IEEE Trans. on PAMI (vol.19, no.11, Nov. 1997)

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

Song Chun Zhu is now with the Computer Science Department, Stanford University, Stanford, CA 94305, and David Mumford is with the Division of Applied Mathematics, Brown University, Providence, RI 02912. This work started when the authors were at Harvard University.

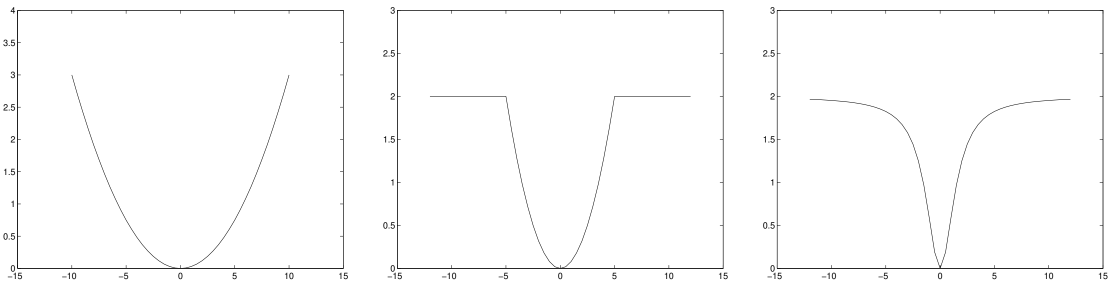
1 Introduction and motivation

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]:

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

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 figure 1. The functions in figure 1b, 1c have flat tails to preserve edges and object boundaries, and thus they are said to have advantages over the quadratic function in figure 1a.



a b c

Figure 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: $\psi(\xi) = a(1 - \frac{1}{1+c\xi^2})$.

These prior models have been motivated by regularization theory [26, 18],[1] phys-

[1]Where the smoothness term is explained as a stabilizer for solving "ill-posed" problems [32].

ical 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],

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

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

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

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

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 equation (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 figure 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 is 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 $NI^{obs} = \{\mathbf{I}_n^{obs}, \quad n = 1, 2.., M\}$ 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 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 $\{\mathbf{I}_n^{obs}, \quad n = 1, 2.., M\}$, 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,

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

where $\forall z \in \mathbf{R}$ and $\delta()$ is a Dirac function with point mass concentrated at 0.

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,*

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

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

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

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 $\{\mathbf{I}_n^{obs}, \ n = 1, 2, ...M\}$, immediately we have two simple solutions. The first one is,

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

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 equation (4) as,

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

The second solution is:

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

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

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

where $<, >$ 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.

1. *The potentials $\psi_1()$, $\psi_2()$ are built on the responses of linear filters.* In equation (7), $\mathbf{I}_n^{obs}$, $n = 1, 2..., M$ 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 $E_p[H^{(\alpha)}(z; \mathbf{I})] = \mu_{obs}^{(\alpha)}(z)$, thus as M increases, $E_p[H^{(\alpha)}(z; \mathbf{I})] \to E_f[H^{(\alpha)}(z; \mathbf{I})]$.

This second property is in general not satisfied by existing prior models in equation (1). $p(\mathbf{I})$ in both cases meets our objective for prior learning, but intuitively they are not good choices. In equation (5), the $\delta()$ filter does not capture spatial structures of larger than one pixels, and in equation (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 at section (3.2).

[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.

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 $\{\mu_{obs}^{(\alpha)}(z), \alpha = 1, 2, ..., K\}$, a maximum entropy distribution is derived which has the following Gibbs form:

$$\begin{aligned} p(\mathbf{I}; \Lambda, S) &= \frac{1}{Z} e^{-U(\mathbf{I}; \Lambda, S)}, && (8) \\ U(\mathbf{I}; \Lambda, S) &= \sum_{\alpha=1}^{K} \sum_{(x,y)} \lambda^{(\alpha)}((F^{(\alpha)} * \mathbf{I})(x, y)). && (9) \end{aligned}$$

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 non-parametric way so that the learned $p(\mathbf{I}; \Lambda, S)$ reproduces the observed statistics:

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

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 $\{\mu_{obs}^{(\alpha)}, \alpha = 1, 2, ..., K\}$, there is an unique solution for $\{\lambda^{(\alpha)}, \alpha = 1, 2, ..., K\}$.*

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],

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

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

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

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 computational too expensive. Instead, in [41] we propose a step-wise greedy procedure for minimizing the KL-distance. We start from $S = \emptyset$ 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* $S = \{F^{(1)}, F^{(2)}, ..., F^{(K)}\}$ *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,*

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

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 $\{\mathbf{I}_n^{syn}, n = 1, 2, ..., M'\}$, and estimate $E_{p(\mathbf{I};\Lambda,S)}[H^{(\beta)}(z; \mathbf{I})]$ by $\mu_{syn}^{(\beta)} = \frac{1}{M'}\sum_{n=1}^{M'} H^{(\beta)}(z; \mathbf{I}_n^{syn})$. 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 figure 2. These images are from various sources, some digitized from books and postcards, and some from a Corel image

