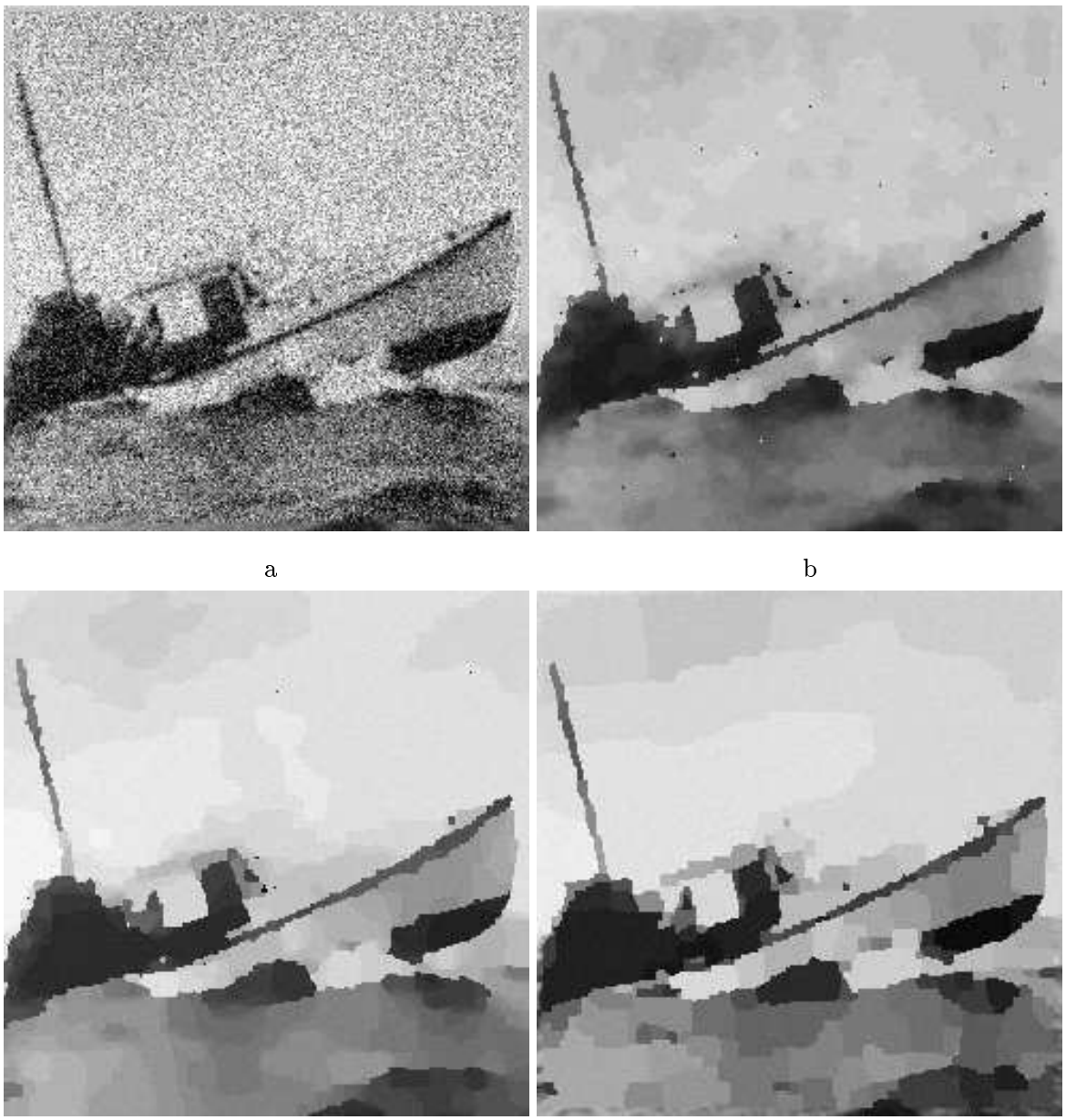
Figure (15) demonstrates a 2D experiment. The original image is the lobster boat displayed in figure (2). It is normalized to have intensity in $[0, 31]$ and Gaussian noises from $N(0, 25)$ are added. The distorted image is displayed in figure (15.a), where we keep the image boundary noise-free for the convenience of boundary condition. The restored images using $p_l(\mathbf{I})$, $p_t(\mathbf{I})$ and $p_s(\mathbf{I})$ are shown in figure (15.a), (15.c), (15.d) respectively. $p_s(\mathbf{I})$ appears to have the best effect in recovering the boat, especially the top of the boat, but it also enhance the water.

