Learning Generic Prior Models for Visual Computation

Learning Generic Prior Models for Visual Computation

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

Abstract

This paper presents a novel theory for learning generic prior models from a set of observed natural images based on a minimax entropy theory that the authors studied in modeling textures. We start by studying the statistics of natural images including the scale invariant properties, then generic prior models were learnt to duplicate the observed statistics. The learned Gibbs distributions confirm and improve the forms of existing prior models. More interestingly inverted potentials are found to be necessary, and such potentials form patterns and enhance preferred image features. The learned model is compared with existing prior models in experiments of image restoration. [1]

1 Introduction and motivation

Many generic smoothness models have been widely used in visual computation ranging from image restoration, motion analysis, to 3D surface reconstruction. For example, In image segmentation (Geman and Geman 1984, Blake and Zisserman 1987, Mumford and Shah 1989), these smoothness prior models take the forms as the following joint probability distribution:

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

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

These prior models enjoy nice explanations in terms of regularization theory (Poggio, Torre, and Koch 1985), physical modeling (Terzopoulos 1983), Bayesian theory (Geman and Geman 1984) and robust statistics (Geiger and Yuille 1991, Black and Rangarajan 1997), there is, however, little rigorous theoretical or empirical justifications for applying these prior models to general images, and the following questions are not answered in the literatures. i). Why are the differential operators good choices in capturing image features? ii). What are the best forms for $p(\mathbf{I})$ and $\psi()$? iii). Real world scenes are observed at arbitrary scales, thus a good prior model should remain the same for image features at multiple scales. However none of the existing prior models on $2D$ images has scale-invariant property, i.e, they are not renormalizable in terms of the renormalization group theory (Wilson 1975).

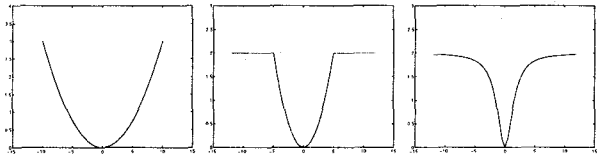


Figure 1: Three existing forms for $\psi()$. a, Quadratic, b, Line process, c T-function.

This paper presents a novel theory for learning generic prior models from a set of observed natural images [2] based on a minimax entropy theory that the authors studied in modeling textures (Zhu, Wu, and Mumford 1996). We start by studying the statistics of natural images including the scale invariant properties, then generic prior models were learnt to duplicate the observed statistics. First, instead of being limited to differential operators, our theory examines whatever filters capture the structures of natural images, such as Gabor filters (Daugman 1985). An information criterion is put forth for choosing the most informative features (or filters) in $p(\mathbf{I})$. Second, unlike previous prior models which subjectively assume some parametric forms for the potential functions $\psi()$,

[1] This work was supported by an NSF grant and an ARO grant to Mumford. For correspondence, contact Song Chun Zhu at zhu@dam.brown.edu, detailed reports are available in http://hrl.harvard.edu/people/postdocs/zhu/zhu.html

[2] Here, natural images refer to an arbitrary view of the world, indoor or outdoor.

1063-6919/97 $10.00

our theory uses non-parametric forms and learns them from observed images. The learned Gibbs distributions confirm and improve the forms of existing prior models. More interestingly inverted potentials are found to be necessary, and such potentials produce patterns and enhance preferred image features. The learned model is compared with existing prior models in experiments of image restoration.

This paper is arranged as follows. Section (2) discusses the objective and theory of learning prior models. Section (3) presents a novel information criterion for model selection. Section (4) and section (5) demonstrate some experiments on the statistics of natural images and prior learning. Section (6) compares different prior models by experiments of image restoration. Finally section (7) concludes with a discussion.