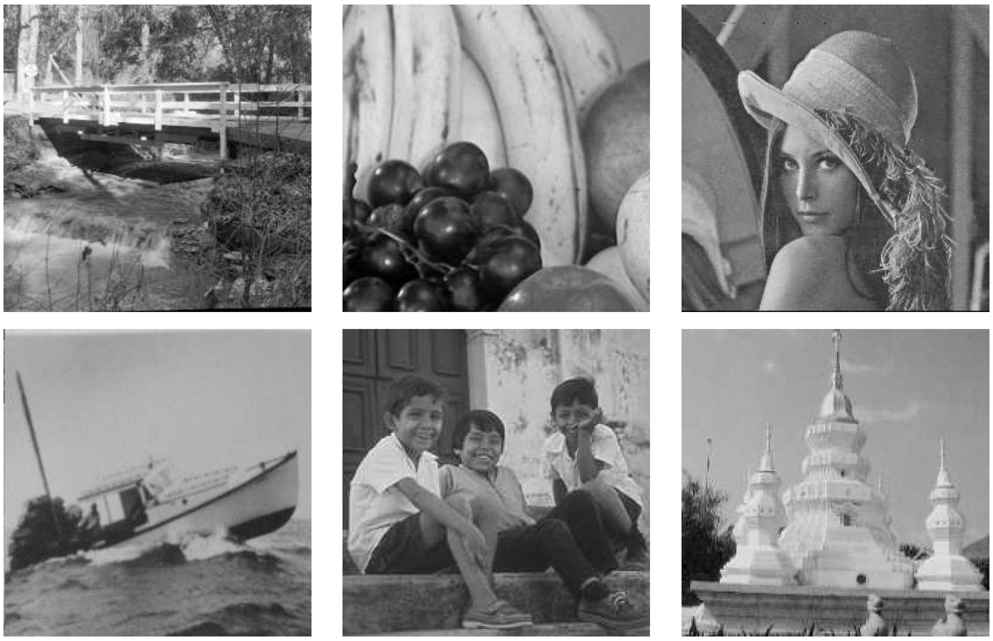


Figure 2: 6 out of the 44 collected natural images.

database. Our database includes both indoor and outdoor pictures, country and urban scenes, and all images are normalized to have intensity between 0 and 31.

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.

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

where $s = \sqrt{2}\sigma$ stands for the scale of the filter. We denote these filters by $LG(s)$. A special filter is $LG(\frac{\sqrt{2}}{2})$, 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.

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

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 figure 3a, and figure 3b is the intensity histogram of an individual image (the temple image in figure 2). It appears that $\mu_{obs}^{(o)}(z)$ is close to a uniform distribution (figure 3c), while the difference between figure 3a and figure 3b is very big. Thus IC for filter $\delta()$ should be small (see table 1).

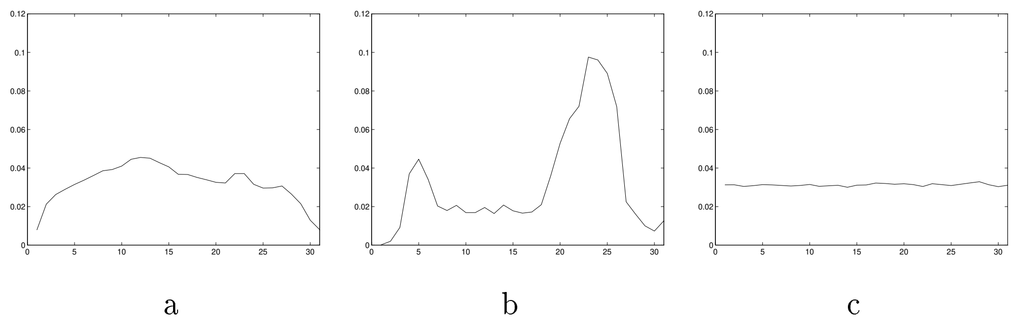
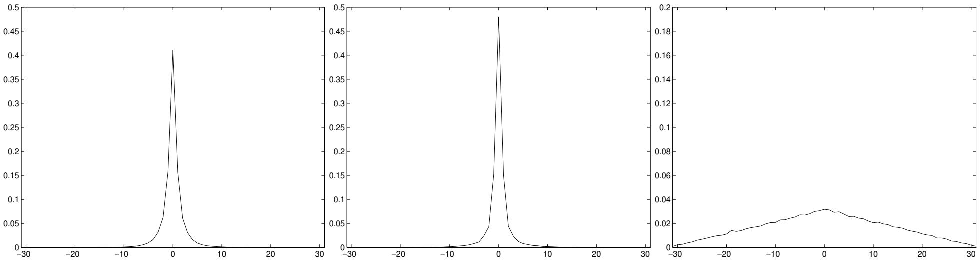


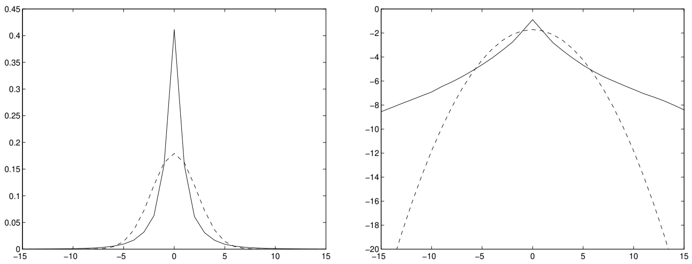
Figure 3: The intensity histograms in domain $[0, 31]$, a, averaged over 44 natural images, b, an individual natural image, c, a uniform noise image.

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. Figure 4a is the average



a b c

Figure 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.



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.

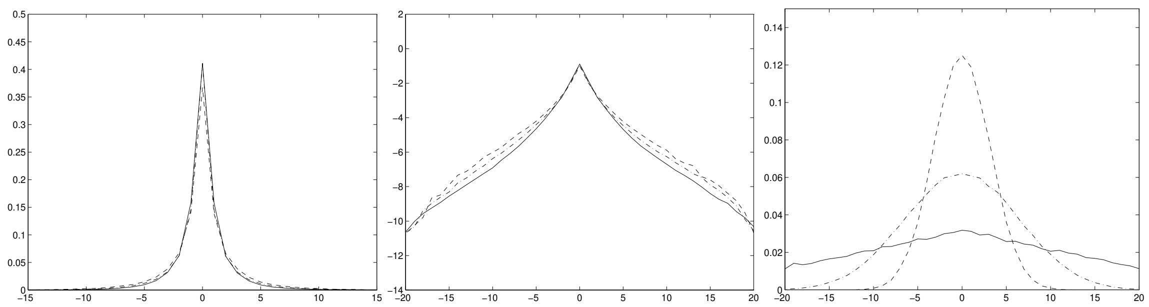
histogram of 44 filtered natural images, figure 4b is the histogram of an individual filtered image (the same image as in figure 3b), and figure 4c is the histogram of a filtered uniform noise image.

