Prior Learning and Gibbs Reaction-Diffusion

Prior Learning and Gibbs Reaction-Diffusion

Song Chun Zhu and David Mumford

Abstract—This article addresses two important themes in early visual computation: First, it presents a novel theory for learning the universal statistics of natural images—a prior model for typical cluttered scenes of the world—from a set of natural images, and, second, it proposes a general framework of designing reaction-diffusion equations for image processing. We start by studying the statistics of natural images including the scale invariant properties, then generic prior models were learned to duplicate the observed statistics, based on the minimax entropy theory studied in two previous papers. The resulting Gibbs distributions have potentials of the form $U(\mathbf{I}; \Lambda, S) = \sum_{\alpha=1}^{K} \sum_{(x,y)} \lambda^{(\alpha)}\left(\left(F^{(\alpha)} * \mathbf{I}\right)(x, y)\right)$ with $S = \{F^{(1)}, F^{(2)}, ..., F^{(K)}\}$ being a set of filters and $\Lambda = \{\lambda^{(1)}(), \lambda^{(2)}(), ..., \lambda^{(K)}()\}$ the potential functions. The learned Gibbs distributions confirm and improve the form of existing prior models such as line-process, but, in contrast to all previous models, *inverted* potentials (i.e., $\lambda(x)$ decreasing as a function of $|x|$) were found to be necessary. We find that the partial differential equations given by gradient descent on $U(\mathbf{I}; \Lambda, S)$ are essentially reaction-diffusion equations, where the usual energy terms produce anisotropic diffusion, while the inverted energy terms produce reaction associated with pattern formation, enhancing preferred image features. We illustrate how these models can be used for texture pattern rendering, denoising, image enhancement, and clutter removal by careful choice of both prior and data models of this type, incorporating the appropriate features.

Index Terms—Visual learning, Gibbs distribution, reaction-diffusion, anisotropic diffusion, texture synthesis, clutter modeling, image restoration.

1 Introduction

In computer vision, many generic prior models have been proposed to capture the universal low level statistics of natural images. These models presume that surfaces of objects be smooth, and adjacent pixels in images have similar intensity values unless separated by edges, and they are applied in vision algorithms ranging from image restoration, motion analysis, to 3D surface reconstruction.

For example, in image restoration, general smoothness models are expressed as probability distributions [9], [4], [20], [11]:

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

where $\mathbf{I}$ is the image, Z is a normalization factor, and $\nabla_x \mathbf{I}(x, y) = \mathbf{I}(x + 1, y) - \mathbf{I}(x, y)$, $\nabla_y \mathbf{I}(x, y) = \mathbf{I}(x, y + 1) - \mathbf{I}(x, y)$ are differential operators. Three typical forms of the potential function $\psi()$ are displayed in Fig. 1. The functions in Fig. 1b and Fig. 1c have flat tails to preserve edges and object boundaries, and thus they are said to have advantages over the quadratic function in Fig. 1a.

These prior models have been motivated by regularization theory [26], [18],[1] physical modeling [31], [4],[2] Bayesian theory [9], [20], and robust statistics [19], [13], [3]. Some connections between these interpretations are also observed in [12], [13] based on effective energy in statistics mechanics. Prior models of this kind are either generalized from traditional physical models [37] or chosen for mathematical convenience. There is, however, little rigorous theoretical or empirical justification for applying these prior models to generic images, and there is little theory to guide the construction and selection of prior models. One may ask the following questions:

1) Why are the differential operators good choices in capturing image features?
2) What is the best form for $p(\mathbf{I})$ and $\psi()$?
3) A relevant fact is that real world scenes are observed at more or less arbitrary scales, thus a good prior model should remain the same for image features at multiple scales. However none of the existing prior models has the scale-invariance property on the 2D image lattice, i.e., is renormalizable in terms of renormalization group theory [36].

In previous work on modeling textures, we proposed a new class of Gibbs distributions of the following form [40], [41],

$$p(\mathbf{I}; \Lambda, S) = \frac{1}{Z} e^{-U(\mathbf{I}; \Lambda, S)} \tag{2}$$

- *S.C. Zhu is with the Computer Science Department, Stanford University, Stanford, CA 94305. E-mail: szhu@cs.stanford.edu.*
- *D. Mumford is with the Division of Applied Mathematics, Brown University, Providence, RI 02912.*

Manuscript received 12 July 1996; revised 15 Sept. 1997. Recommended for acceptance by B. Vemuri.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 105703.

1. Where the smoothness term is explained as a stabilizer for solving "ill-posed" problems [32].
2. If $\psi()$ is quadratic, then variational solutions minimizing the potential are splines, such as flexible membrane or thin plate models.

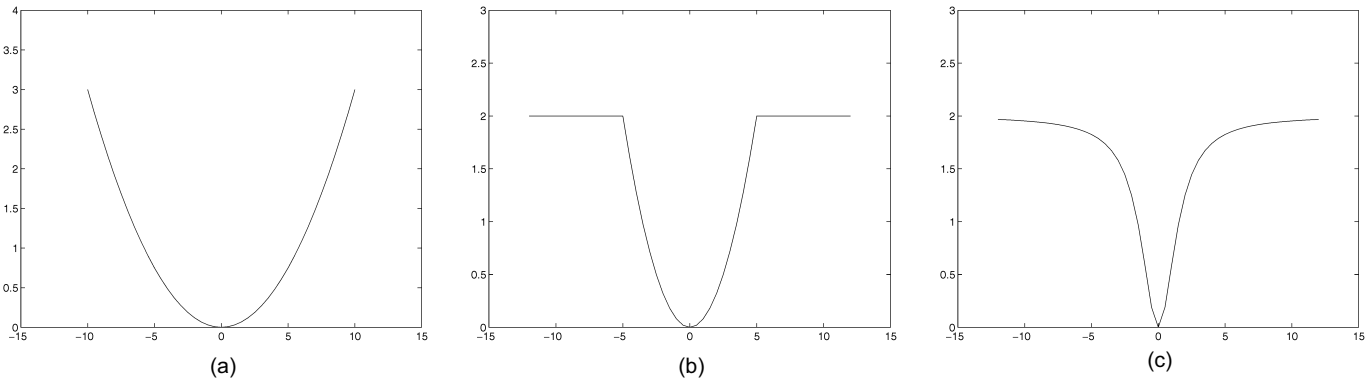


Fig. 1. Three existing forms for $\psi()$. (a) Quadratic: $\psi(\xi) = a\xi^2$. (b) Line process: $\psi(\xi) = a\min(\theta^2, \xi^2)$. (c) T-function: $\psi(\xi) = a\left(1 - \frac{1}{1+c\xi^2}\right)$.

$$U(\mathbf{I}; \Lambda, S) = \sum_{\alpha=1}^{K} \sum_{(x,y)} \lambda^{(\alpha)}\left(\left(F^{(\alpha)} * \mathbf{I}\right)(x, y)\right). \quad (3)$$

In the above equation, $S = \{F^{(1)}, F^{(2)}, ..., F^{(K)}\}$ is a set of linear filters, and $\Lambda = \{\lambda^{(1)}(), \lambda^{(2)}(), ..., \lambda^{(K)}()\}$ is a set of potential functions on the features extracted by S. The central property of this class of models is that they can reproduce the marginal distributions of $F^{(\alpha)} * \mathbf{I}$ estimated over a set of the training images $\mathbf{I}$—while having the maximum entropy—and the best set of features $\{F^{(1)}, F^{(2)}, ..., F^{(K)}\}$ is selected by minimizing the entropy of $p(\mathbf{I})$ [41]. The conclusion of our earlier papers is that, for an appropriate choice of a small set of filters S, random samples from these models can duplicate very general classes of textures—as far as normal human perception is concerned. Recently, we found that similar ideas of model inference using maximum entropy have also been used in natural language modeling [1].

In this paper, we want to study to what extent probability distributions of this type can be used to model *generic* natural images, and we try to answer the three questions raised above.

We start by studying the statistics of a database of 44 real world images, and then we describe experiments in which Gibbs distributions in the form of (2) were constructed to duplicate the observed statistics. The learned potential functions $\lambda^{(\alpha)}()$, $\alpha = 1, 2, ..., K$ can be classified into two categories: *diffusion terms* which are similar to Fig. 1c, and *reaction terms* which, in contrast to all previous models, have *inverted* potentials (i.e., $\lambda(x)$ decreasing as a function of $|x|$).

We find that the partial differential equations given by gradient descent on $U(\mathbf{I}; \Lambda, S)$ are essentially reaction-diffusion equations, which we call the *Gibbs Reaction and Diffusion Equations* (Grade). In Grade, the diffusion components produce denoising effects which are similar to the anisotropic diffusion [25], while reaction components form patterns and enhance preferred image features.

The learned prior models are applied to the following applications.

First, we run the Grade starting with white noise images and demonstrate how Grade can easily generate canonical texture patterns, such as leopard blobs and zebra stripe, as the Turing reaction-diffusion equations do [34], [38]. Thus our theory provides a new method for designing PDEs for pattern synthesis.

Second, we illustrate how the learned models can be used for denoising, image enhancement, and clutter removal by careful choice of both prior and noise models of this type, incorporating the appropriate features extracted at various scales and orientations. The computation simulates a stochastic process—the Langevin equations—for sampling the posterior distribution.

This paper is arranged as follows: Section 2 presents a general theory for prior learning. Section 3 demonstrates some experiments on the statistics of natural images and prior learning. Section 4 studies the reaction-diffusion equations. Section 5 demonstrates experiments on denoising, image enhancement and clutter removal. Finally, Section 6 concludes with a discussion.

2 Theory of Prior Learning

2.1 Goal of Prior Learning and Two Extreme Cases

We define an image $\mathbf{I}$ on an $N \times N$ lattice L to be a function such that for any pixel (x, y), $\mathbf{I}(x, y) \in \mathcal{L}$, and $\mathcal{L}$ is either an interval of $\mathbf{R}$ or $\mathcal{L} \subset \mathbf{Z}$. We assume that there is an underlying probability distribution $f(\mathbf{I})$ on the image space $\mathcal{L}^{N^2}$ for general natural images—arbitrary views of the world. Let $NI^{obs} = \left\{\mathbf{I}_n^{obs}, n = 1, 2, \ldots, M\right\}$ be a set of observed images which are independent samples from $f(\mathbf{I})$. *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 $p(\mathbf{I})$* as an estimation of $f(\mathbf{I})$, *so that $p(\mathbf{I})$, as a prior model, will bias vision algorithms against image features which are not typical in natural images, such as noise distortion and blurring.* For this objective, it is reasonable to assume that any image features have an equal chance to occur at any location, so $f(\mathbf{I})$ is translation invariant with respect to (x, y). We will discuss the limits of this assumption in Section 6.

To study the properties of images $\left\{\mathbf{I}_n^{obs}, n = 1, 2, \ldots, M\right\}$, we start from exploring a set of linear filters $S = \{F^{(\alpha)}, \alpha = 1, 2, ..., K\}$ which are characteristic of the observed images. The statistics extracted by S are the empirical marginal distributions (or histograms) of the filter responses.

DEFINITION 1. *Given a probability distribution $f(\mathbf{I})$, the marginal distribution of $f(\mathbf{I})$ with respect to $F^{(\alpha)}$ is*:

$$f^{(\alpha)}(z) = \int \cdot \int_{F^{(\alpha)} * \mathbf{I}(x,y) = z} f(\mathbf{I}) d\mathbf{I} = E_f\Big[\delta\big(z - F^{(\alpha)} * \mathbf{I}(x,y)\big)\Big]$$
$$\forall (x,y) \in L$$

where $\forall z \in \mathbf{R}$ and $\delta()$ is a Dirac function with point mass concentrated at zero.