7 Discussion

In this paper, some statistics properties of natural images are studied and a general theory is proposed, based on a maximum entropy principle, for learning generic prior models for natural images. A novel information criterion is also studied for model selection by minimizing a Kullback-Leibler information distance. This information criterion is different from the Akaike information criterion (Akaike 1976), the latter is used for limiting the complexity of auto regression models. Compared with the previously used prior models, our theory answers two questions–how to choose filters or features and what is the potential functions suggested by real images. Moreover we discuss the scale invariant properties of both natural images and prior model. We observe that the models used in RBF learning and image coding theory are two extreme cases of our theory.

We argue that the same strategy developed in this paper can be used in other applications. For example, learning prior models for MRI images in medical image processing, and for 3D surface reconstruction, the targets are no longer natural images, and different prior models shall be learned according to our theory.

There is a major drawback in our experiments. The natural images that we collect



a b

c d

Figure 15: a. The noise distorted image, b. c. d. are respectively restored images by prior models $p_l(\mathbf{I})$, $p_t(\mathbf{I})$ and $p_s(\mathbf{I})$.

are subject to the influences of many artificial factors, such as noise distortions, and γ-effect of cameras. The exact natural images are never available to our experiments, and people have different opinions about what natural images should look like. Changing the database from country scenes to urban scenes shall change the statistics we observe. In fact, all unsupervised learning methods suffer from the same problems. The only way for us to collect a good database is to select images which look "natural" subjectively. We abandon images with noticeable noises, and images which are over-exposed or under-exposed. Some preprocessing of images are possible, for example people take the logarithm of images as inputs (Ruderman and Bialek 1994) or calibrate the camera carefully, but we argue that these methods are less relevant. We are learning a probability model for $\mathbf{I}$ in image space $\mathbf{L}^{|\mathbf{S}|}$, and the goal of learning prior models is to look for common structures and their statistics which bias a vision algorithm in image reconstruction. As long as the images in the database looks natural to human perception, the statistics of those differential operators shall be less influenced by the noises and γ-effects.

Furthermore, although the synthesized images bear important features of natural images, they are still far from realistic ones. In other words, these generic prior models can do very little beyond image restoration. This is mainly due to the fact that all generic prior models are assumed to be translation invariant. This homogeneity assumption is unrealistic, since for a given image, local statistics varies from locations to locations. Although table 1 suggests that $p_s(\mathbf{I})$ is close to the underlying probability $f(\mathbf{I})$, however for a constrained set of images, more specific probability model is called for. For example, some realistic texture patterns are synthesized using similar method (Zhu, Wu, Mumford 1996), where a probability model is tuned to each observed texture pattern.

Many work in wavelet coding (Coifman and Wickerhauser 1992, Donoho 1995, Simoncelli and Adelson 1996) adopt different set of wavelets to represent the signals at different locations, but they do not discuss how to build a prior model. Instead in all the wavelet coding literatures, prior models are subjectively chosen such as the L_p-norm (Donahue and Geiger 1993). A common assumption for those work is that all subbands are independent of each other.

Recently, Ruderman studied correlations of $\log \mathbf{I}$, and explained certain scale invariance by synthesizing images with square-shaped regions of various sizes. With carefully designed probability models about intensity inside region and between regions, the corre-

lation of the synthesized images is reported to be close to the observed (Ruderman 1996). This paper studies the first order statistics and learns $p_s(\mathbf{I})$ without resorting to high level concepts.

We call the generic prior models studied in this paper *the first level prior*. All the above discussion points to a more sophisticated prior model, which we call *second level prior*. It shall incorporate concepts like object geometry, and such prior model is used in image segmentation (Zhu and Yuille 1995). It is our hope that this article will stimulate further investigations along this direction to build more realistic prior models.

Acknowledgments. It is a pleasure to acknowledge many stimulating discussions with Y. Wu and S. Geman, B. Gidas. This work is supported by an ARO grant.