The average histogram in figure 4a is very different from a Gaussian distribution. To see this, figure 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, figure 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 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 $N/2^s \times N/2^s$.



a b c

Figure 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).

For the filter ∇_x, let $\mu_{x,s}(z)$ be the average histogram of $\nabla_x \mathbf{I}_n^{[s]}$, over $n = 1, 2, ..., 44$. Figure 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 figure 6c.

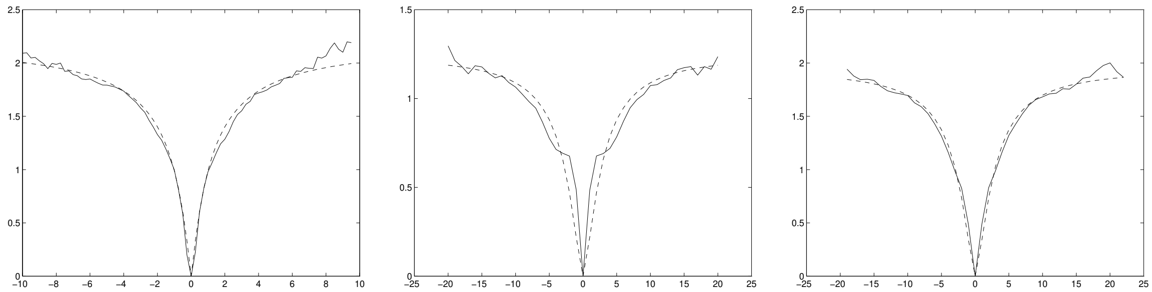
The difference between them are still small. Similar results are observed for $\mu_{y,s}(z)$ $s = 0, 1, 2$, the average histograms of $\nabla_y \mathbf{I}_n^{obs}$. In contrast, figure 6b plots the histograms of $\nabla_x \mathbf{I}^{[s]}$ with $\mathbf{I}^{[s]}$ being a uniform noise image at scales $s = 0, 1, 2$.

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, 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.

Experiment I.



a b c

Figure 7: The three learned 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. $\psi_3(x) = 1.95(1 - 1/(1 + (\mid x \mid /2.8)^{1.5})$

We start from $S = \emptyset$ 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

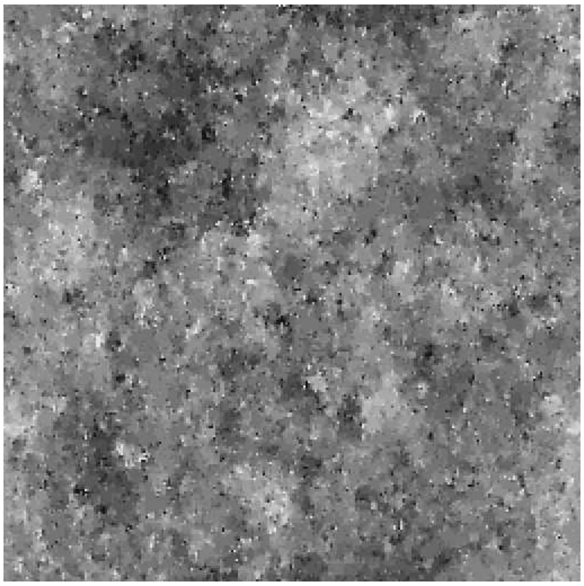


Figure 8: A typical sample of $p_3(\mathbf{I})$ (256×256 pixels).

$F^{(1)}$. 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 0, 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 ICs 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 figure 7. Since $\mu^{(1)}_{obs}(z) = 0$ if $\mid z \mid \geq 9.5$,[4] and $\mu^{(\alpha)}_{obs}(z) = 0$ if $\mid z \mid \geq 22$ for $\alpha = 2,3$, we only plot $\lambda^{(1)}(z)$ for $z \in [-9.5, 9.5]$ and $\lambda^{(2)}(z), \lambda^{(3)}(z)$ for $z \in [-22, 22]$. These three curves are fitted with the functions $\psi_1(z) = 2.1(1 - 1/(1 + (\mid z \mid /4.8)^{1.32})$, $\psi_2(z) = 1.25(1 - 1/(1 + (\mid z \mid /2.8)^{1.5})$,

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

			$p_0(\mathcal{I})$		$p_1(\mathcal{I})$		$p_3(\mathcal{I})$		$p_s(\mathcal{I})$	
Filter	Filter Size	AIF	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.

$\psi_3(z) = 1.95(1 - 1/(1 + (\mid z \mid /2.8)^{1.5})$ respectively. A synthesized image sampled from $p_3(\mathbf{I})$ is displayed in figure 8.

So far, we have used three filters to characterize the statistics of natural images, and the synthesized image in figure 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 list in table 1 in column $p_3(\mathbf{I})$.