DEFINITION 2. *Given a linear filter $F^{(\alpha)}$ and an image $\mathbf{I}$, the empirical marginal distribution (or histogram) of the filtered image $F^{(\alpha)} * \mathbf{I}(x, y)$ is*:

$$H^{(\alpha)}(z; \mathbf{I}) = \frac{1}{|L|} \sum_{(x,y) \in L} \delta\big(z - F^{(\alpha)} * \mathbf{I}(x,y)\big).$$

We compute the histogram averaged over all images in NI^{obs} as the *observed statistics*,

$$\mu_{obs}^{(\alpha)}(z) = \frac{1}{M} \sum_{n=1}^{M} H^{(\alpha)}\big(z; \mathbf{I}_n^{obs}\big), \quad \alpha = 1, 2, \ldots, K.$$

If we make a good choice of our database, then we may assume that $\mu_{obs}^{(\alpha)}(z)$ is an unbiased estimate for $f^{(\alpha)}(z)$, and as $M \to \infty$, $\mu_{obs}^{(\alpha)}(z)$ converges to $f^{(\alpha)}(z) = E_f[H^{(\alpha)}(z; \mathbf{I})]$.

Now, to learn a prior model from the observed images $\{I_n^{obs}, n = 1, 2, \ldots, M\}$, immediately we have two simple solutions. The first one is:

$$p(\mathbf{I}) = \prod_{(x,y) \in L} \mu_{obs}^{(o)}\big(\mathbf{I}(x,y)\big), \tag{4}$$

where $\mu_{obs}^{(o)}$ is the observed average histogram of the image intensities, i.e., the filter $F^{(o)} = \delta$ is used. Taking $\psi_1(z) = -\log \mu_{obs}^{(o)}(z)$, we rewrite (4) as

$$p(\mathbf{I}) = \frac{1}{Z} e^{-\sum_{(x,y)} \psi_1(\mathbf{I}(x,y))}. \tag{5}$$

The second solution is:

$$p(\mathbf{I}) = \frac{1}{M} \sum_{n=1}^{M} \delta\big(\mathbf{I} - \mathbf{I}_n^{obs}\big). \tag{6}$$

Let $\big\|\mathbf{I}_n^{obs}\big\|^2 = c_n$ for $n = 1, 2, \ldots, M$, (6) becomes

$$p(\mathbf{I}) = \frac{1}{M} e^{-\sum_{n=1}^{M} \psi_2\left(\langle \mathbf{I}_n^{obs}, \mathbf{I} \rangle - c_n\right)}, \tag{7}$$

where $\langle , \rangle$ is inner product, $\psi_2(0) = 0$, and $\psi_2(x) = \infty$ if $x \neq 0$, i.e., $\psi_2()$ is similar to the Potts model [37].

These two solutions stand for two typical mechanisms for constructing probability models in the literature. The first is often used for image coding [35], and the second is a special case of the learning scheme using radial basis functions (RBF) [27].[3] Although the philosophies for learning these two prior models are very different, we observe that they share two common properties.

3. 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.

1) *The potentials $\psi_1()$, $\psi_2()$ are built on the responses of linear filters.* In (7), $\mathbf{I}_n^{obs}$, $n = 1, 2, \ldots M$ are used as linear filters of size $N \times N$ pixels, which we denote by $F^{(obsn)} = \mathbf{I}_n^{obs}$.

2) For each filter $F^{(\alpha)}$ chosen, $p(\mathbf{I})$ in both cases duplicates the observed marginal distributions. It is trivial to prove that in both cases $E_p\Big[H^{(\alpha)}\big(z; \mathbf{I}\big)\Big] = \mu_{obs}^{(\alpha)}(z)$, thus as M increases, $E_p\Big[H^{(\alpha)}\big(z; \mathbf{I}\big)\Big] \to E_f\Big[H^{(\alpha)}\big(z; \mathbf{I}\big)\Big]$.

This second property is in general not satisfied by existing prior models in (1). $p(\mathbf{I})$ in both cases meets our objective for prior learning, but intuitively they are not good choices. In (5), the $\delta()$ filter does not capture spatial structures of larger than one pixel, and in (7), filters $F^{(obsn)}$ are too specific to predict features in unobserved images.

In fact, the filters used above lie in the two extremes of the spectrum of all linear filters. As discussed by Gabor [7], 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. Filter $F^{(obsn)}$ includes a specific combination of all the components in both space and frequency. A quantitative analysis of the goodness of these filters is given in Table 1 in Section 3.2.

2.2 Learning Prior Models by Minimax Entropy

To generalize the two extreme examples, it is desirable to compute a probability distribution which duplicates the observed marginal distributions for an arbitrary set of filters, linear or nonlinear. This goal is achieved by a minimax entropy theory studied for modeling textures in our previous papers [40], [41].

Given a set of filters $\{F^{(\alpha)}, \alpha = 1, 2, ..., K\}$, and observed statistics $\big\{\mu_{obs}^{(\alpha)}(z), \alpha = 1, 2, \ldots, K\big\}$, a maximum entropy distribution is derived which has the following Gibbs form:

$$p\big(\mathbf{I}; \Lambda, S\big) = \frac{1}{Z} e^{-U(\mathbf{I}; \Lambda, S)} \tag{8}$$

$$U\big(\mathbf{I}; \Lambda, S\big) = \sum_{\alpha=1}^{K} \sum_{(x,y)} \lambda^{(\alpha)}\Big(\big(F^{(\alpha)} * \mathbf{I}\big)(x,y)\Big) \tag{9}$$

In the above equation, we consider linear filters only, and $\Lambda = \{\lambda^{(1)}(), \lambda^{(2)}(), ..., \lambda^{(K)}()\}$ is a set of potential functions on the features extracted by S. In practice, image intensities are discretized into a finite gray levels, and the filter responses are divided into a finite number of bins, therefore $\lambda^{(\alpha)}()$ is approximated by a piecewisely constant functions, i.e., a vector, which we denote by $\lambda^{(\alpha)}$, $\alpha = 1, 2, ..., K$.

The $\lambda^{(\alpha)}$s are computed in a nonparametric way so that the learned $p(\mathbf{I}; \Lambda, S)$ reproduces the observed statistics:

$$E_{p(\mathbf{I}; \Lambda, S)}\Big[H^{(\alpha)}\big(z; \mathbf{I}\big)\Big] = \mu_{obs}^{(\alpha)}(z) \quad \alpha = 1, 2, \ldots, K, \forall z.$$

Therefore as far as the selected features and their statistics are concerned, we cannot distinguish between $p(\mathbf{I}; \Lambda, S)$ and the "true" distribution $f(\mathbf{I})$.

Unfortunately, there is no simple way to express the $\lambda^{(\alpha)}$s in terms of the $\mu_{obs}^{(\alpha)}$s as in the two extreme examples. To compute $\lambda^{(\alpha)}$s, we adopted the Gibbs sampler [9], which simulates an inhomogeneous Markov chain in image space $\mathcal{L}^{|N^2|}$. This Monte Carlo method iteratively samples from the distribution $p(\mathbf{I}; \Lambda, S)$, followed by computing the histogram of the filter responses for this sample and updating the $\lambda^{(\alpha)}$ to bring the histograms closer to the observed ones. For a detailed account of the computation of $\lambda^{(\alpha)}$s, the readers are referred to [40], [41].

In our previous papers, the following two propositions are observed.

PROPOSITION 1. *Given a filter set S, and observed statistics* $\{\mu_{obs}^{(\alpha)}, \alpha = 1, 2, \ldots, K\}$, *there is an unique solution for* $\{\lambda^{(\alpha)}, \alpha = 1, 2, \ldots, K\}$.

PROPOSITION 2. $f(\mathbf{I})$ *is determined by its marginal distributions, thus* $p(\mathbf{I}) = f(\mathbf{I})$ *if it reproduces all the marginal distributions of linear filters.*

But for computational reasons, it is often desirable to choose a small set of filters which most efficiently capture the image structures. Given a set of filters S, and an ME distribution $p(\mathbf{I}; \Lambda, S)$, the goodness of $p(\mathbf{I}; \Lambda, S)$ is often measured by the Kullback-Leibler information distance between $p(\mathbf{I}; \Lambda, S)$ and the ideal distribution $f(\mathbf{I})$ [17],

$$KL\big(f(\mathbf{I}), p(\mathbf{I}; \Lambda, S)\big) = \int \cdot \int f(\mathbf{I}) \log \frac{f(\mathbf{I})}{p(\mathbf{I}; \Lambda, S)} d\mathbf{I} = E_f\big[\log f(\mathbf{I})\big] - E_f\big[\log p(\mathbf{I}; \Lambda, S)\big]$$

Then for a fixed model complexity K, the best feature set S^* is selected by the following criterion:

$$S^* = \arg\min_{|S|=K} KL\big(f(\mathbf{I}), p(\mathbf{I}; \Lambda, S)\big),$$

where S is chosen from a general filter bank B such as Gabor filters at multiple scales and orientations.

Enumerating all possible sets of features S in the filter bank and comparing their entropies is computationally too expensive. Instead, in [41] we propose a stepwise greedy procedure for minimizing the KL-distance. We start from $S = \varnothing$ and $p(\mathbf{I}; \Lambda, S)$ a uniform distribution, and introduce one filter at a time. Each added filter is chosen to maximally decrease the KL-distance, and we keep doing this until the decrease is smaller than a certain value.

In the experiments of this paper, we have used a simpler measure of the "information gain" achieved by adding one new filter to our feature set S. This is roughly an L^1-distance (vs. the L^2-measure implicit in the Kullback-Leibler distance), the readers are referred to [42] for a detailed account.

DEFINITION 3. *Given* $S = \{F^{(1)}, F^{(2)}, \ldots, F^{(K)}\}$ *and* $p(\mathbf{I}; \Lambda, S)$ *defined above, the information criterion* (*IC*) *for each filter* $F^{(\beta)} \in B / S$ *at step* $K + 1$ *is*:

$$IC = \frac{1}{2M}\sum_{n=1}^{M}\left\|H^{(\beta)}\big(z; \mathbf{I}_n^{obs}\big) - E_{p(\mathbf{I};\Lambda,S)}\Big[H^{(\beta)}\big(z; \mathbf{I}\big)\Big]\right\| - \frac{1}{2M}\sum_{n=1}^{M}\left\|H^{(\beta)}\big(z; \mathbf{I}_n^{obs}\big) - \mu_{obs}^{(\alpha)}(z)\right\|$$

We call the first term "average information gain" (*AIG*) *by choosing* $F^{(\beta)}$, *and the second term "average information fluctuation"* (*AIF*).

Intuitively, *AIG* measures the average error between the filter responses in the database and the marginal distribution of the current model $p(\mathbf{I}; \Lambda, S)$. In practice, we need to sample $p(\mathbf{I}; \Lambda, S)$, thus synthesize images $\{\mathbf{I}_n^{syn}, n = 1, 2, \ldots, M'\}$, and estimate $E_{p(\mathbf{I};\Lambda,S)}[H^{(\beta)}(z; \mathbf{I})]$ by $\mu_{syn}^{(\beta)} = \frac{1}{M'}\sum_{n=1}^{M'} H^{(\beta)}\big(z; \mathbf{I}_n^{syn}\big)$. For a filter $F^{(\beta)}$, the bigger *AIG* is, the more information $F^{(\beta)}$ captures as it reports the error between the current model and the observations. *AIF* is a measure of disagreement between the observed images. The bigger *AIF* is, the less their responses to $F^{(\alpha)}$ have in common.