2 Learning prior models by maximum entropy

We define an image $\mathbf{I}$ on an $N \times N$ lattice L, and we assume that there is an underlying joint probability distribution $f(\mathbf{I})$ on the image space for general natural images – arbitrary views of the world. Let $NI^{obs} = \{\mathbf{I}_n^{obs}, \ n = 1, 2.., M\}$ be a set of observed images which are independent samples from $f(\mathbf{I})$. Then *the objective of learning a generic prior model is to look for common features and their statistics from the observed natural images, based on which a model $p(\mathbf{I})$ is inferred as an estimate to $f(\mathbf{I})$. $p(\mathbf{I})$, as a prior model, will bias vision algorithms against image features which are not typical in natural images, such as noise distortions and blurring.* For purpose of learning a generic prior model, it is reasonable to assume that any image features have equal chance to occur at any location, so $f(\mathbf{I})$ is translation invariant with respect to (x, y).

To study the properties of images $\{\mathbf{I}_n^{obs}, \ n = 1, 2.., M\}$, we start from exploring a set of linear filters $\{F^{(\alpha)}, \alpha = 1, 2, ..., K\}$ which are characteristic of the observation.

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

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

where $\delta()$ is a Dirac function with point mass concentrated at0. $\mid L \mid$ is the size of the image lattice

We compute the average histogram of all observed images as the *observed statistics*,

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

We note that $\mu_{obs}^{(\alpha)}(z)$ is an unbiased estimate for $E_f[H^{(\alpha)}(z; \mathbf{I})]$, and the latter is an 1D marginal distribution of $f(\mathbf{I})$.

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

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

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

where $\Lambda = \{\lambda^{(1)}(), ..., \lambda^{(K)}()\}$ is a set of potential functions on the features extracted by S. In practice, the filter responses are divided into a finite number of bins, thus $\lambda^{(\alpha)}()$ is approximated by a piecewise 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 learnt $p(\mathbf{I}, ; \Lambda, S)$ can reproduce the observed statistics:

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

So 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. We adopted the Gibbs sampler (Geman and Geman 1984), which simulates an inhomogeneous Markov chain in image space (Winkler 1995). 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 synthesized histograms closer to the observed ones. For a detailed account of the computation of $\lambda^{(\alpha)}$'s, the readers are referred to (Zhu, Wu and Mumford 1996).

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 *As $M \to \infty$, $K \to \infty$, with only linear filters used, $p(\mathbf{I}; \Lambda, S)$ converges to the underlying distribution $f(\mathbf{I})$.*

We shall discuss how to choose filters in the next section

3 Information criterion for model selection

We notice that the statistics of natural images vary from image to image. For each image $\mathbf{I}_n^{obs}$ or a group of images in a given domain, it is desirable to have a specific underlying distribution $f_n(\mathbf{I})$. Given a set S, and an ME distribution $p(\mathbf{I};\Lambda,S)$, the goodness of $p(\mathbf{I};\Lambda,S)$ with respect to $\mathbf{I}_n^{obs}$ depends on S, and is often measured by the Kullback-Leibler information distance between $f_n(\mathbf{I})$ and $p(\mathbf{I};\Lambda,S)$ (Kullback and Leibler 1951),

$$KL(f_n(\mathbf{I}),p(\mathbf{I};\Lambda,S)) = \int \cdot \int f_n(\mathbf{I}) \log \frac{f_n(\mathbf{I})}{p(\mathbf{I};\Lambda,S)} d\mathbf{I}.$$

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

$$S^* = \arg\min_{|S|=K} D_K = \arg\min_{|S|=K} \frac{1}{M}\sum_{n=1}^{M} KL(f_n(\mathbf{I}),p(\mathbf{I};\Lambda,S))$$

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 a filter bank and comparing their entropies is computational too expensive. Instead, we propose a stepwise greedy procedure for minimizing the average KL-distance. We start from $S=\emptyset$ and $p(\mathbf{I};\Lambda,S)$ a uniform distribution, then it sequentially introduces one filter at a time. At each time the added filter leads to the maximum decrease in the average KL-distance, and keep doing this until the decrease is smaller than a certain value.