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 4 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,

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

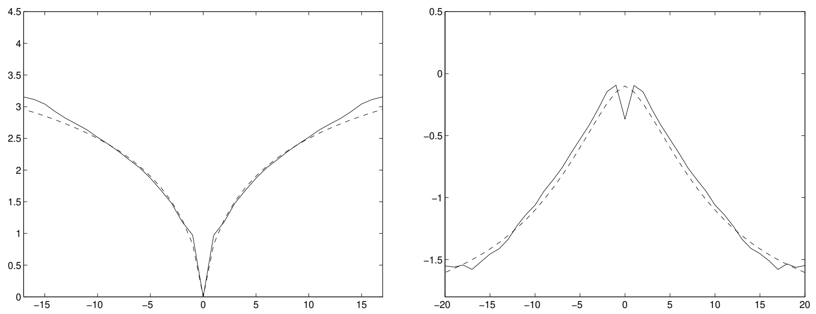
where L_s is the image lattice at level s, and $\bar{\mu}(z)$ is the average of the observed histograms of $\nabla_x \mathbf{I}^{[s]}$ and $\nabla_y \mathbf{I}^{[s]}$ on all 44 natural images at all scales. This results in a maximum entropy distribution $p_s(\mathbf{I})$ with energy of the following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

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

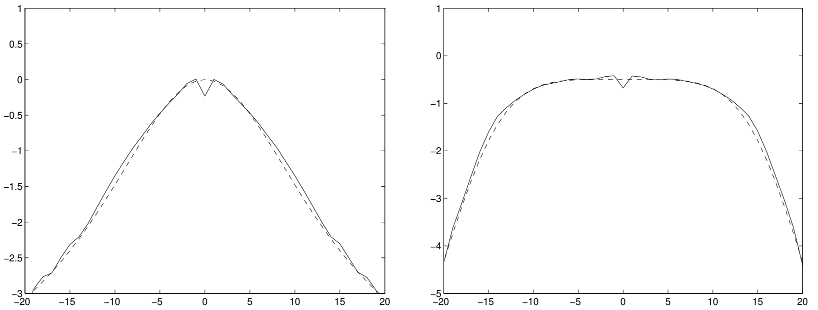
Figure 9 displays $\lambda_{x,s}()$, $s = 0, 1, 2, 3$. At the beginning of the learning process, all $\lambda_{x,s}()$, $s = 0, 1, 2, 3$ are of the form displayed in figure 7 with low values around zero to encourage smoothness. As the learning proceeds, gradually $\lambda_{x,3}()$ turns "upside down" 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 $\lambda_{y,s}()$, $s = 0, 1, 2, 3$. Figure 11 is a typical sample image from $p_s(\mathbf{I})$. To demonstrate it has scale invariant properties, in figure 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 6000 sweeps.

Remark 1. In figure 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

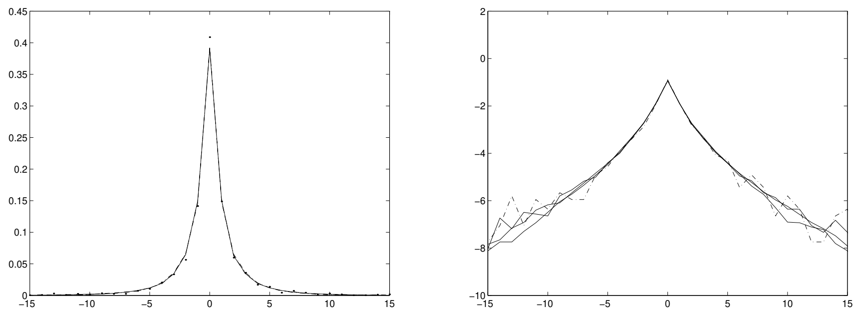


a b



c d

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



a b

Figure 10: a. The histograms of the synthesized image at 4 scales–almost indistinguishable. b. The logarithm of histograms in a.