3 EXPERIMENTS ON NATURAL IMAGES

This section presents experiments on learning prior models, and we start from exploring the statistical properties of natural images.

3.1 Statistic of Natural Images

It is well known that natural images have statistical properties distinguishing them from random noise images [28], [6], [24]. In our experiments, we collected a set of 44 natural images, six of which are shown in Fig. 2. These images are from various sources, some digitized from books and postcards, and some from a Corel image database. Our database includes both indoor and outdoor pictures, country and urban scenes, and all images are normalized to have intensity between zero and 31.

Fig. 2. Six out of the 44 collected natural images.

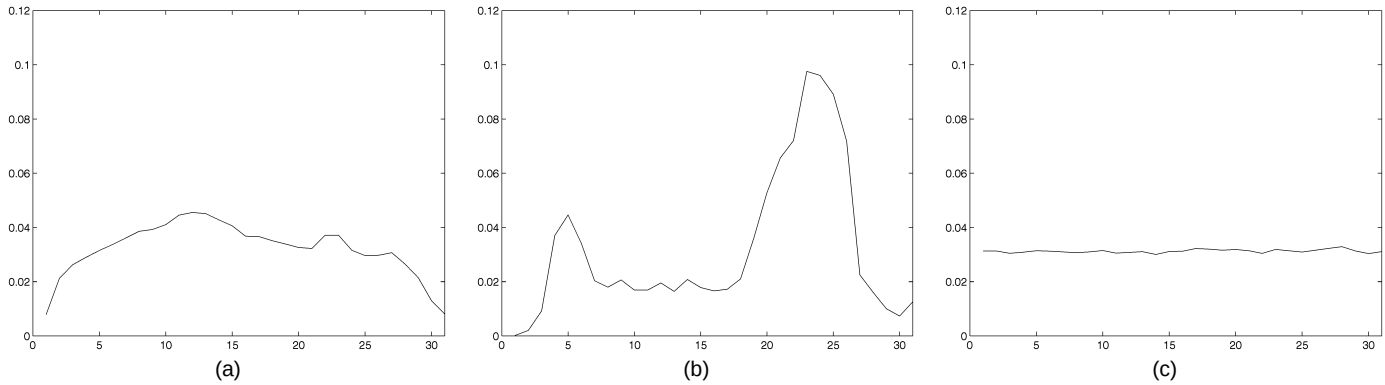


Fig. 3. The intensity histograms in domain [0, 31]. (a) Averaged over 44 natural images. (b) An individual natural image. (c) A uniform noise image.

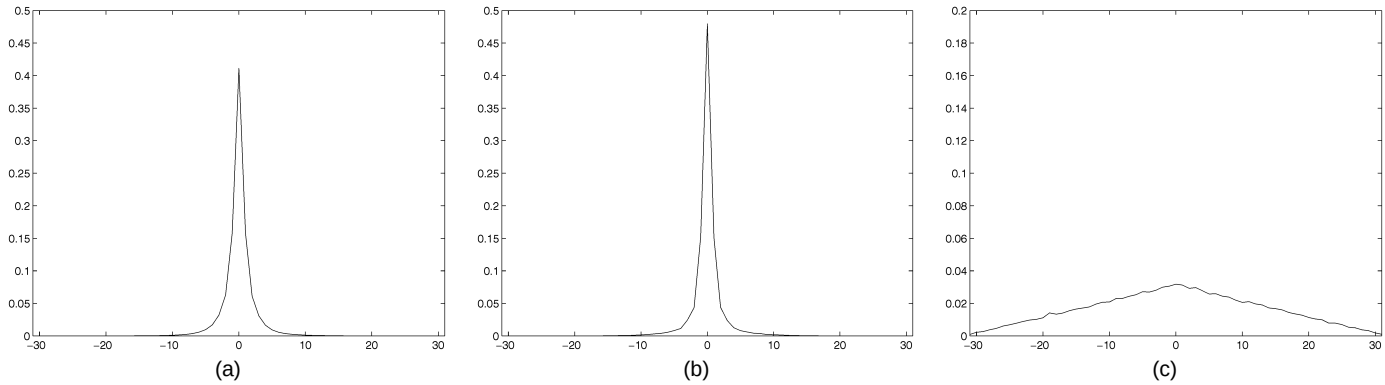


Fig. 4. The histograms of $\nabla_x \mathbf{I}$ plotted in domain [–30, 30]. (a) Averaged over 44 natural images. (b) An individual natural image. (c) A uniform noise image.

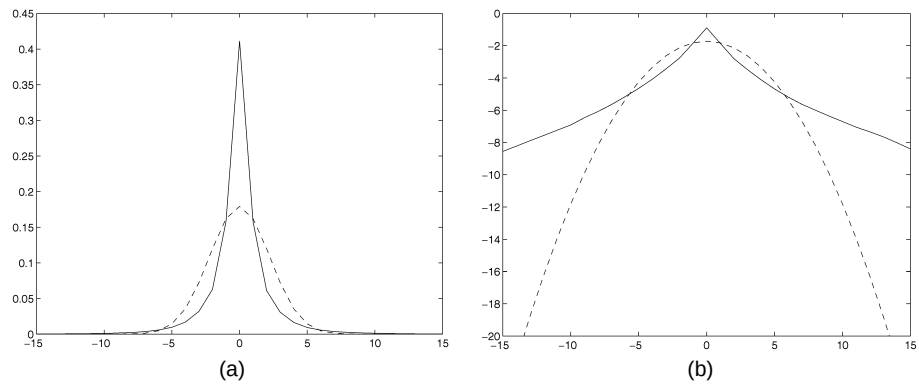


Fig. 5. (a) The histogram of $\nabla_x \mathbf{I}$ plotted against Gaussian curve (dashed) of same mean and variance in domain [–15, 15]. (b) The logarithm of the two curves in (a).

As stated in Proposition 2, marginal distributions of linear filters alone are capable of characterizing $f(\mathbf{I})$. In the rest of this paper, we shall only study the following bank B of linear filters.

1) An intensity filter $\delta()$.
2) Isotropic center-surround filters, i.e., the Laplacian of Gaussian filters.

$$LG(x, y, s) = const \cdot \left(x^2 + y^2 - s^2\right) e^{-\frac{x^2+y^2}{s^2}}, \qquad (10)$$

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

$$G(x, y, s, \theta) = const \circ Rot(\theta) \circ e^{-\frac{1}{2s^2}\left(4x^2+y^2\right)} e^{-i\frac{2\pi}{s}x} \qquad (11)$$

It is a sine wave at frequency $\frac{2\pi}{s}$ modulated by an

elongated Gaussian function, and rotated at angle θ. We denote the real and image parts of $G(x, y, s, \theta)$ by $Gcos(s, \theta)$ and $Gsin(s, \theta)$. Two special $Gsin(s, \theta)$ filters are the gradients ∇_x, ∇_y.

4) We will approximate large scale filters by filters of small window sizes on the high level of the image pyramid, where the image in one level is a "blown-down" version (i.e., averaged in 2×2 blocks) of the image below.

We observed three important aspects of the statistics of natural images.

First, for some features, the statistics of natural images vary widely from image to image. We look at the $\delta()$ filter as in Section 2.1. The average intensity histogram of the 44 images $\mu_{obs}^{(o)}$ is plotted in Fig. 3a, and Fig. 3b is the intensity histogram of an individual image (the temple image in Fig. 2). It appears that $\mu_{obs}^{(o)}(z)$ is close to a uniform distribution (Fig. 3c), while the difference between Fig. 3a and Fig. 3b is very big. Thus *IC* for filter $\delta()$ should be small (see Table 1).

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

The average histogram in Fig. 4a is very different from a Gaussian distribution. To see this, Fig. 5a plots it against a Gaussian curve (dashed) of the same mean and same variance. The histogram of natural images has higher kurtosis and heavier tails. Similar results are reported in [6]. To see the difference of the tails, Fig. 5b plots the logarithm of the two curves.

Third, the statistics of natural images are essentially scale invariant with respect to some features. As an example, we look at filters ∇_x and ∇_y. For each image $\mathbf{I}_n^{obs} \in NI^{obs}$, we build a pyramid with $\mathbf{I}_n^{[s]}$ being the image at the sth 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 the filter ∇_x, let $\mu_{x,s}(z)$ be the average histogram of $\nabla_x \mathbf{I}_n^{[s]}$, over $n = 1, 2, ..., 44$. Fig. 6a plots $\mu_{x,s}(z)$, for $s = 0, 1, 2$, and they are almost identical. To see the tails more clearly, we display $\log \mu_{x,s}(z)$, $s = 0, 1, 2$ in Fig. 6c. The differences between them are still small. Similar results are observed for $\mu_{y,s}(z)$, $s = 0, 1, 2$, the average histograms of $\nabla_y \mathbf{I}_n^{obs}$. In contrast, Fig. 6b plots the histograms of $\nabla_x \mathbf{I}^{[s]}$ with $\mathbf{I}^{[s]}$ being a uniform noise image at scales $s = 0, 1, 2$.

Combining the second and the 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 and across scales $s = 0, 1, 2$. The scale invariant property of natural images is largely caused by the following facts:

1) natural images contains objects of all sizes and
2) natural scenes are viewed and made into images at arbitrary distances.

3.2 Empirical Prior Models

In this section, we learn the prior models according to the theory proposed in Section 2 and analyze the efficiency of the filters quantitatively.

3.2.1 Experiment I

We start from $S = \varnothing$ and $p_0(\mathbf{I})$ a uniform distribution. We compute the *AIF*, *AIG*, and *IC* for all filters in our filter bank. We list the results for a small number of filters in Table 1. The filter Δ has the biggest *IC* (= 0.642), thus is chosen as $F^{(1)}$. An ME distribution $p_1(\mathbf{I}; \Lambda, S)$ is learned, and the information criterion for each filter is shown in the column headed $p_1(\mathbf{I})$ in Table 1. We notice that the *IC* for the filter Δ drops to near zero, and *IC* also drops for other filters because these filters are in general not independent of Δ. Some small filters like $LG(1)$ have smaller *IC*s than others, due to higher correlations between them and Δ.

The big filters with larger *IC* are investigated in Experiment II. In this experiment, we choose both ∇_x and ∇_y to be $F^{(2)}$, $F^{(3)}$ as in other prior models. Therefore, a prior model $p_3(\mathbf{I})$ is learned with potential:

$$U_3(\mathbf{I}; \Lambda, S) = \sum_{(x,y)} \lambda^{(1)}(\Delta\mathbf{I}(x,y)) + \lambda^{(2)}(\nabla_x \mathbf{I}(x,y)) + \lambda^{(3)}(\nabla_y \mathbf{I}(x,y)).$$

$\lambda^{(\alpha)}(z)$, $\alpha = 1, 2, 3$ are plotted in Fig. 7. Since $\mu_{obs}^{(1)}(z) = 0$ if $|z| \geq 9.5$,[4] and $\mu_{obs}^{(\alpha)}(z) = 0$ if $|z| \geq 22$ for $\alpha = 2, 3$, we only plot $\lambda^{(1)}(z)$ for $z \in [-9.5, 9.5]$ and $\lambda^{(2)}(z)$, $\lambda^{(3)}(z)$ for $z \in [-22, 22]$. These three curves are fitted with the functions $\psi_1(z) = 2.1(1 - 1/(1 + (|z|/4.8)^{1.32})$, $\psi_2(z) = 1.25(1 - 1/(1 + (|z|/2.8)^{1.5})$, $\psi_3(z) = 1.95(1 - 1/(1 + (|z|/2.8)^{1.5})$, respectively. A synthesized image sampled from $p_3(\mathbf{I})$ is displayed in Fig. 8.