Appendix Proof of theorem 2

By definition,

$$p_0(\mathbf{I}) = \frac{1}{Z_0} e^{-\sum_{\alpha=1}^{k} <\lambda_0^{(\alpha)}, H^{(\alpha)}>}, \quad p(\mathbf{I}) = \frac{1}{Z} e^{-\sum_{\alpha=1}^{k} <\lambda^{(\alpha)}, H^{(\alpha)}>}.$$

Since $E_{p_0}[H^{(\alpha)}] = h_0^{(\alpha)}$, for $\alpha = 1, 2, ..., k$, we have the Kullback-Leibler distance,

$$I(h) = E_{p_0}[\log p_0(\mathbf{I})] + \log Z + \sum_{\alpha=1}^{k} < \lambda^{(\alpha)}, h_0 >$$

where h_0, $\lambda_0^{(\alpha)}$ and $p_0(I)$ are fixed, $\lambda^{(\alpha)} = \lambda^{(\alpha)}(h^{(1)}, h^{(2)}, ..., h^{(k)})$, for $\alpha = 1, 2, ..., k$, and $Z = Z(\lambda^{(1)}, ..., \lambda^{(k)})$. Then

$$\begin{aligned} \frac{\partial I(h)}{\partial h^{(\beta)}} &= \frac{1}{Z}\frac{\partial Z}{\partial h^{(\beta)}} + \sum_{\alpha=1}^{k} < \frac{\partial \lambda^{(\alpha)}}{\partial h^{(\beta)}}, h_0 > \\ &= \frac{1}{Z}\sum_{\alpha=1}^{k} < \frac{\partial Z}{\partial \lambda^{(\alpha)}}, \frac{\partial \lambda^{(\alpha)}}{\partial h^{(\beta)}} > + \sum_{\alpha=1}^{k} < \frac{\partial \lambda^{(\alpha)}}{\partial h^{(\beta)}}, h_0 > \end{aligned}$$

Since $-\frac{1}{Z}\frac{\partial Z}{\partial \lambda^{(\alpha)}} = h^{(\alpha)}$ (see property i of the partition function), therefore,

$$\frac{\partial I(h)}{\partial h^{(\beta)}} = \sum_{\alpha=1}^{k} < \frac{\partial \lambda^{(\alpha)}}{\partial h^{(\beta)}}, \quad h_0^{(\alpha)} - h^{(\alpha)} > .$$

So $\frac{\partial I(h_0)}{\partial h^{(\beta)}} = 0, \quad \beta = 1, 2, ..., k$. Taking the second derivative at $h = h_0$, we have,

$$\frac{\partial^2 I(h_0)}{\partial h^{(\beta)} \partial h^{(\gamma)}} = -\frac{\partial \lambda^{(\gamma)}}{\partial h^{(\beta)}} = [\frac{\partial h^{(\beta)}}{\partial \lambda^{(\gamma)}}]^{-1}. \tag{18}$$

Again, since $= h^{(\beta)} = -\frac{1}{Z}\frac{\partial Z}{\partial \lambda^{(\beta)}}$, we have,

$$\begin{aligned}\frac{\partial h^{(\beta)}}{\partial \lambda^{(\gamma)}} &= \frac{1}{Z}\frac{\partial^2 Z}{\partial h^{(\gamma)}\partial h^{(\beta)}} - \frac{1}{Z^2}\frac{\partial Z}{\partial h^{(\gamma)}}\frac{\partial Z}{\partial h^{(\beta)}} \\ &= E_p[H^{(\beta)}H^{(\gamma)}] - h^{(\beta)}h^{(\gamma)} \\ &= E_p[(H^{(\beta)} - h^{(\beta)})^T(H^{(\gamma)} - h^{(\gamma)})].\end{aligned}$$

Let $Var(h_0)$ be matrix of $\frac{\partial^2 I(h_0)}{\partial h^{(\beta)}\partial h^{(\gamma)}}$, by Taylor expansion at h_0,

$$I(p_0, p) = I(p_0, p_0) + \sum_{\alpha=1}^{k} < \frac{dI(h_0)}{d\,h^{(\alpha)}}, h - h_0 > + \frac{1}{2}(h - h_0)Var^{-1}(h_0)(h - h_0) + o((h - h_0)^3)$$

The conclusion follows $I(p_0, p_0) = 0$ and $\frac{dI(h_0)}{dh} = 0$ □

References

[1] H. Akaike, "Canonical correlation analysis of time series and the use of an information criterion. *Systems Identification: Advances and Case Studies*, eds. R. K. Mehra etc. Academic Press, New York, 1976.

[2] P. N. Belhumeur, "A binocular stereo algorithm for reconstructing sloping, creased and broken surfaces, in the presence of half-occlusion", *Proc. ICCV*, Berlin, 1993.

[3] M. J. Black and A. Rangarajan, "The outlier process: unifying line processes and robust statistics." *Proc. of CVPR* Seattle, Washington, 1994.