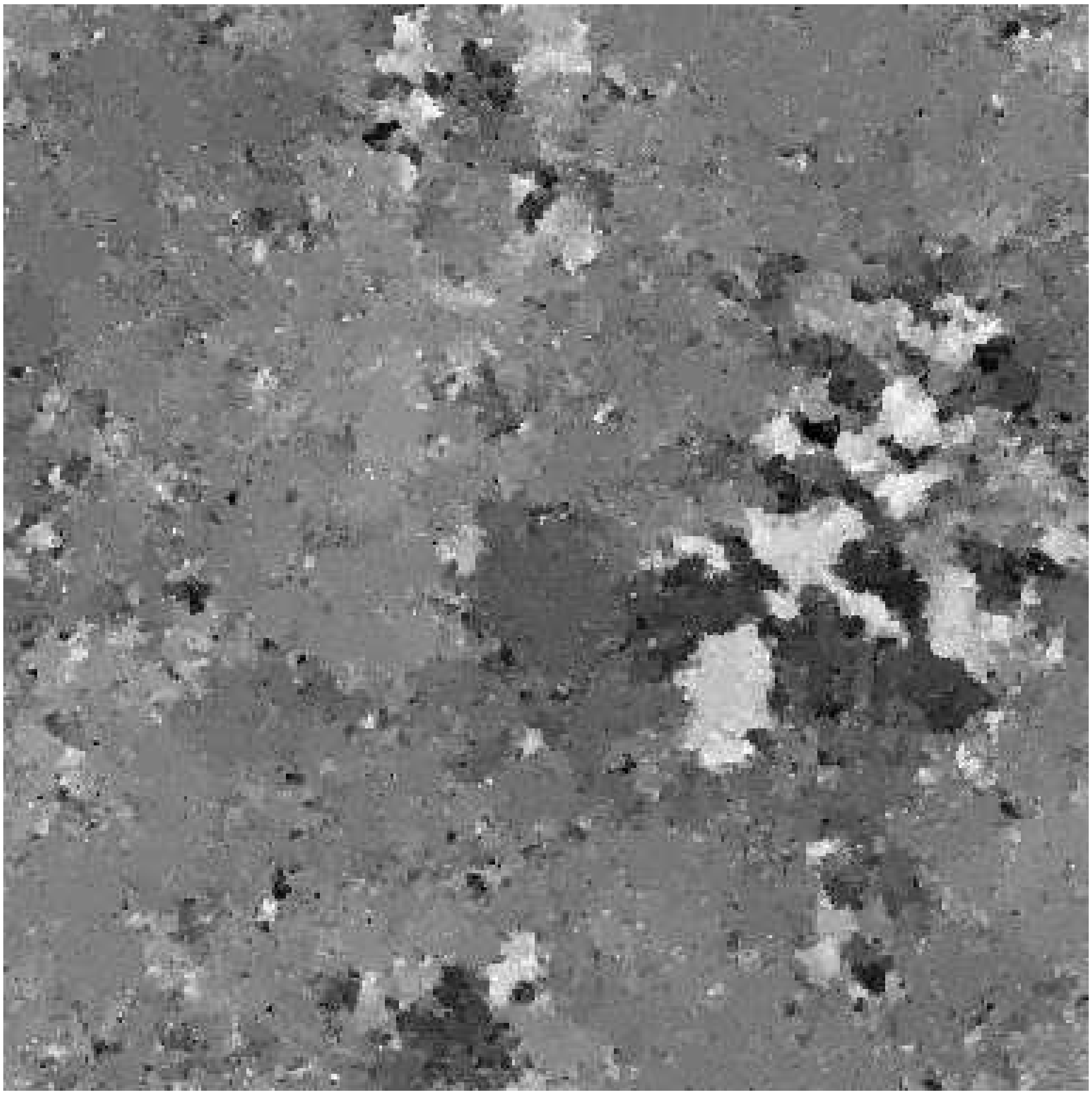


Figure 11: A typical sample of $p_s(\mathbf{I})$ (384×384 pixels).

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 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 $p_s(\mathbf{I})$, $\lambda_{x,0}()$ 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 figure 11, and the intensity contrasts are also higher at the boundary. These appearances are missing in figure 8.

Remark 2. Based on the image in figure 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, ICs 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. Similar phenomenon was observed in our learned texture models [40].

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

$$\begin{aligned}\phi(\xi) &= a(1-\frac{1}{1+(|\xi-\xi_0|/b)^\gamma}) \quad a>0 \\ \psi(\xi) &= a(1-\frac{1}{1+(|\xi-\xi_0|/b)^\gamma}) \quad a<0\end{aligned}$$

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,

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

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 $\nabla_x, \nabla_y, \Delta$ 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.

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:

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

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 equation (14) the Gibbs Diffusion And Reaction Equation (GRADE).

Since the computation of equation (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,

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

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

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,

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

where div is the divergence operator, i.e., $\mathrm{div}(\vec{V}) = \nabla_x P + \nabla_y Q$ for $\vec{V} = (P, Q)$. Perona and Malik defined the *heat conductivity* $c(x,y,t)$ as functions of local gradients, for example:

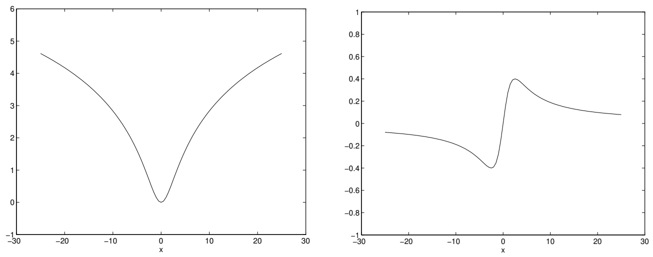
$$\mathbf{I}_t = \nabla_x(\frac{1}{1 + (\mathbf{I}_x/b)^2}\mathbf{I}_x) + \nabla_y(\frac{1}{1 + (\mathbf{I}_y/b)^2}\mathbf{I}_y), \tag{16}$$

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

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

where $\lambda(\xi) = a\log(1 + (\xi/b)^2)$ and $\lambda'(\xi) = a\frac{\xi}{1+(\xi/b)^2}$ are plotted in figure 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.



a $\lambda(\xi)$ b $\lambda^{'}(\xi)$

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

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

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

and a divergence operator,

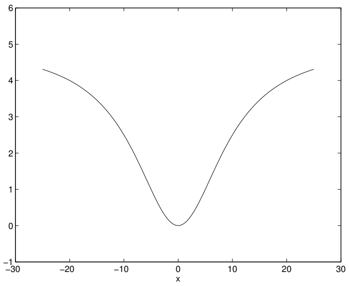
$$\mathrm{div} = F_{-}^{(1)} * \ + F_{-}^{(2)} * \ + \cdots + F_{-}^{(K)} * .$$

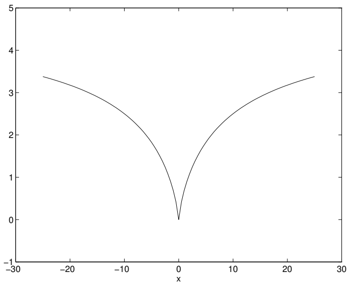
Thus equation (14) can be written as,

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

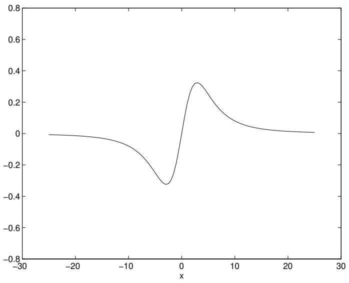
Compared to equation (15) which transfers the "heat" among adjacent pixels, equation (17) transfers the "heat" in many directions in a graph, and the conductivities are defined as functions of the local patterns not just the local gradients.