So far, we have used three filters to characterize the statistics of natural images, and the synthesized image in Fig. 8 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 preserve intensity breaks, they only generate small speckles instead of big regions and long edges as one may expect. Based on this synthesized image, we compute the *AIG* and *IC* for all filters, and the results are listed in Table 1 in column $p_3(\mathbf{I})$.

4. In fact, $\mu_{obs}^{(1)}(\Delta\mathbf{I}) = 0$ if $\mu_{obs}^{(1)}(\Delta\mathbf{I}) < \frac{1}{N^2}$, with $N \times N$ being the size of synthesized image.

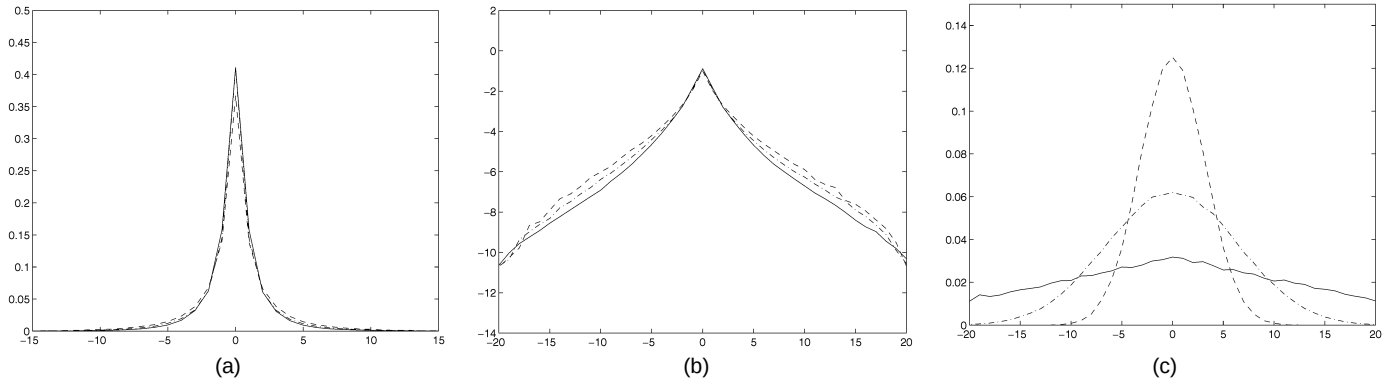


Fig. 6. (a) $\mu_{x,s}(z)$, $s = 0, 1, 2$. (b) $\log\mu_{x,s}(z)$, $s = 0$ (solid), $s = 1$ (dash-dotted), and $s = 2$ (dashed). (c) Histograms of a filtered uniform noise image at scales: $s = 0$ (solid curve), $s = 1$ (dash-dotted curve), and $s = 2$ (dashed curve).

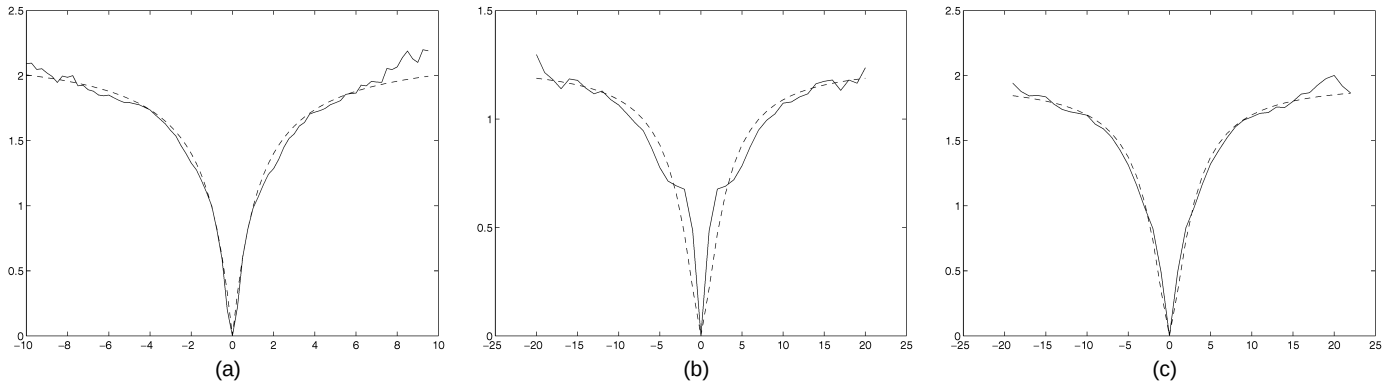


Fig. 7. The three learned potential functions for filters. (a) Δ. (b) ∇_x. (c) ∇_y. Dashed curves are the fitting functions: (a) $\psi_1(x) = 2.1(1-1) / \left(1+\left(|x| / 4.8\right)^{1.32}\right)$. (b) $\psi_2(x) = 1.25(1-1) / \left(1+\left(|x| / 2.8\right)^{1.5}\right)$. (c) $\psi_3(x) = 1.95(1-1) / \left(1+\left(|x| / 2.8\right)^{1.5}\right)$.

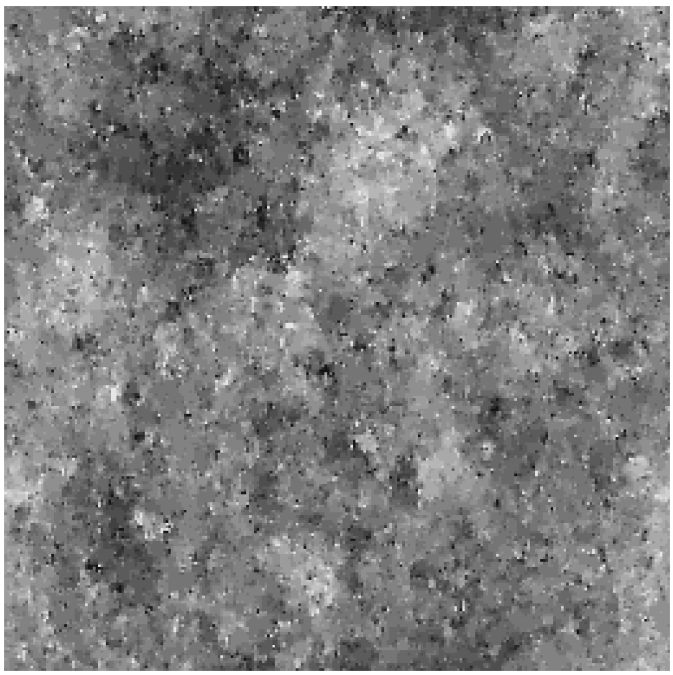


Fig. 8. A typical sample of p_3(I) (256 × 256 pixels).

3.2.2 Experiment II

It is clear that we need large-scale filters to do better. Rather than using the large scale Gabor filters, we chose to use ∇_x and ∇_y on four different scales and impose explicitly the scale invariant property that we find in natural images. Given an image $\mathbf{I}$ defined on an $N \times N$ lattice L, we build a pyramid in the same way as before. Let $\mathbf{I}^{[s]}$, $s = 0, 1, 2, 3$ be four layers of the pyramid. Let $H_{x,s}(z, x, y)$ denote the histogram of $\nabla_x\mathbf{I}^{[s]}(x, y)$ and $H_{y,s}(z, x, y)$ the histogram of $\nabla_y\mathbf{I}^{[s]}(x, y)$.

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

$$E_{p(\mathbf{I})}\left[H_{x,s}(z,x,y)\right] = \overline{\mu}(z), \quad \forall z \ \forall(x,y) \in L_s, \ s = 0,1,2,3$$

$$E_{p(\mathbf{I})}\left[H_{y,s}(z,x,y)\right] = \overline{\mu}(z), \quad \forall z \ \forall(x,y) \in L_s, \ s = 0,1,2,3$$

where L_s is the image lattice at level s, and $\overline{\mu}(z)$ is the average of the observed histograms of $\nabla_x\mathbf{I}^{[s]}$ and $\nabla_y\mathbf{I}^{[s]}$ on all 44 natural images at all scales. This results in a maximum entropy distribution $p_s(\mathbf{I})$ with energy of the following form,

$$U_s(\mathbf{I}) = \sum_{s=0}^{3}\sum_{(x,y)\in L_s}\lambda_{x,s}\left(\nabla_x\mathbf{I}^{[s]}(x,y)\right) + \lambda_{y,s}\left(\nabla_y\mathbf{I}^{[s]}(x,y)\right). \quad (12)$$

TABLE 1
THE INFORMATION CRITERION FOR FILTER SELECTION

Filter	Filter Size	AIF	$p_0(\mathcal{I})$		$p_1(\mathcal{I})$		$p_3(\mathcal{I})$		$p_s(\mathcal{I})$	
			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
$Gcos(2, 0°)$	5x5	0.119	0.716	0.597	0.239	0.119	0.181	0.062	0.197	0.078
$Gcos(2, 90°)$	5x5	0.155	0.673	0.518	0.238	0.083	0.188	0.034	0.189	0.035
$Gsin(2, 0°)$	5x5	0.125	0.573	0.447	0.344	0.219	0.170	0.045	0.141	0.015
$Gsin(2, 90°)$	5x5	0.126	0.666	0.540	0.241	0.115	0.156	0.030	0.154	0.028
$Gcos(4, 0°)$	7x7	0.133	0.569	0.436	0.321	0.187	0.203	0.070	0.164	0.031
$Gcos(4, 90°)$	7x7	0.144	0.545	0.401	0.304	0.160	0.214	0.070	0.183	0.039
$Gsin(4, 0°)$	7x7	0.124	0.555	0.431	0.334	0.209	0.186	0.063	0.149	0.025
$Gsin(4, 90°)$	7x7	0.131	0.535	0.405	0.322	0.191	0.196	0.065	0.163	0.033
$Gcos(6, 0°)$	11x11	0.125	0.384	0.259	0.340	0.216	0.220	0.095	0.145	0.020
$Gcos(6, 90°)$	11x11	0.129	0.398	0.269	0.340	0.211	0.229	0.100	0.165	0.035
$Gsin(6, 0°)$	11x11	0.123	0.398	0.275	0.360	0.237	0.211	0.089	0.139	0.016
$Gsin(6, 90°)$	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	$\frac{1}{M}$

Fig. 9 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 form displayed in Fig. 7 with low values around zero to encourage smoothness. As the learning proceeds, gradually $\lambda_{x,3}()$ turns "upside down" with smaller values at the two tails. Then $\lambda_{x,2}()$ and $\lambda_{x,1}()$ turn upside down one by one. Similar results are observed for $\lambda_{y,s}()$, $s = 0, 1, 2, 3$. Fig. 11 is a typical sample image from $p_s(\mathbf{I})$. To demonstrate it has scale invariant properties, in Fig. 10 we show the histograms $H_{x,s}$ and log $H_{x,s}$ of this synthesized image for $s = 0, 1, 2, 3$.

The learning process iterates for more than 10,000 sweeps. To verify the learned $\lambda()$s, we restarted a homogeneous Markov chain from a noise image using the learned model, and found that the Markov chain goes to a perceptually similar image after 6,000 sweeps.