Let S be the currently selected set, and $p(\mathbf{I};\Lambda,S)$ the ME distribution duplicating the observed statistics. For the next step, let $S_+ = S \cup \{F^{(\beta)}\}$ be a new feature set, and $p(\mathbf{I};\Lambda_+,S_+)$ the new ME distribution. Our greedy procedure chooses the next filter by minimizing the following information criterion IC^*,

$$F^{(K+1)} = \arg\min_{F^{(\beta)}\in B/S} D_K - D_{K+1}$$

To compute the above equation, for each $f_n(\mathbf{I})$ and S we introduce a new ME distribution $p(\mathbf{I};\Lambda_n^*,S)$ which reproduces the statistics of $\mathbf{I}_n^{obs}$,

$$E_{p(\mathbf{I};\Lambda_n^*,S)}[H^{(\alpha)}(z;\mathbf{I})] = H^{(\alpha)}(z;\mathbf{I}_n^{obs})$$

$H^{(\alpha)}(z;\mathbf{I}_n^{obs})$ is a closer estimate to the marginal distribution $E_{f_n}[H^{(\alpha)}(z;\mathbf{I})]$ than $\mu_{obs}^{(\alpha)}(z)$. In the following we assume $E_{f_n}[H^{(\alpha)}(z;\mathbf{I})] = H^{(\alpha)}(z;\mathbf{I}_n^{obs})$. Thus $f_n(\mathbf{I})$ is better estimated by $p(\mathbf{I};\Lambda_n^*,S)$ than by $p(\mathbf{I};\Lambda,S)$, as stated in the following proposition.

Proposition 3 *Given two ME distributions $p(\mathbf{I};\Lambda,S)$ and $p(\mathbf{I};\Lambda_n^*,S)$ defined above, $KL(f_n(\mathbf{I}),p(\mathbf{I};\Lambda,S)) = KL(f_n(\mathbf{I}),p(\mathbf{I};\Lambda_n^*,S)) + KL(p(\mathbf{I};\Lambda_n^*,S),p(\mathbf{I};\Lambda,S))$.*

[Proof]. The proof follows proposition 4 in (Zhu, Wu, and Mumford 1996).

By proposition 3, we obtain

$$\begin{aligned} & KL(f_n(\mathbf{I}),p(\mathbf{I};\Lambda,S)) - KL(f_n(\mathbf{I}),p(\mathbf{I};\Lambda_+,S_+)) \\ = \; & KL(p(\mathbf{I};\Lambda_{n+}^*,S_+),p(\mathbf{I};\Lambda_+,S_+)) \\ - \; & KL(p(\mathbf{I};\Lambda_{n+}^*,S_+),p(\mathbf{I};\Lambda_n^*,S)) \\ - \; & KL(p(\mathbf{I};\Lambda_n^*,S),p(\mathbf{I};\Lambda,S)) \end{aligned} \quad (4)$$

The following proposition measures the distance of two ME distributions in terms of the difference of their marginal distributions.

Proposition 4 *Let $p_0(\mathbf{I})$ and $p(\mathbf{I})$ be two ME distributions, $E_{p_0(\mathbf{I})}[H^{(\alpha)}(z;\mathbf{I})] = h_0^{(\alpha)}$ and $E_{p(\mathbf{I})}[H^{(\alpha)}(z;\mathbf{I})] = h^{(\alpha)}$ for $\alpha = 1,2,...,k$. Denote $h_0 = (h_0^{(1)},h_0^{(2)},...,h_0^{(k)})$ and $h = (h^{(1)},h^{(2)},...,h^{(k)})$. Fixing h_0, $KL(p_0(\mathbf{I}),p(\mathbf{I}))$ is a function of h, and $KL(p_0(\mathbf{I}),p(\mathbf{I})) = (h-h_0)Var^{-1}(h^*)(h-h_0)^T$, where $Var(h^*)$ is a variance matrix of h^* and h^* lies between h_0 and h.*

[Proof] The proof follows proposition 3 in (Zhu,Wu and Mumford 1996).