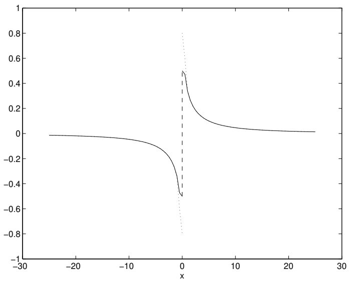
Second, in figure 13, $\phi(\xi)$ has round tip for $\gamma \geq 1$, and a cusp occurs at $\xi = 0$ for $0 < \gamma < 1$ which leaves $\phi^{'}(0)$ undefined, i.e. $\phi^{'}(0)$ can be any value in $(-\infty, \infty)$ as shown by the dotted curves in figure 13d. An interesting fact is that the potential function learned from real world images does have a cusp as shown in figure 9a where the best fit is $\gamma = 0.7$. This cusp forms because 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 $\gamma \geq 1$.





a $\phi(\xi),\ \gamma \geq 1$ b $\phi(\xi),\ \gamma < 1$





c $\phi^{'}(\xi),\ \gamma \geq 1$ d $\phi^{'}(\xi),\ \gamma < 1$

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

By continuity, $\phi^{'}(\xi)$ can be assigned any value in the range $[-\omega, \omega]$ for $\xi \in [-\epsilon, \epsilon]$ and $\omega = c\frac{\epsilon^{\gamma-1}}{(1+(\epsilon/b)^{\gamma})^2}$. 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 $||\omega F_{-}^{(\alpha)}(u-x, v-y)||$. 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.

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. (I) 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]. (II) 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 2 or 3 filters using the GRADE. We run equation (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 figure 14.

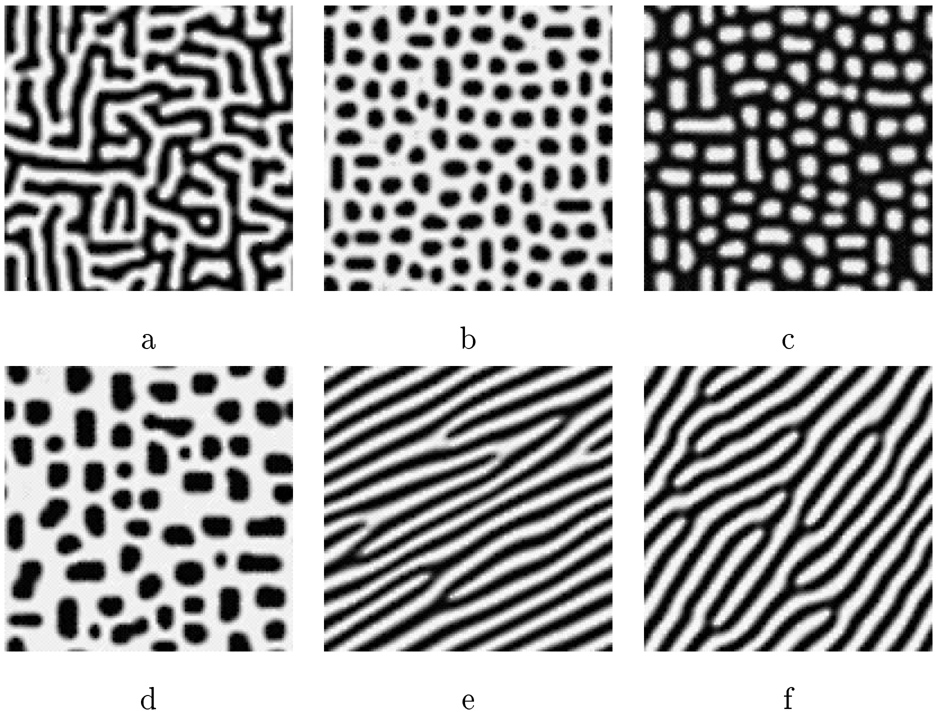


Figure 14: Leopard blobs and zebra stripes synthesized by GRADEs.

For all the six patterns in figure 14, we choose $F_0^{(1)} = \Delta$ the Laplacian of Gaussian filter at level 0 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 figure 14 a,b,c we choose isotropic center-surround filter $LG(\sqrt{2})$ of widow size 7×7 pixels as the reaction filter $F_1^{(2)}$ at level 1 of the image pyramid, and we set $(a = -6.0, b = 10, \gamma = 2.0)$ 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 part as shown in figure 14a. As ξ_o becomes negative, black blobs begin to form as shown in figure 14b where $\xi_o = -6$, and positive ξ_o generates white blobs in the black background as shown in figure 14c where $\xi_o = 6$. The general smoothness appearance of the images is attributed to the diffusion filter. Figure 14d is generated with two reaction filters – $LG(\sqrt{2})$ at level 1 and level 2 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 1 as the reaction filter $F_1^{(2)}$ for figure 14e where $(a = -3.5, b = 10, \gamma = 2.0, \xi_o = 0)$ for $\psi_1^{(2)}(\xi)$. Figure 14f is generated with two reaction filters $Gcos(4, 30^o), Gcos(4, 60^o)$ at level 1 where $(a = -3.5, b = 10, \gamma = 2.0, \xi_o = -3)$ 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 figure 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.

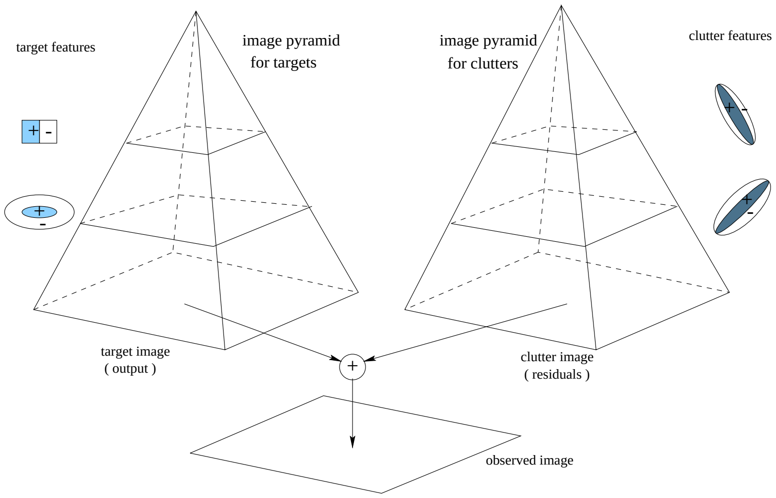


Figure 15: The computational scheme for removing noise and clutter.