[4] A. Blake and A. Zisserman. *Visual Reconstruction.* MIT press, 1987.

[5] R. R. Coifman and M. V. Wickerhauser, Entropy-based algorithms for best basis selection. *IEEE Trans. on Information Theory.* vol. 38, pp.713-718, 1992.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., 1985.

[7] J. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of Optical Society of America.* vol. 2, No. 7, pp1160-1169. 1985.

[8] M. J. Donahue and D. Geiger, " Template matching and function decomposition using non-minimal spanning sets", Tech. Report. Siemens, 1993.

[9] D. L. Donoho, De-Noising by soft thresholding, *IEEE Trans. on Information Theory.* vol.41, pp.613-627, 1995.

[10] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells", *J. of Optical soc. America, A*, vol.4, No. 12, 1987.

[11] D. Gabor, "Theory of communication." IEE Proc.vol 93, no.26. 1946.

[12] S. Geman and D. Geman. "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images". *IEEE Trans. on PAMI* 6(7), pp 721-741, 1984.

[13] S. Geman and D. McClure, "Statistical methods for tomographic image reconstructions", *Bulletin of the International Statistical Institute*, **52**, 4-20. 1987.

[14] B. Gidas, "A renormalization group approach to image processing problems". *IEEE Trans. on PAMI*, vol. 11, No.2, Feb. 1989.

[15] U. Grenander, *Lectures in Pattern Theory.* vol. 1, Springer-verlag, New York, 1976.

[16] S. Mallat, "A theory for multi-resolution signal decomposition: the wavelet representation", *IEEE trans on Pattern Analysis and Machine Intelligence*, vol. 11, No.7, 674-693, 1989.

[17] D. Mumford and J. Shah. "Optimal approximations by piecewise smooth functions and associated variational problems." *Comm. Pure Appl. Math.,* **42**, pp 577-684, 1989.

[18] B. A. Olshausen and D. J. Field, "Natural image statistics and efficient coding", *Proc. of workshop on Information Theory and the Brain*, Setpember, 1995.

[19] T. Poggio, V. Torre and C. Koch, "Computational vision and regularization theory", *Nature*, vol. 317, pp 314-319, 1985.

[20] T. Poggio and F. Girosi, "Networks for approximation and learning", *Proc. of IEEE*, vol.78, 1481-1497, 1990.

[21] J. G. Propp and D. B. Wilson, "Exact sampling with coupled Markov chains and applications to statistical mechanics", Tech. report, Math Dept. MIT. 1995.

[22] D. L. Ruderman and Bialek, "Statistics of natural images: scaling in the woods", *Phys. Rev. Letter*, 73:814-817, 1994.

[23] D. L. Ruderman, "Origins of scaling in natural images". *Proc. of IS&T/SPIE Symposium on Electronic Imaging.* 1996.

[24] J. Shah, "A common framework for curve evolution, segmentation, and anisotropic diffusion", *Proc. of CVPR*, San Fran. 1996.

[25] E. P. Simoncelli and E. H. Adelson, "Noise removal via Bayesian wavelet coring", *Int'l Conf. on Image Processing, Switzerland, 1996.*

[26] D. Terzopoulos, "Multilevel computational processes for visual surface reconstruction". *Computer Visiaon, Graphics, and Image Processing*, **24**, 52-96, 1983.

[27] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-posed Problems*, 1906, (Translated version), V.H.Winston & Sons, 1977.

[28] A. B. Watson, "Efficiency of model human image code", *Journal of Optical Society of America A*. Vol.4, No.12, 1987.

[29] K. Wilson, "The renormalization group: critical phenonmena and the Knodo problem," Rev. Mod. Phys., Vol.47, pp.773-840, 1975.

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

[31] L. Younes, "Estimation and annealing for Gibbs fields", *Annales de l'Institut Henri Poincare, Section B, Calcul des Probabilities et Statistique* Vol.24 pp269-294, 1988.

[32] S. C. Zhu and A. L. Yuille "Region Competition: unifying snakes, region growing, and Bayes/MDL for multi-band image segmentation". To appear in *IEEE Trans.on PAMI.* Sept. 1996.

[33] S. C. Zhu, Y. N. Wu and D. B. Mumford. "Filters, Random Fields, and Minimax Entropy (FRAME): Towards a unified theory for texture modeling". *Proc. Comp. Vision and Patt. Recog.*, San Fran, 1996.