In practice, for computational convenience, we use the L_1 norm distance to replace the quadratic term,

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

In summary, if we use the L_1 norm distance, together with equation (4), we approximate IC^* by IC defined below.

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

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

In practice, we need to sample $p(\mathbf{I}; \Lambda, S)$, thus synthesize images $\{\mathbf{I}_n^{syn}, n = 1, 2, ..., M'\}$, and use averaged histogram of these synthesized images to estimate $E_{p(\mathbf{I};\Lambda,S)}[H^{(\beta)}(z; \mathbf{I})]$. For a filter $F^{(\beta)}$, the bigger AIG is, the more information $F^{(\beta)}$ captures, as it measures the error between the current model and the observations. AIF is a measure of disagreement between observed images. The bigger AIF is, the less common $F^{(\alpha)}$ is shared by all images.

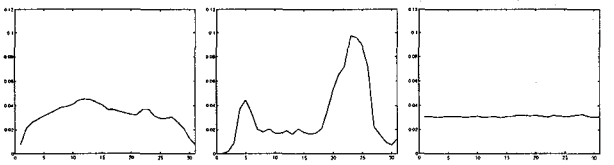
4 Statistics of natural images

Figure 2: 4 out of the 44 collected natural images.

We start from studying the statistical properties of natural images. We collect a set of 44 natural images, four of which are shown in figure 2, and these 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 histograms of linearly filtered images.

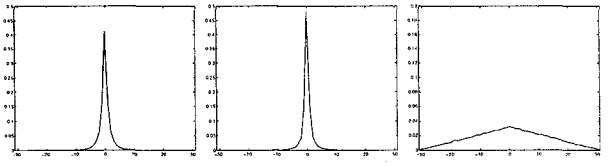
Firstly, for some features, the statistics of natural images vary largely from image to image. As an example, we study the $\delta()$ filter, the filter response is the intensity itself. The average intensity histogram of the 44 images μ_{obs}^{o} is plotted in figure (3.a), and figure (3.b) is the intensity histogram of an individual image (the temple image in figure (2)). It appears that μ_{obs}^{o} is close to an uniform distribution (figure (3.c)), whereas the difference between figure (3.a) and figure (3.b) is very big. Thus IC for filter $\delta()$ is very small.



a b c

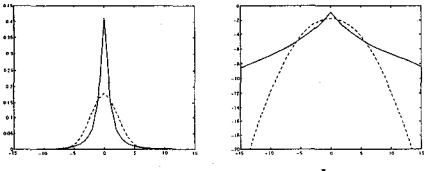
Figure 3: The intensity histograms, a, averaged over 44 natural images, b, an individual natural image, c, an uniform noise image.

Secondly, for some other filters, the histograms of filtered images are amazingly consistent across all 44 natural images, and they are very different from those of noise images. Therefore the IC is relatively large for these features. For example, we look at filter ∇_x and the histograms are plotted in figure (4). The average histogram in figure (4.a) is very different from a Gaussian distribution. Figure (5.a) plots it against a Gaussian curve (dashed one) of the same mean and same variance. The histogram of natural images has higher kurtosis and heavier tails. Similar results are reported in (Field 1994). To see the difference of the tails, we plot the logarithm of the two curves in figure (5.b).



a b c

Figure 4: The histograms of $\nabla_x \mathbf{I}$, a. averaged over 44 natural images, b, an individual natural image (the same image as in figure (3.b)), c, an uniform noise image.



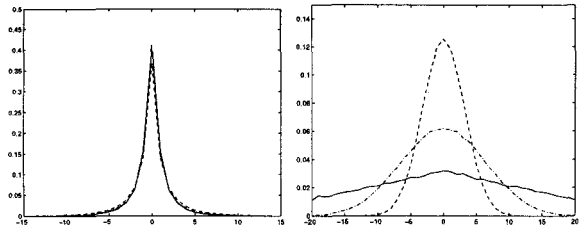
a b

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

Thirdly, the statistics of natural images are scale invariant with respect to some features. We look at

filter ∇_x again. Let $\mu^{[0]}_{xobs}$ be the average histogram of filtered images, then we scale down all observed images from $N \times N$ pixels to $N/2 \times N/2$ pixels by averaging 2×2 pixels, and we compute the average histogram $\mu^{[1]}_{xobs}$ over these scaled images. Continuing this process we obtained $\mu^{[s]}_{xobs}$ where $s = 2, 3, ...$ is the index of the layer or scale in the image pyramid.