3.2.2.1 Remark 1

In Fig. 9, we notice that $\lambda_{x,s}()$ are inverted, i.e., decreasing functions of $|z|$ for $s = 1, 2, 3$, distinguishing this model from other 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 $\lambda_{x,s}(z) = 0$ for $|z| > 31$, so $p_s(\mathbf{I})$ is still well-defined. Second, such inverted potentials have significant meaning in visual computation. In image restoration, when a high intensity difference $\nabla_x \mathbf{I}^{[s]}(x, y)$ is present, it is very likely to be noise if $s = 0$. However this is not true for $s = 1, 2, 3$. Additive noise can hardly pass to the high layers of the pyramid because at each layer the 2×2 averaging operator reduces the variance of the noise by four 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 $p_s(\mathbf{I})$, $\lambda_{x,0}()$ suppresses noise at the first layer, while $\lambda_{x,s}()$, $s = 1, 2, 3$ encourages sharp edges to form, and thus enhances blurred boundaries. We notice that regions of various scales emerge in Fig. 11, and the intensity contrasts are also higher at the boundary. These appearances are missing in Fig. 8.

3.2.2.2 Remark 2

Based on the image in Fig. 11, we computed *IC* and *AIG* for all filters and list them under column $p_s(\mathbf{I})$ in Table 1. We also compare the two extreme cases discussed in Section 2.1. For the $\delta()$ filter, *AIF* is very big, and *AIG* is only slightly bigger than *AIF*. Since all the prior models that we learned have no preference about the image intensity domain, 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*. For filter $\mathbf{I}^{(obsi)}$, $AIF = \frac{M-1}{M}$, i.e., the biggest among all filters, and $AIG \to 1$. In both cases, *IC*s are the two smallest.

4 GIBBS REACTION-DIFFUSION EQUATIONS

4.1 From Gibbs Distribution to Reaction-Diffusion Equations

The empirical results in the previous section suggest that the forms of the potentials $\lambda^{(\alpha)}(z)$ learned from images of real world scenes can be divided into two classes: upright curves $\lambda(z)$ for which $\lambda()$ is an even function increasing as $|z|$ increases and inverted curves for which the opposite happens. A similar phenomenon was observed in our learned texture models [40].

In Fig. 9, $\lambda_{x,s}(z)$ are fit to the family of functions (see the dashed curves),

$$\phi(\xi) = a\left(1 - \frac{1}{1 + \left(\left|\xi - \xi_0\right| / b\right)^{\gamma}}\right) \quad a > 0$$

$$\psi(\xi) = a\left(1 - \frac{1}{1 + \left(\left|\xi - \xi_0\right| / b\right)^{\gamma}}\right) \quad a < 0$$

ξ_0, b are, respectively, the translation and scaling constants, and $\|a\|$ weights the contribution of the filter.

In general, the Gibbs distribution learned from images in a given application has potential function of the following form,

$$U(\mathbf{I}; \Lambda, S) = \sum_{\alpha=1}^{n_d} \sum_{(x,y)} \phi^{(\alpha)}\left(F^{(\alpha)} * \mathbf{I}(x, y)\right) + \sum_{\alpha=n_d+1}^{K} \sum_{(x,y)} \psi^{(\alpha)}\left(F^{(\alpha)} * \mathbf{I}(x, y)\right) \quad (13)$$

Note that the filter set is now divided into two parts $S = S_d \cup S_r$, with $S_d = \{F^{(\alpha)}, \alpha = 1, 2, .., n_d\}$ and $S_r = \{F^{(\alpha)}, \alpha = n_d + 1, ..., K\}$. In most cases S_d consists of filters such as ∇_x, ∇_y, Δ which capture the general smoothness of images, and S_r contains filters which characterize the prominent features of a class of images, e.g., Gabor filters at various orientations and scales which respond to the larger edges and bars.

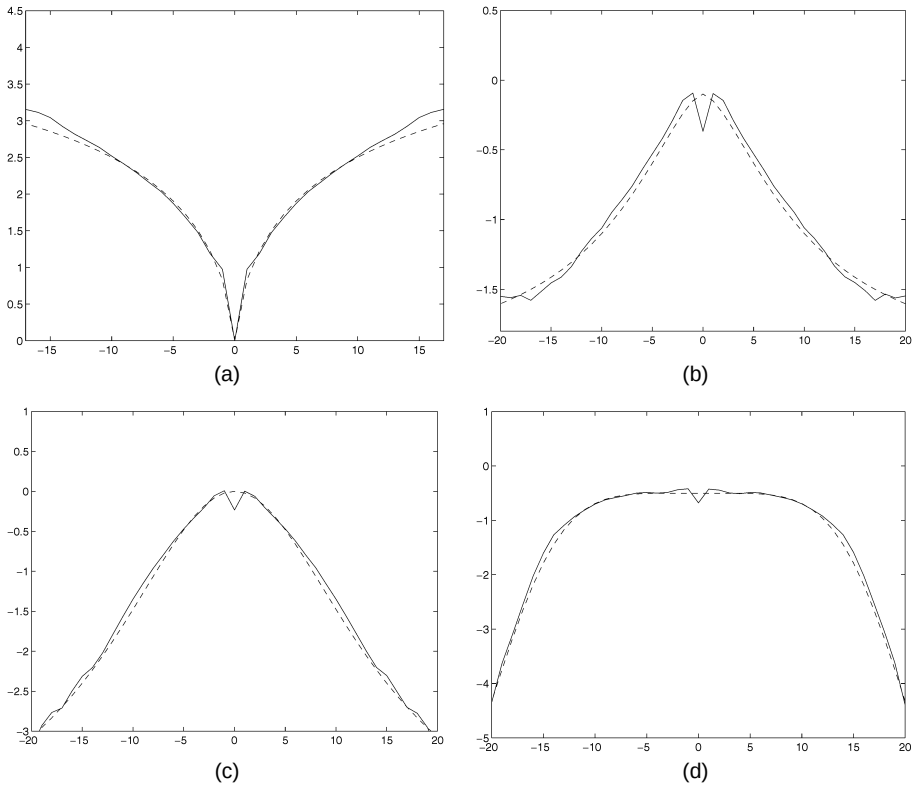


Fig. 9. Learned potential functions $\lambda_{\chi,s}()$, s = 0, 1, 2, 3. The dashed curves are fitting functions: $\phi(\xi) = a(1 - 1/(1 + (|\xi|/b)^{\gamma})$. (a) ($a$ = 5, b = 10, ξ_o = 0, γ = 0.7). (b) (a = –2.0, b = 10, ξ_o = 0, γ = 1.6). (c) (a = –4.8, b = 15, ξ_o = 0, γ = 2.0). (d) (a = –10.0, b = 22, ξ_o = 0, γ = 5.0).

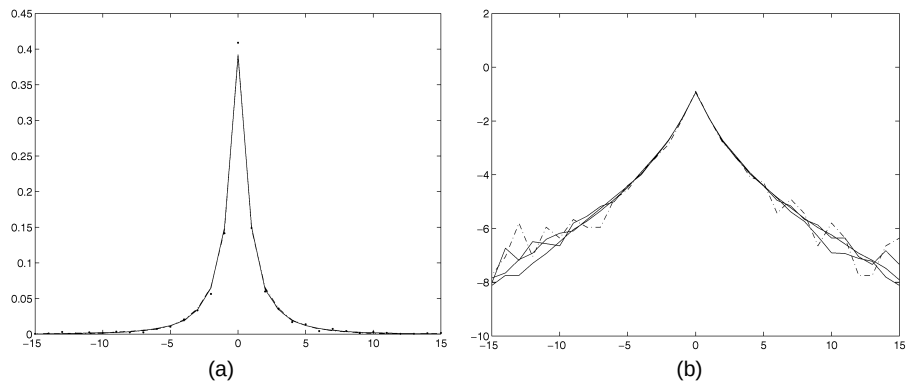


Fig. 10. (a) The histograms of the synthesized image at four scales—almost indistinguishable. (b) The logarithm of histograms in Fig. 10a.

Instead of defining a whole distribution with U, one can use U to set up a variational problem. In particular, one can attempt to minimize U by gradient descent. This leads to a non-linear parabolic partial differential equation:

$$\mathbf{I}_t = \sum_{\alpha=1}^{n_d} F_-^{(\alpha)} * \phi^{(\alpha)'}\left(F^{(\alpha)} * \mathbf{I}\right) + \sum_{\alpha=n_d+1}^{K} F_-^{(\alpha)} * \psi^{(\alpha)'}\left(F^{(\alpha)} * \mathbf{I}\right) \qquad (14)$$

with $F_-^{(\alpha)}(x, y) = -F_-^{(\alpha)}(-x, -y)$. Thus starting from an input image $\mathbf{I}(x, y, 0) = \mathbf{I}^{in}$, the first term diffuses the image by reducing the gradients, while the second term forms patterns as the reaction term. We call (14) the Grade.

Since the computation of (14) involves convolving twice for each of the selected filters, a conventional way for efficient computation is to build an image pyramid so that filters with large scales and low frequencies can be scaled down into small ones in the higher level of the image pyramid. This is appropriate especially when the filters are selected from a bank of multiple scales, such as the Gabor filters and the wavelet transforms. We adopt this representation in our experiments.

For an image $\mathbf{I}$, let $\mathbf{I}^{[s]}$ be an image at level s = 0, 1, 2, ... of a pyramid, and $\mathbf{I}[0] = \mathbf{I}$, the potential function becomes

$$U(\mathbf{I}; \Lambda, S) = \sum_s \sum_\alpha \sum_{(x,y)\in L_s} \phi_s^{(\alpha)}\left(F_s^{(\alpha)} * \mathbf{I}^{[s]}(x, y)\right) + \sum_s \sum_\alpha \sum_{(x,y)\in L_s} \psi_s^{(\alpha)}\left(F_s^{(\alpha)} * \mathbf{I}^{[s]}(x, y)\right)$$

We can derive the Grade equations similarly for this pyramidal representation.

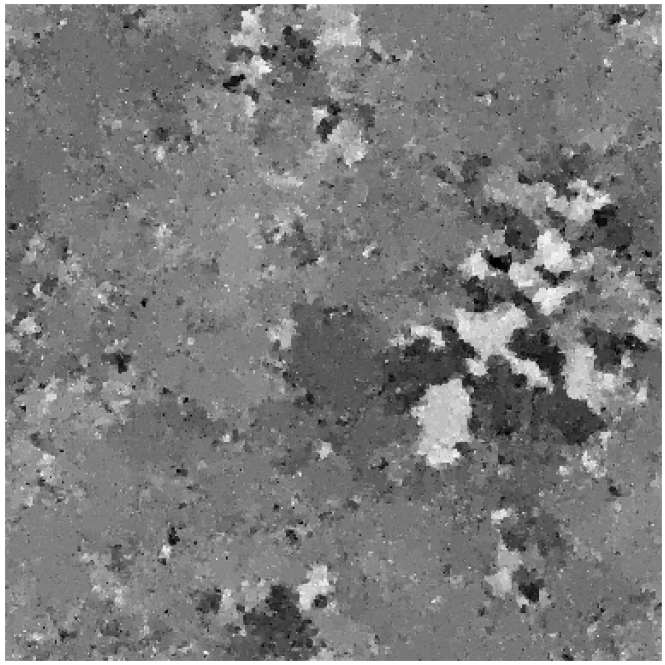


Fig. 11. A typical sample of $p_S(\mathbf{I})$ (384 × 384 pixels).

4.2 Anisotropic Diffusion and Gibbs Reaction-Diffusion

This section compares Grades with previous diffusion equations in vision.