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.,

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

$T(t)$ is the "temperature" which controls the magnitude of the random fluctuation. Under mild conditions on U^*, equation (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.

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:

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

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

$$U_s(\mathbf{I}) = \text{the 4-scale energy in equation (12)} \tag{22}$$

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

Figure 16 demonstrates the results: the original image is the lobster boat displayed in figure 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 figure 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 figure 16b, 16c, 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.

Experiment II

In many applications, i.i.d. Gaussian models for distortions are not sufficient. For example, in figure 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

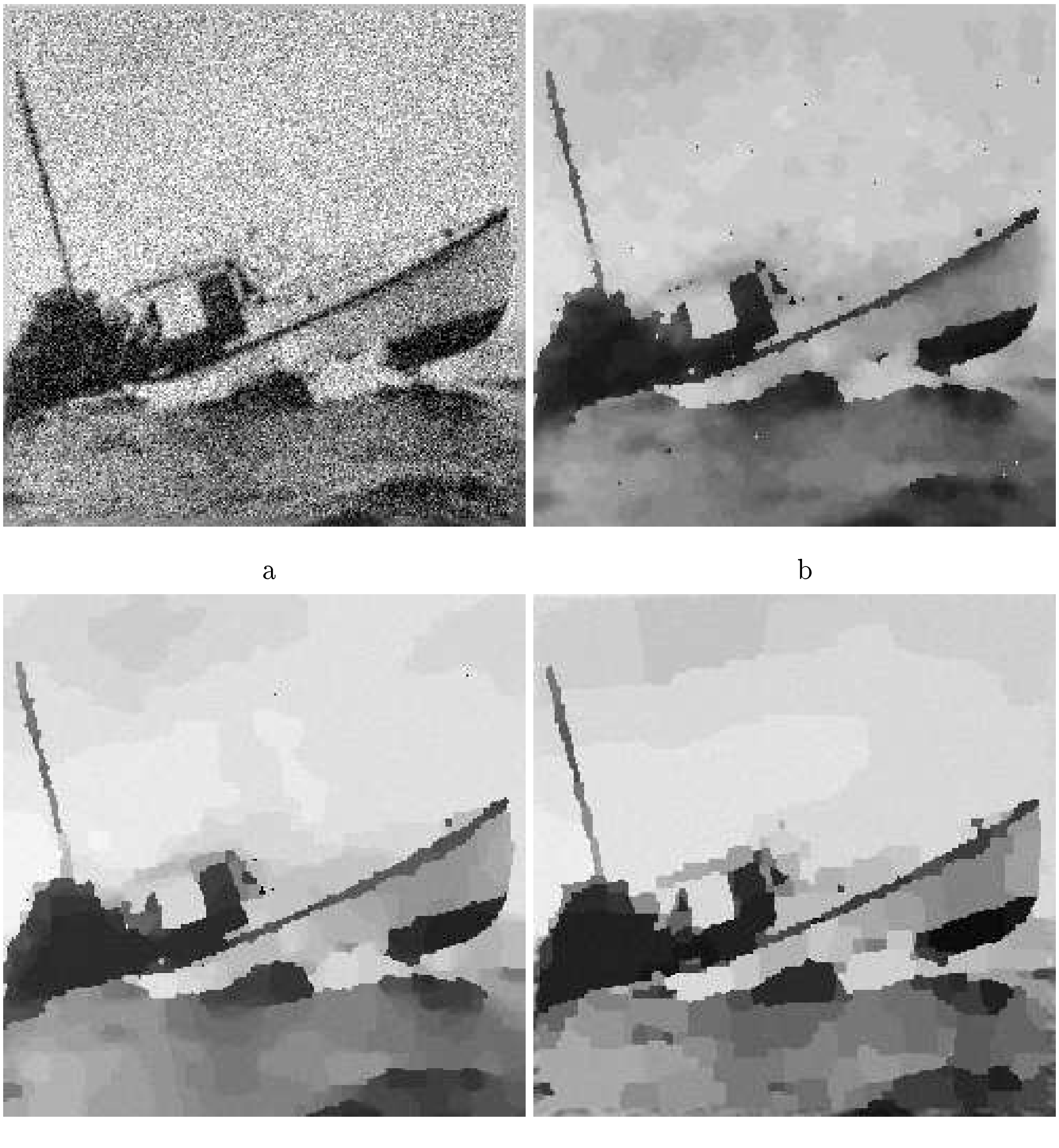
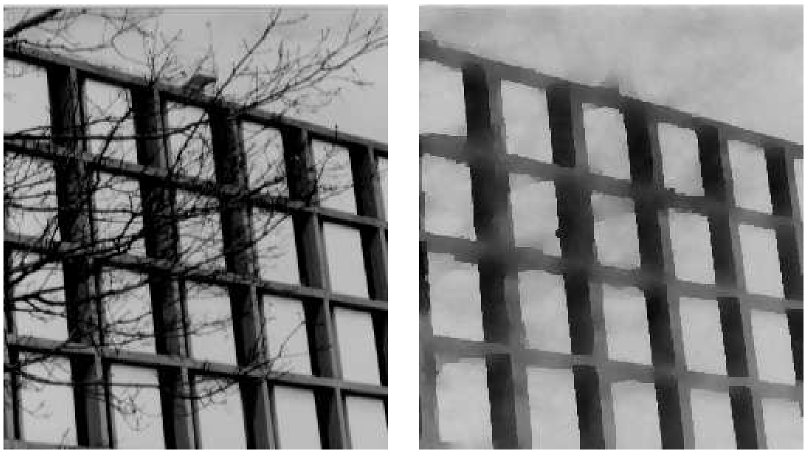


Figure 16: a. The noise distorted image, b. c. d. are respectively restored images by prior models $p_l(\mathbf{I})$, $p_t(\mathbf{I})$ and $p_s(\mathbf{I})$.



a b

Figure 17: a. the observed image, b, the restored image using 6 filters.