Figure (6.a) plots $\mu^{[s]}_{xobs}$, for $s = 0, 1, 2$, and they are almost identical. In contrast, figure (6.b) plots the histograms of $\nabla_x \mathbf{I}$ with $\mathbf{I}$ being an uniform noise image at scales $s = 0, 1, 2$. Similar results are observed for filter ∇_y.

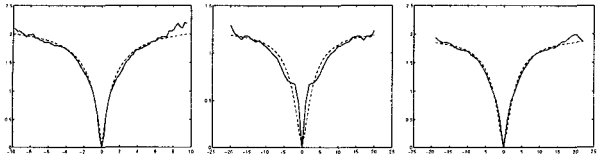


a b

Figure 6: a. $\mu^{[s]}_{xobs}$ $s = 0, 1, 2$. b. histograms of a filtered uniform noise image at scales: $s = 0$ (solid curve), $s = 1$ (dash-dotted curve), and $s = 2$ (dashed curve).

5 Simulations of prior learning

This section briefly presents the experiments on learning generic prior models. We first choose a general filter bank to characterize the interesting features of natural images. This filter bank includes the intensity filter $\delta()$, the Laplacian of Gaussian filters at various scales, and Gabor filters with both sine and cosine components at various scales and orientations.



a b c

Figure 7: The three potential functions: a. $\lambda^{(1)}()$, b. $\lambda^{(2)}()$, c. $\lambda^{(3)}()$. Dashed curves are the fitting functions $\psi()$ with parameters (a, c, γ) being: a. $(2.1, 4.8, 1.32)$, b. $(1.25, 2.8, 1.5)$, and c. $(1.95, 2.8, 1.5)$

According the information criterion discussed before, we found that $\Delta, \nabla_x, \nabla_y$ are sequentially the first three important filters whose IC are the biggest among all filters in the filter bank, where Δ is the second differential operator whose impulse response is a 3×3 window $[0, 1, 0; 1, -4, 1; 0, 1, 0]$. Detailed account for the IC's of each filter is referred to (Zhu and Mumford 1996).

Thus a prior model is learned as following,

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

The potential functions $\lambda^{(\alpha)}(), \alpha = 1, 2, 3$ are plotted in figure (7). $\lambda^{(\alpha)}(), \alpha = 1, 2, 3$ are well-fit by a family of functions $\psi(x) = a(1 - 1/(1 + (\mid x \mid /c)^{\gamma})$ with (a, c, γ) being parameters. A synthesized image sampled from $p_3(\mathbf{I})$ is displayed in figure (8).

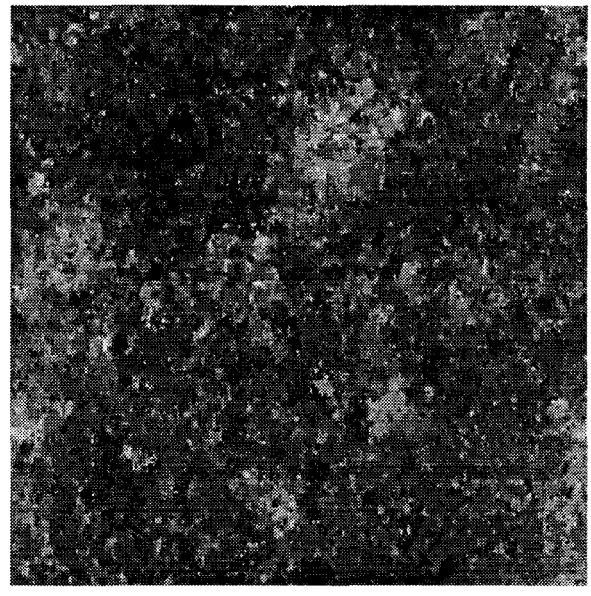


Figure 8: A typical sample of $p_3(\mathbf{I})$ (256×256 pixels).