In [25], [23], anisotropic diffusion equations for generating image scale spaces are introduced in the following form,

$$\mathbf{I}_t = \text{div}(c(x, y, t)\nabla\mathbf{I}), \qquad \mathbf{I}(x, y, 0) = \mathbf{I}^{in}, \tag{15}$$

where div is the divergence operator, i.e.,

$$\text{div}\left(\vec{V}\right) = \nabla_x P + \nabla_y Q$$

for $\vec{V} = \left(P, Q\right)$. Perona and Malik defined the *heat conductivity* $c(x, y, t)$ as functions of local gradients, for example:

$$\mathbf{I}_t = \nabla_x\left(\frac{1}{1+\left(\mathbf{I}_x / b\right)^2}\mathbf{I}_x\right) + \nabla_y\left(\frac{1}{1+\left(\mathbf{I}_y / b\right)^2}\mathbf{I}_y\right) \tag{16}$$

Equation (16) minimizes the energy function in a continuous form,

$$U(\mathbf{I}) = \int\int \lambda\left(\nabla_x\mathbf{I}(x, y)\right) + \lambda\left(\nabla_y\mathbf{I}(x, y)\right)dxdy$$

where $\lambda(\xi) = a\log(1 + (\xi/b)^2)$ and $\lambda'(\xi) = a\frac{\xi}{1+(\xi/b)^2}$ are plotted in Fig. 12. Similar forms of the energy functions are widely used as prior distributions [9], [4], [20], [11], and they can also be equivalently interpreted in the sense of robust statistics [13], [3].

In the following, we address three important properties of the Gibbs reaction-diffusion equations.

First, we note that (14) is an extension to (15) on a discrete lattice by defining a vector field,

$$\vec{V}(x, y) = \left(\phi^{(1)'}(\ ), \ldots, \phi^{(n_d)'}(\ ), \psi^{(n_d+1)'}(\ ), \ldots, \psi^{(K)'}(\ )\right)$$

and a divergence operator,

$$\text{div} = F_-^{(1)}* + F_-^{(2)}* + \cdots + F_-^{(K)}*.$$

Thus (14) can be written as,

$$\mathbf{I}_t = \text{div}\left(\vec{V}\right). \tag{17}$$

Compared to (15), which transfers the "heat" among adjacent pixels, (17) transfers the "heat" in many directions in a graph, and the conductivities are defined as functions of the local patterns not just the local gradients.

Second, in Fig. 13, $\phi(\xi)$ has round tip for $\gamma \geq 1$, and a cusp occurs at $\xi = 0$ for $0 < \gamma < 1$ which leaves $\phi'(\xi)$ undefined, i.e., $\phi'(\xi)$ can be any value in $(-\infty, \infty)$ as shown by the dotted curves in Fig. 13d. An interesting fact is that the potential function learned from real world images does have a cusp as shown in Fig. 9a, where the best fit is $\gamma = 0.7$. This cusp forms because a large part of objects in real world images have flat intensity appearances, and $\phi(\xi)$ with $\gamma < 1$ will produce piecewise constant regions with much stronger forces than $\gamma \geq 1$.

By continuity, $\phi'(\xi)$ can be assigned any value in the range $[-\omega, \omega]$ for $\xi \in [-\epsilon, \epsilon]$ and $\omega = c\frac{\epsilon^{\gamma-1}}{\left(1+(\epsilon / b)^{\gamma}\right)^2}$. In numerical simulations, for $\xi \in [-\omega, \omega]$, we take

$$\phi'(\xi) = \begin{cases} +\omega & \text{if } \sigma < -\omega \\ -\sigma & \text{if } \sigma \in [-\omega, \omega] \\ -\omega & \text{if } \sigma > \omega \end{cases}$$

where σ is the summation of the other terms in the differential equation whose values are well defined. Intuitively when $\gamma < 1$ and $\xi = (F^{(\alpha)} * \mathbf{I})(x, y) = 0$, $\phi^{(\alpha)'}(0)$ forms an attractive basin in its neighborhood $\mathcal{N}^{(\alpha)}(x, y)$ specified by the filter window of $F^{(\alpha)}$. For a pixel $(u, v) \in \mathcal{N}^{(\alpha)}(x, y)$, the depth of the attractive basin is $\left\|\omega F_-^{(\alpha)}(u - x, v - y)\right\|$. If a pixel is involved in multiple zero filter responses, it will accumulate the depth of the attractive basin generated by each filter. Thus unless the absolute value of the driving force from other well-defined terms is larger than the total depth of the attractive basin at (u, v), $\mathbf{I}(u, v)$ will stay unchanged. In the image restoration experiments in later sections, $\gamma < 1$ shows better performance in forming piecewise constant regions.

Third, the learned potential functions confirmed and improved the existing prior models and diffusion equations, but, more interestingly, reaction terms are first discovered, and they can produce patterns and enhance preferred features. We will demonstrate this property in the experiments below.

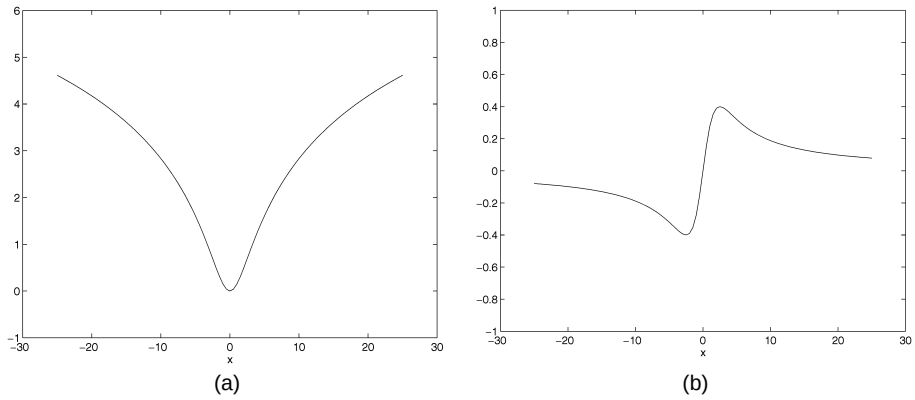


Fig. 12. (a) $\lambda(\xi) = a\log(1 + (\xi/b)^2)$. (b) $\lambda'(\xi) = a\frac{\xi}{1+(\xi/b)^2}$.

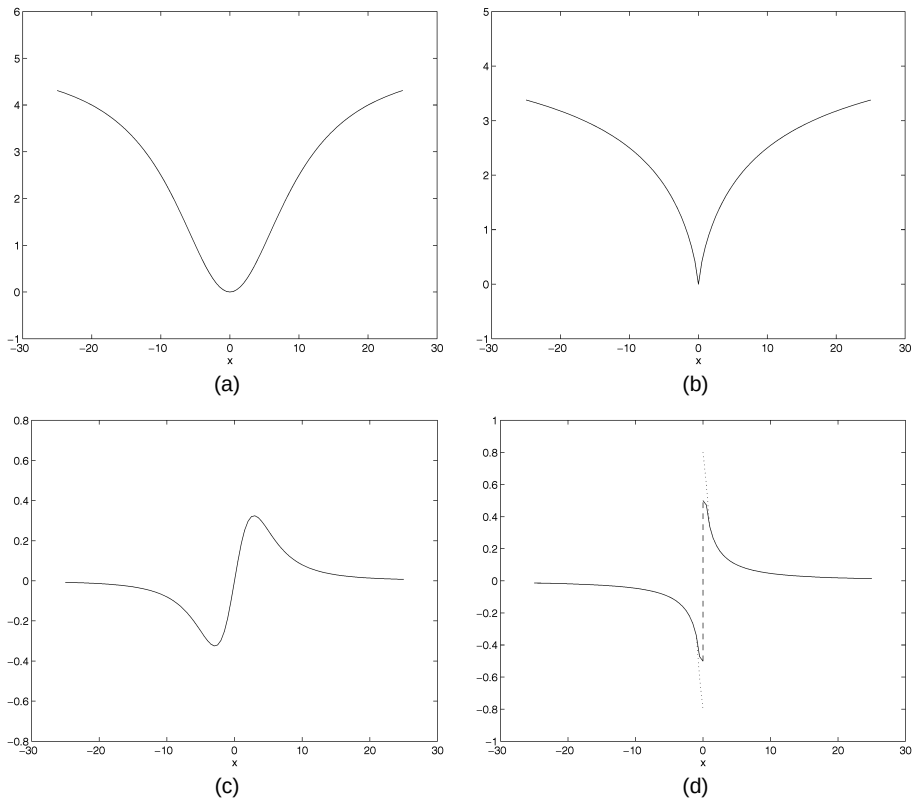


Fig. 13. The potential function $\phi(\xi) = -a\frac{1}{1+(\|\xi\|/b)^\gamma} + a, \phi'(\xi)$. (a) and (c) $\gamma = 2.0$. (b) and (d) $\gamma = 0.8$. (a) $\phi(\xi), \gamma \geq 1$. (b) $\phi(\xi), \gamma < 1$. (c) $\phi'(\xi), \gamma \geq 1$. (d). $\phi'(\xi), \gamma < 1$

4.3 Gibbs Reaction-Diffusion for Pattern Formation

In the literature, there are many nonlinear PDEs for pattern formation, of which the following two examples are interesting:

1) The Turing reaction-diffusion equation which models the chemical mechanism of animal coats [33], [21]. Two canonical patterns that the Turing equations can synthesize are leopard blobs and zebra stripes [34], [38]. These equations are also applied to image processing such as image halftoning [29], and a theoretical analysis can be found in [15].
2) The Swindale equation which simulates the development of the ocular dominance stripes in the visual cortex of cats and monkey [30]. The simulated patterns are very similar to the zebra stripes.

In this section, we show that these patterns can be easily generated with only two or three filters using the Grade. We run (14) starting with $\mathbf{I}(x, y, 0)$ as a uniform noise image, and Grade converges to a local minimum. Some synthesized texture patterns are displayed in Fig. 14.

For all six patterns in Fig. 14, we choose $F_0^{(1)} = \Delta$ the Laplacian of Gaussian filter at level zero of the image pyramid as the only diffusion filter, and we fix $a = 5$, $b = 10$, $\xi_o = 0$, $\gamma = 1.2$ for $\phi_0^{(1)}(\xi)$. For the three patterns in Fig. 14a, Fig. 14b, and Fig. 14c, we choose isotropic center-surround

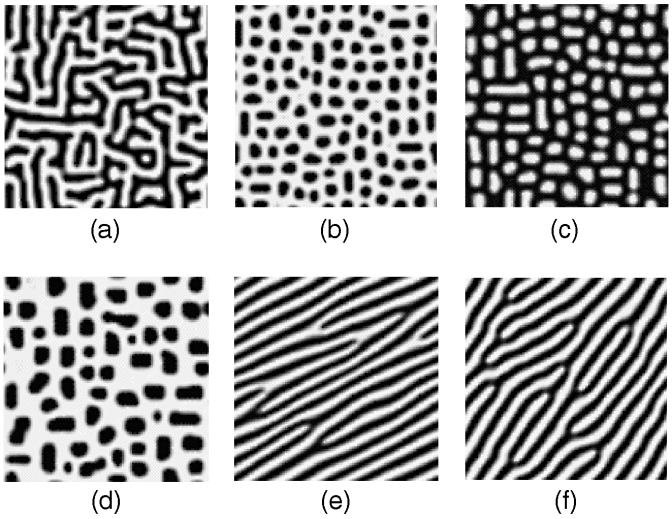


Fig. 14. Leopard blobs and zebra stripes synthesized by Grades.

filter $LG(\sqrt{2})$ of widow size 7×7 pixels as the reaction filter $F_1^{(2)}$ at level one of the image pyramid, and we set (a = –6.0, b = 10, γ = 2.0) for $\psi_1^{(2)}(\xi)$. The differences between these three patterns are caused by ξ_o in $\psi_1^{(2)}(\xi)$. $\xi_o = 0$ forms the patterns with symmetric appearances for both black and white parts as shown in Fig. 14a. As ξ_o becomes negative, black blobs begin to form as shown in Fig. 14b, where $\xi_o = -6$, and positive ξ_o generates white blobs in the black background as shown in Fig. 14c, where $\xi_o = 6$. The general smoothness appearance of the images is attributed to the diffusion filter. Fig. 14d is generated with two reaction filters: $LG(\sqrt{2})$ at level one and level two, respectively, therefore the Grade creates blobs of mixed sizes. Similarly we selected one cosine Gabor filter $Gcos(4, 30^o)$ (size 7×7 pixels oriented at 30^o) at level one as the reaction filter $F_1^{(2)}$ for Fig. 14e where (a = –3.5, b = 10, γ = 2.0, ξ_o = 0) for $\psi_1^{(2)}(\xi)$. Fig. 14f is generated with two reaction filters $Gcos(4, 30^o)$, $Gcos(4, 60^o)$ at level one, where (a = –3.5, b = 10, γ = 2.0, ξ_o = –3) for $\psi_1^{(2)}(\xi)$ and $\psi_1^{(3)}(\xi)$.