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., 0 for sky. In this case, since the trees are always darker than the build, 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 equation (18). We recovered the underlying images by maximizing a posteriori distribution using the stochastic process.

For example, figure 17b is computed using 6 filters with 2 filters for $\mathbf{I}$: $\{\nabla_{x,0}, \nabla_{y,0}\}$, and 4 filters for $\mathbf{I_C}$: $\{\delta, \nabla_x, \nabla_y, Gcos(2, 30^o)\}$, i.e., the potential for $\mathbf{I_C}$ is,

$$U_{\mathbf{C}}(\mathbf{I}) = \sum_{(x,y)} \phi(\nabla_x \mathbf{I}(x,y)) + \phi(\nabla_y \mathbf{I}(x,y)) + \phi^*(\mathbf{I}(x,y)) + \psi^*(Gcos * \mathbf{I}(x,y)).$$

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

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

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.

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

Figure 18b is computed using 8 filters with 4 filters for $\mathbf{I}$ and 4 filters for $\mathbf{I_C}$. 13 filters are used for figure 19 where the clutter is much heavier.

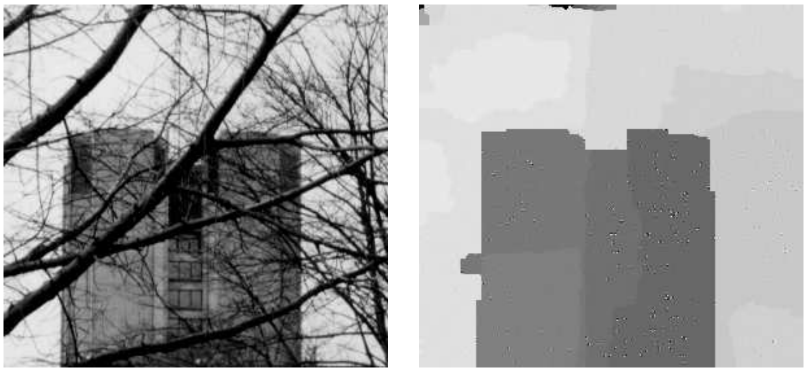
As a comparison, we run the anisotropic diffusion process [25] on figure 19a, and images at iterations $t = 50, 100, 300$ are displayed in figure 20. As we can see that as $t \to \infty$, $\mathbf{I}(t)$ becomes a flat image. A robust anisotropic diffusion equation is recently reported in [2].

a b

Figure 18: a. an observed image, b. the restored image using 8 filters.

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



a b

Figure 19: a. the observed image, b. the restored image using 13 filters.

a (t=50) b (t=100) c (t=300)

Figure 20: Images by anisotropic diffusion at iteration $t = 50, 100, 300$ for comparison.

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 priors 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.

References

[1] A. Berger, V. Della Pietra, and S. Della Pietra, "A maximum entropy approach to natural language processing", *Computational Linguistics*, 22(1), 1996.

[2] Black, M. Sapiro, G. Marimont, D and Heeger, D. "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*, 19(1), 1996.

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

[5] 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.

[6] 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.

[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", in *Markov Random Fields – Theory and Applications* Academic Press, Inc, 1993.

[9] 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.

[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. PAMI*, vol.14, pp.367-383, 1992.

[12] D. Geiger and F. Girosi, "Parallel and deterministic algorithms for MRFs: surface reconstruction", *IEEE Trans PAMI*, 13(5):401-412, 1991.

[13] D. Geiger and A. L. Yuille. "A common framework for image segmentation". *Inter.J.of Computer Vision* 6(3), pp 227-243 1991.

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

[15] P. Grindrod, *The theory and applications of reaction-diffusion equations*, 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, pp189-224, 1995.

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

[18] J. Marroguin, S. Mitter, and T. Poggio, "Probabilistic solution of ill-posed problems in computational vision", *J. Amer.Stat. Assoc.* 82(397), 1987.

[19] P. Meer, D. Mintz, D.Y. Kim, and A. Rosenfeld, "Robust regression methods for computer vision: a review", *Int'l J. of Computer Vision*, 6(1), 1991.

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

[21] J. D. Murray, "A pre-pattern formation mechanism for mammalian coat markings", *Journal of Theoretical Biology*, vol. 88, pp 161-199, 1981.

[22] W. Niessen etc. "A general framework for geometry-driven evolution equations", *Int'l J. computer Vision*, 21(3), 187-205,1997.

[23] M. Nitzberg and T. Shiota, "Nonlinear image filtering with edge and corner enhancement" *IEEE. Trans. PAMI*, vol.14, Aug. 1992.

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

[25] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion" *IEEE Trans. on PAMI*, vol.12, No.7, 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. of IEEE*, vol.78, 1481-1497, 1990.

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

[29] A. Sherstinsky and R. Picard, "M-lattice: from morphogenesis to image processing" *IEEE Trans. on Image Processing*, vol.5, no.7, July 1996.

[30] N. V. Swindale, "A model for the formation of ocular dominance stripes", *Proc. R. Soc. Lond.* B 208, pp243-264, 1980.

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

[32] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-posed Problems*, 1906, (Translated version), 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", *Journal of Optical Society of America A.* Vol.4, No.12, 1987.

[36] K. Wilson, "The renormalization group: critical phenonmena 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*, 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, Bayes/MDL for multi-band image segmentation". *IEEE Trans.on PAMI.* Vol.18, No.9, 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.*, 1996a. (to appear in IJCV).

[41] S. C. Zhu, Y. N. Wu and D. B. Mumford. "Minimax entropy principle and its application to texture modeling", *Neural Computation*, (to appear) 1996b.

[42] S. C. Zhu and D. B. Mumford, "Learning generic prior models for visual computation" Harvard Robotics Lab TR-96-05, a short version appeared in *CVPR97*.