Although we have used the three most informative filters to extract the structures and statistics of natural images, the synthesized image according to model $p_3(\mathbf{I}; \Lambda)$, as shown 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 encourage intensity breaks, but it only generates small speckles instead of big regions and long edges as one may expected. We also sampled the distributions with potential function shown in figure (1), the sampled images have even less features.

In the next experiment, we shall study a prior model which has the scale invariant property with respect to filters ∇_x and ∇_y as we find in natural images.

Given an image $\mathbf{I}$ defined on an $N \times N$ lattice L. We build an image pyramid by scaling the images as before with $\mathbf{I}^{[s]}, s = 0, 1, 2, 3$ being four layers of the pyramid. We set $\mathbf{I}^{[0]} = \mathbf{I}$, and $\mathbf{I}^{[s]}$ is defined on lattice $L^{[s]}$, which is of size $N/2^s \times N/2^s$ pixels. Let $H_x(z; \mathbf{I}^{[s]})$ denotes the histogram of $\nabla_x \mathbf{I}^{[s]}$ for $s = 0, 1, 2, 3$, and $\mu^{[s]}_{x,obs}(z)$ the average of $H_x(z; \mathbf{I}^{obs[s]}_n)$. Similarly we define and $H_y(z; \mathbf{I}^{[s]})$ and $\mu^{[s]}_{y,obs}(z)$.

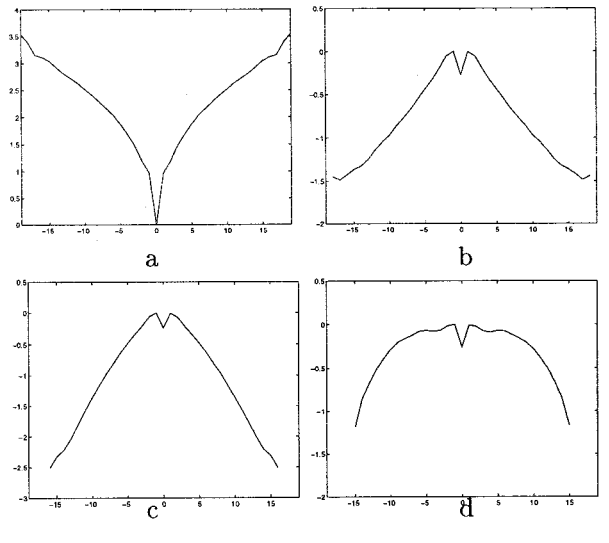


Figure 9: Learned potential functions: a. $\lambda^{(0)}_x$, b. $\lambda^{(1)}_x$, c. $\lambda^{(2)}_x$, d. $\lambda^{(3)}_x$.

We ask for a probability model $p(\mathbf{I})$ which reproduces the observed statistics over 4 scales:

$$\begin{aligned} E_p[H_x(z; \mathbf{I}^{[s]})] &= \mu^{[s]}_{x,obs}(z), \quad s = 0, 1, 2, 3. \forall z \\ E_p[H_y(z; \mathbf{I}^{[s]})] &= \mu^{[s]}_{y,obs}(z), \quad s = 0, 1, 2, 3. \forall z \end{aligned}$$

This results in an ME distribution:

$$\begin{aligned} p_s(\mathbf{I}; \Lambda, S) &= \frac{1}{Z} e^{-\sum_{s=0}^{3} U_s(\mathbf{I}^{[s]})} \\ U_s(\mathbf{I}^{[s]}) &= \sum_{(x,y) \in L_s} \lambda^{[s]}_x(\nabla_x \mathbf{I}^{[s]}_{(x,y)}) + \lambda^{[s]}_y(\nabla_y \mathbf{I}^{[s]}_{(x,y)}). \end{aligned}$$

Figure 9 displays the learned potential functions $\lambda^{[s]}_x(), s = 0, 1, 2, 3$. Similar results are observed for $\lambda^{[s]}_y(), s = 0, 1, 2, 3$. Figure (10) is a typical sample image from $p_s(\mathbf{I}; \Lambda)$, which has almost identical histogram for filtered images across 4 scales.