It seems that the leopard blobs and zebra stripes are among the most canonical patterns which can be generated with easy choices of filters and parameters. As shown in [40], the Gibbs distribution are capable of modeling a large variety of texture patterns, but filters and different forms for $\psi(\xi)$ have to be learned for a given texture pattern.

5 Image Enhancement and Clutter Removal

So far we have studied the use of a single energy function $U(\mathbf{I})$ either as the log likelihood of a probability distribution at $\mathbf{I}$ or as a function of $\mathbf{I}$ to be minimized by gradient descent. In image processing, we often need to model both the underlying images and some distortions, and to maximize a *posterior* distribution. Suppose the distortions are additive, i.e., an input image is,

$$\mathbf{I}^{\text{in}} = \mathbf{I} + \mathbf{C}.$$

In many applications, the distortion images $\mathbf{C}$ are often not i.i.d. Gaussian noise, but clutter with structures such as trees in front of a building or a military target. Such clutter will be very hard to handle by edge detection and image segmentation algorithms.

We propose to model clutter by an extra Gibbs distribution, which can be learned from some training images by the minimax entropy theory as we did for the underlying image $\mathbf{I}$. Thus an extra pyramidal representation for $\mathbf{I}^{\text{in}} - \mathbf{I}$ is needed in a Gibbs distribution form as shown in Fig. 15. The resulting posterior distributions are still of the Gibbs form with potential function,

$$U^*(\mathbf{I}) = U_{\mathbf{C}}(\mathbf{I}^{\text{in}} - \mathbf{I}; \Lambda_{\mathbf{C}}, S_{\mathbf{C}}) + U(\mathbf{I}; \Lambda, S), \tag{18}$$

where $U_{\mathbf{C}}()$ is the potential of the clutter distribution.

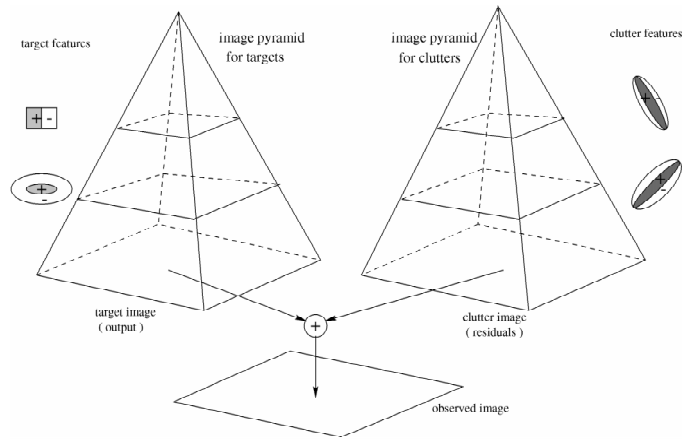


Fig. 15. The computational scheme for removing noise and clutter.

Thus the MAP estimate of $\mathbf{I}$ is the minimum of U^*. In the experiments which follow, we use the Langevin equation for minimization, a variant of simulated annealing:

$$d\mathbf{I}_t = -\nabla U^*(\mathbf{I})dt + \sqrt{2T(t)}dw_t \tag{19}$$

where $w(x, y, t)$ is the standard Brownian motion process, i.e.,

$$w(x, y, t) \sim N(\mu(x, y), |t|), \quad dw_t = \sqrt{dt}N(0, 1).$$

$T(t)$ is the "temperature" which controls the magnitude of the random fluctuation. Under mild conditions on U^*, (19) approaches a global minimum of U^* at a low temperature. The analyses of convergence of the equations can be found in [14], [10], [8]. The computational load for the annealing process is notorious, but, for applications like denoising, a fast decrease of temperature may not affect the final result very much.

5.1 Experiment I

In the first experiment, we take $U_{\mathbf{C}}$ to be quadratic, i.e., $\mathbf{C}$ to be an i.i.d. Gaussian noise image. We first compare the performance of the three prior models $p_l(\mathbf{I})$, $p_t(\mathbf{I})$, and $p_s(\mathbf{I})$ whose potential functions are, respectively:

$$U_l(\mathbf{I}) = \psi_l(\nabla_x \mathbf{I}) + \psi_l(\nabla_y \mathbf{I}), \quad \psi_l(\xi) = a\min(\theta^2, \xi^2) \quad (20)$$

$$U_t(\mathbf{I}) = \psi_t(\nabla_x \mathbf{I}) + \psi_t(\nabla_y \mathbf{I}), \quad \psi_t(\xi) = a\xi^2/(1 + c\xi^2) \quad (21)$$

$$U_s(\mathbf{I}) = \text{the four-scale energy in (12)} \quad (22)$$

$\psi_l()$ and $\psi_t()$ are the line-process and T-function displayed in Fig. 1b and Fig. 1c, respectively.

Fig. 16 demonstrates the results: The original image is the lobster boat displayed in Fig. 2. It is normalized to have intensity in [0, 31] and Gaussian noise from $N(0, 25)$ are added. The distorted image is displayed in Fig. 16a, 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 Fig. 16b, Fig. 16c, and Fig. 16d, respectively. $p_s(\mathbf{I})$, which is the only model with a reaction term, appears to have the best effect in recovering the boat, especially the top of the boat, but it also enhances the water.

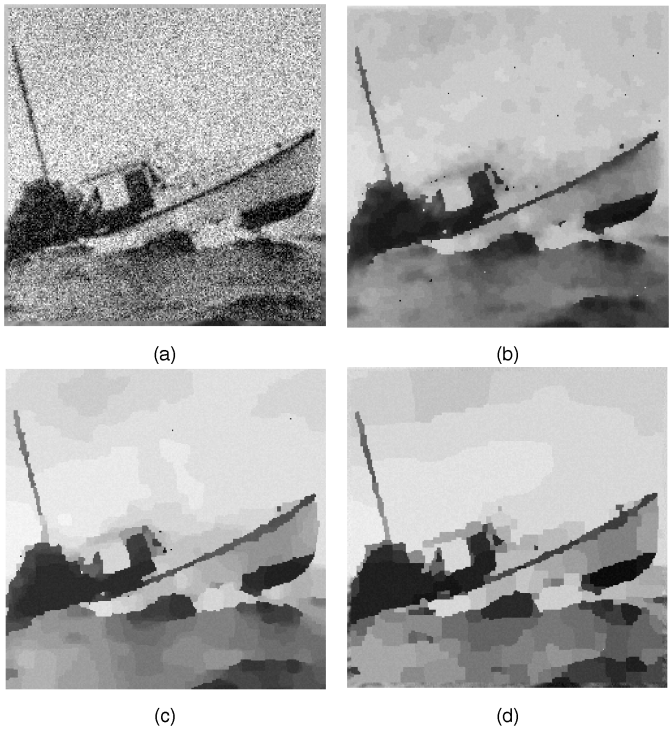


Fig. 16. (a) The noise distorted image. (b)-(d) Restored images by prior models $p_l(I)$, $p_t(I)$, and $p_s(I)$, respectively.

5.2 Experiment II

In many applications, i.i.d. Gaussian models for distortions are not sufficient. For example, in Fig. 17a, the tree branches in the foreground will make image segmentation and object recognition extremely difficult, because they cause strong edges across the image. Modeling such clutter is a challenging problem in many applications. In this paper, we only consider clutter as two-dimensional pattern, despite its geometry and 3D structure.

We collected a set of images of buildings and a set of images of trees all against clean background—the sky. For the tree images, we translate the image intensities to [–31, 0], i.e., zero for sky. In this case, since the trees are always darker than the building, thus the negative intensity will approximately take care of the occlusion effects. We learn the Gibbs distributions for each set respectively in the pyramid, then such models are respectively adopted as the prior distribution and the likelihood as in (18). We recovered the underlying images by maximizing a posteriori distribution using the stochastic process.

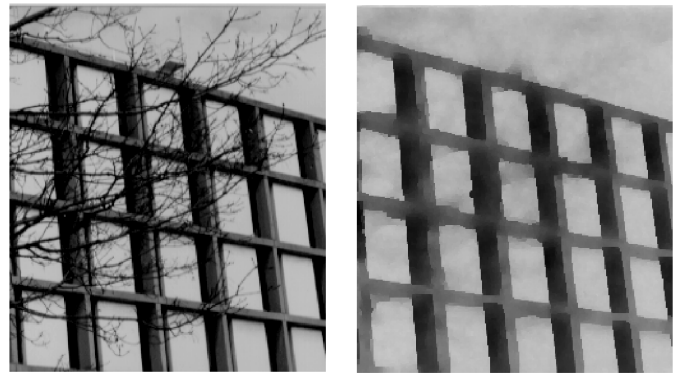


Fig. 17. (a) The observed image. (b) The restored image using six filters.

For example, Fig. 17b is computed using six filters with two filters for $\mathbf{I}$: $\{\nabla_{x,0}, \nabla_{y,0}\}$, and four filters for $\mathbf{I}_\mathbf{C}$: $\{\delta, \nabla_x, \nabla_y, Gcos(2, 30^o)\}$, i.e., the potential for $\mathbf{I}_\mathbf{C}$ is:

$$U_\mathbf{C}(\mathbf{I}) = \sum_{(x,y)} \phi\left(\nabla_x \mathbf{I}(x,y)\right) + \phi\left(\nabla_y \mathbf{I}(x,y)\right) + \phi^*\left(\mathbf{I}(x,y)\right) + \psi^*\left(G\cos * \mathbf{I}(x,y)\right)$$

In the above equation, $\phi^*(\xi)$ and $\psi^*(\xi)$ are fit to the potential functions learned from the set of tree images:

$$\phi^*(\xi) = \begin{cases} a_1\left(1 - \dfrac{1}{1 + \left(\left|\xi - \xi_o\right| / b\right)^\gamma}\right) & \xi < \xi_o \\ a_2\left(1 - \dfrac{1}{1 + \left(\left|\xi - \xi_o\right| / b\right)^\gamma}\right) & \xi \geq \xi_o, \ a_2 > a_1 > 0 \end{cases}$$

So, the energy term $\phi^*(\mathbf{I}(x, y))$ forces zero intensity for the clutter image while allowing for large negative intensities for the dark tree branches.