In contrast to existing prior models in vision, the learned model in figure (9) has **inverted** potentials $\lambda^{[s]}_x()$ for $s = 1, 2, 3$. Such potentials have significant meanings in visual computation. In image restoration,

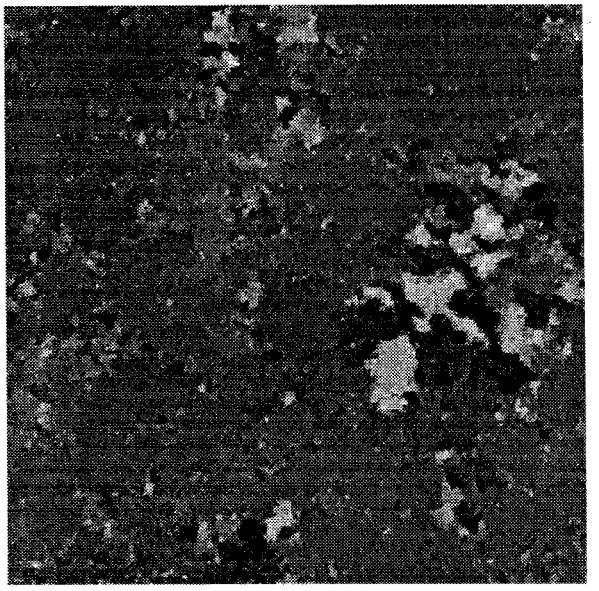


Figure 10: A typical sample of $p_s(\mathbf{I}; \Lambda)$ (384 × 384 pixels).

when a high intensity difference $\nabla_x \mathbf{I}^{[s]}(x, y)$ presents, it is very likely to be noise if $s = 0$. However this is not true for $s = 1, 2, 3$. Additive noises can hardly pass to the high layers of the pyramid because at each layer the 2×2 averaging operator reduces the variance of noise by 4 times. When $\nabla_x \mathbf{I}^{[s]}(x, y)$ is large for $s = 1, 2, 3$, it is more likely to be a true edge and object boundary. So in $p_s(\mathbf{I})$, $\lambda^{[0]}_x()$ suppresses noise at the first layer, whereas $\lambda^{[s]}_x(), s = 1, 2, 3$ encourage sharp edges to form, and thus enhance blurred boundaries. We notice that figure (10) shows regions of various scales, and the intensity contrasts are also higher at the boundary. These are missing in figure (8). Model $p_s(\mathbf{I})$ further leads to the study a new class of reaction-diffusion equations with the inverted terms produce reaction and form patterns. For more detailed account, the readers are referred to (Zhu and Mumford 1997).

6 Comparison of prior models

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

Figure (11.a) shows an input image $\mathbf{I}^d$, and it is the lobster boat image in figure (2) distorted with additive i.i.d. Gaussian noises. Thus the data model $p(\mathbf{I}^d \mid \mathbf{I})$ is known to be Gaussian. Then given a prior model $p(\mathbf{I})$,

by Bayesian rule we restore an image $\mathbf{I}$ by maximizing *a posteriori* probability (MAP) $p(\mathbf{I}^d \mid \mathbf{I})p(\mathbf{I})$. Following (Geman and Geman 1984), we use simulated annealing to compute the MAP-estimate. The restored images using $p_l(\mathbf{I})$, $p_t(\mathbf{I})$ and $p_s(\mathbf{I})$ are shown in figure (11.b), (11.c), (11.d) respectively. $p_s(\mathbf{I})$, which is the only model with the inverted potential terms, appears to have the best effect in recovering the boat, especially the top of the boat, but it also enhances the water.

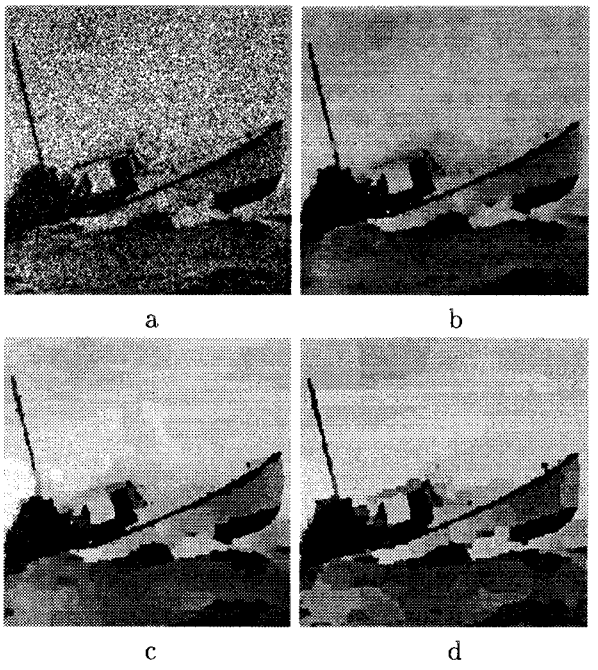


Figure 11: a. The noise distorted image, b. c. d. are respectively restored images by prior models $p_l(\mathbf{I})$, $p_t(\mathbf{I})$ and $p_s(\mathbf{I})$.

Due to space limitation, more experiments, such as clutter removal etc, are referred to (Zhu and Mumford 1997).

7 Discussion

In this paper, a general theory is proposed for learning generic prior models for natural images. We argue that the same strategy can be used in other applications. For example, learning prior models for MRI images, and for 3D surfaces, where prior models of different forms are expected.

An important fact in the learned prior models is the inverted potentials associated with reaction, pattern formation and feature enhancement. Although the synthesized images bear important features of natural images, they are still far from realistic ones. In other words, these generic prior models can do very little beyond image restoration. This is mainly due to the fact that all generic prior models are assumed to be translation invariant. This homogeneity assumption is unrealistic.

We call the generic prior models studied in this paper *the first level prior*. A more sophisticated prior model, which we call *second level prior*, should incorporate concepts like object geometry, and such prior model is used in image segmentation (Zhu and Yuille 1996). It is our hope that this article will stimulate further investigations along this direction to build more realistic prior models.

References

[1] M. J. Black and A. Rangarajan, "On the unification of line processes, outlier rejection, ... ", *IJCV* 19(1), 1996.

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

[3] J. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by 2D visual cortical filters. *JOSA*. vol. 2, No. 7, 1985.

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

[5] D.Geiger and A.L.Yuille. "A common framework for image segmentation". *IJCV* 6(3),1991.

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

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

[8] D. Mumford and J. Shah. "Optimal approx. by piecewise smooth functions and associated variational problems." *Comm. Pure & Appl. Math.*, **42**, 1989.

[9] T. Poggio, V. Torre and C. Koch, "Comput. vision and regularization theory", *Nature*, vol. 317, 1985.

[10] D. Terzopoulos, "Multilevel computational processes for visual surface reconstruction". *CVGIP*, **24**, 1983.

[11] K. Wilson, "The renormalization group: critical phenonmena and the Knodo problem," Rev. Mod. Phys., Vol.47, 1975.

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

[13] S. C. Zhu and A. L. Yuille "Region Competition: unifying snakes, ...". *IEEE Trans.on PAMI.* Vol.18, No.9, 1996.

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

[15] S. C. Zhu, and D. B. Mumford., "Prior learning and Gibbs reaction-diffusion", *Technical report*, 1997. To appear in PAMI.