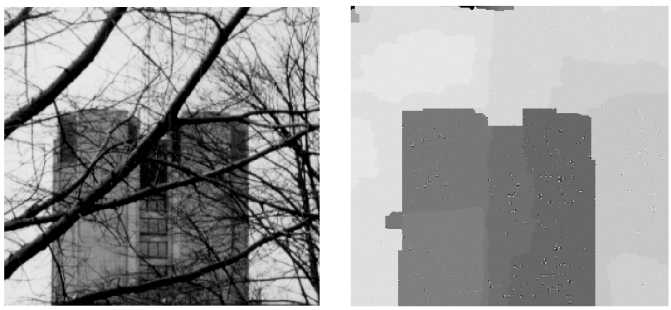
$$\psi^*(\xi) = \begin{cases} a\left(1 - \dfrac{1}{1 + \left(\left|\xi - \xi_o\right| / b\right)^\gamma}\right) & \xi < \xi_o, \quad a > 0 \\ 0 \end{cases}$$

Fig. 18b is computed using eight filters with four filters for $\mathbf{I}$ and four filters for $\mathbf{I}_\mathbf{C}$. Thirteen filters are used for Fig. 19 where the clutter is much heavier.

As a comparison, we run the anisotropic diffusion process [25] on Fig. 19a, and images at iterations $t = 50, 100, 300$ are displayed in Fig. 20. As we can see that as $t \to \infty$, $\mathbf{I}(t)$ becomes a flat image. A robust anisotropic diffusion equation is recently reported in [2].

(a) (b)

Fig. 18. (a) An observed image. (b) The restored image using eight filters.



(a) (b)

Fig. 19. (a) The observed image. (b) The restored image using 13 filters.

(a) (b) (c)

Fig. 20. Images by anisotropic diffusion at iteration (a) $t = 50$, (b) $t = 100$, and (c) $t = 300$ for comparison.

6 Conclusion

In this paper, we studied the statistics of natural images, based on which a novel theory is proposed for learning the generic prior model—the universal statistics of real world scenes. We argue that the same strategy developed in this paper can be used in other applications. For example, learning probability models for MRI images and 3D depth maps.

The learned prior models demonstrate some important properties such as the "inverted" potentials terms for patterns formation and image enhancement. The expressive power of the learned Gibbs distributions allow us to model structured noise–clutter in natural scenes. Furthermore, our prior learning method provides a novel framework for designing reaction-diffusion equations based on the observed images in a given application, without modeling the physical or chemical processes as people did before [33].

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, and this homogeneity assumption is unrealistic. We call the generic prior models studied in this paper *the first-level prior*. A more sophisticated prior model should incorporate concepts like object geometry, and we call such prior models *second-level priors*. Diffusion equations derived from this second-level prior are studied in image segmentation [39], and in scale space of shapes [16]. A discussion of some typical diffusion equations is given in [22]. It is our hope that this article will stimulate further investigations on building more realistic prior models as well as sophisticated PDEs for visual computation.

Acknowledgment

This work was started when the authors were at Harvard University. This research was supported by a U.S. National Science Foundation grant and a grant from ARO. We thank Y.N. Wu and S. Geman for valuable discussion.

References

[1] A. Berger, V. Della Pietra, and S. Della Pietra, "A Maximum Entropy Approach to Natural Language Processing," *Computational Linguistics*, vol. 22, no. 1, 1996.

[2] M. Black, G. Sapiro, D. Marimont, and D. Heeger, "Robust Anisotropic Diffusion," *IEEE Trans. Image Processing*, to appear.

[3] M.J. Black and A. Rangarajan, "On the Unification of Line Processes, Outlier Rejection, and Robust Statistics With Applications in Early Vision," *Int'l J. Computer Vision*, vol. 19, no. 1, 1996.

[4] A. Blake and A. Zisserman, *Visual Reconstruction*. Cambridge, Mass.: MIT Press, 1987.

[5] J. Daugman, "Uncertainty Relation for Resolution in Space, Spatial Frequency, and Orientation Optimized by Two-Dimensional Visual Cortical Filters," *J. Optical Soc. America*, vol. 2, no. 7, pp. 1,160-1,169, 1985.

[6] D.J. Field, "Relations Between the Statistics of Natural Images and the Response Properties of Cortical Cells," *J. Optical Soc. America, A*, vol. 4, no. 12, 1987.

[7] D. Gabor, "Theory of Communication," *IEE Proc*., vol. 93, no. 26, 1946.

[8] S.B. Gelfand and S.K. Mitter, "On Sampling Methods and Annealing Algorithms," *Markov Random Fields—Theory and Applications*. New York: Academic Press, 1993.

[9] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 6, no. 7, pp. 721-741, July 1984.

[10] S. Geman and C. Hwang, "Diffusion for Global Optimization," *SIAM J. Control and Optimization*, vol. 24, no. 5, 1986.

[11] D. Geman and G. Reynoids, "Constrained Restoration and the Recover of Discontinuities," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, pp. 367-383, 1992.

[12] D. Geiger and F. Girosi, "Parallel and Deterministic Algorithms for MRFs: Surface Reconstruction," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 5, pp. 401-412, May 1991.

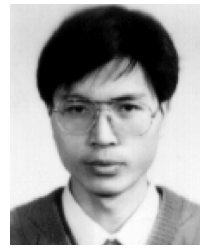
[13] D. Geiger and A.L. Yuille, "A Common Framework for Image Segmentation," *Int'l J. Computer Vision*, vol. 6, no. 3, pp. 227-243, 1991.

[14] B. Gidas, "A Renormalization Group Approach to Image Processing Problems," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 2, Feb. 1989.

[15] P. Grindrod, *The Theory and Applications of Reaction-Diffusion Equations*. New York: Oxford Univ. Press, 1996.

[16] B. Kimia, A. Tannebaum, and S. Zucker, "Shapes, Shocks, and Deformations I: The Components of Two-Dimensional Shape and

the Reaction-Diffusion Space," *Int'l J. Computer Vision*, vol. 15, pp. 189-224, 1995.
[17] S. Kullback and R.A. Leibler, "On Information and Sufficiency," *Annual Math. Stat.*, vol. 22, pp. 79-86, 1951.
[18] J. Marroguin, S. Mitter, and T. Poggio, "Probabilistic Solution of Ill-Posed Problems in Computational Vision," *J. Am. Statistical Assoc.*, vol. 82, no. 397, 1987.
[19] P. Meer, D. Mintz, D.Y. Kim, and A. Rosenfeld, "Robust Regression Methods for Computer Vision: A Review," *Int'l J. Computer Vision*, vol. 6, no. 1, 1991.
[20] D. Mumford and J. Shah, "Optimal Approximations by Piecewise Smooth Functions and Associated Variational Problems," *Comm. Pure Applied Math.*, vol. 42, pp. 577-684, 1989.
[21] J.D. Murray, "A Pre-Pattern Formation Mechanism for Mammalian Coat Markings," *J. Theoretical Biology*, vol. 88, pp. 161-199, 1981.
[22] W. Niessen, B. Romeny, L. Florack, and M. Viergever, "A General Framework for Geometry-Driven Evolution Equations," *Int'l J. Computer Vision*, vol. 21, no. 3, pp. 187-205, 1997.
[23] M. Nitzberg and T. Shiota, "Nonlinear Image Filtering With Edge and Corner Enhancement," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, no. 8, pp. 826-833, Aug. 1992.
[24] B.A. Olshausen and D.J. Field, "Natural Image Statistics and Efficient Coding," *Proc. Workshop on Information Theory and the Brain*, Sept. 1995.
[25] P. Perona and J. Malik, "Scale-Space and Edge Detection Using Anisotropic Diffusion," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629-639, July 1990.
[26] T. Poggio, V. Torre, and C. Koch, "Computational Vision and Regularization Theory," *Nature*, vol. 317, pp. 314-319, 1985.
[27] T. Poggio and F. Girosi, "Networks for Approximation and Learning," *Proc. IEEE*, vol. 78, pp. 1,481-1,497, 1990.
[28] D.L. Ruderman and W. Bialek, "Statistics of Natural Images: Scaling in the Woods," *Phys. Rev. Letter*, vol. 73, pp. 814-817, 1994.
[29] A. Sherstinsky and R. Picard, "M-Lattice: From Morphogenesis to Image Processing," *IEEE Trans. Image Processing*, vol. 5, no. 7, July 1996.
[30] N.V. Swindale, "A Model for the Formation of Ocular Dominance Stripes," *Proc. Royal Soc. London* B, vol. 208, pp. 243-264, 1980.
[31] D. Terzopoulos, "Multilevel Computational Processes for Visual Surface Reconstruction," *Computer Vision, Graphics, and Image Processing*, vol. 24, pp. 52-96, 1983.
[32] A.N. Tikhonov and V.Y. Arsenin, *Solutions of Ill-Posed Problems*. V.H. Winston & Sons, 1977.
[33] A.M. Turing, "The Chemical Basis of Morphogenesis," *Philosophy Trans. Royal Soc. London*, vol. 237, no. B, pp. 37-72, 1952.
[34] G. Turk, "Generating Textures on Arbitrary Surfaces Using Reaction-Diffusion," *Computer Graphics*, vol. 25, no. 4, 1991.
[35] A.B. Watson, "Efficiency of Model Human Image Code," *J. Optical Soc. America A*, vol. 4, no. 12, 1987.
[36] K. Wilson, "The Renormalization Group: Critical Phenomena and the Knodo Problem," *Rev. Mod. Phys.*, vol. 47, pp. 773-840, 1975.
[37] G. Winkler, *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*. New York: Springer-Verlag, 1995.
[38] A. Witkin and M. Kass, "Reaction-Diffusion Textures," *Computer Graphics*, vol. 25, no. 4, 1991.
[39] S.C. Zhu and A.L. Yuille, "Region Competition: Unifying Snakes, Region Growing, and Bayes/MDL for Multi-Band Image Segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 9, pp. 884-900, Sept. 1996.
[40] 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. CVPR*, 1996.
[41] S.C. Zhu, Y.N. Wu, and D.B. Mumford, "Minimax Entropy Principle and Its Application to Texture Modeling," *Neural Computation*, vol. 9, no. 8, Nov. 1997.
[42] S.C. Zhu and D.B. Mumford, "Learning Generic Prior Models for Visual Computation," Harvard Robotics Lab, Technical Report TR-96-05 (a short version appeared in *CVPR97*).



Song Chun Zhu received his BS degree in computer science from the University of Science and Technology of China in 1991. He received his MS and PhD degrees in computer science from Harvard University in 1994 and 1996, respectively. From 1996 to 1997, he worked in the Division of Applied Math at Brown University, and he joined the Computer Science Department at Stanford University in 1997. His research is concentrated in the areas of computer and human vision, statistical modeling, and pattern theory.

David Mumford received his AB degree in mathematics from Harvard University in 1957 and his PhD degree also in mathematics from Harvard in 1961. He continued at Harvard as instructor, 1961; associate professor, 1963; and professor in 1967. He was chairman of the department from 1981 to 1984 and has held visiting appointments at the Institute for Advanced Study (Princeton), Warwick University, the Tata Institute of Fundamental Science (Bombay), the Institut des Hautes Etudes Scientifiques, and Isaac Newton Institute of Mathematical Sciences (Cambridge). He was Higgins Professor in Mathematics until his retirement in 1997. He was appointed University Professor at Brown in 1996. He received the Fields Medal in 1974 and DSc(hon) at Warwick in 1983 and is a member of the National Academy of Science. He is president of the International Mathematical Union (1995-1998). Professor Mumford worked in the area of algebraic geometry up to 1983. Since then he has been studying the application of mathematical ideas to the modeling of intelligence, both theoretically and in animals.