Towards a Mathematical Theory of Primal Sketch and Sketchability

Cheng-en Guo[1], Song-Chun Zhu[1,2] and Ying Nian Wu[2]
[1]Computer Science Department
[2]Statistics Department
University of California, Los Angles
Los Angeles, CA 90095

Abstract

In this paper, we present a mathematical theory for Marr's primal sketch. We first conduct a theoretical study of the descriptive Markov random field model and the generative wavelet/sparse coding model from the perspective of entropy and complexity. The competition between the two types of models defines the concept of "sketchability", which divides image into texture and geometry. We then propose a primal sketch model that integrates the two models and, in addition, a Gestalt field model for spatial organization. We also propose a sketching pursuit process that coordinates the competition between two pursuit algorithms: the matching pursuit [8] and the filter pursuit [12], that seek to explain the image by bases and filters respectively. The model can be used to learn a dictionary of image primitives, or textons in Julesz's language, for natural images. The primal sketch model is not only parsimonious for image representation, but produces meaningful sketches over a large number of generic images.

1. Introduction

Objects in natural scenes appear at a wide range of scales and our perception changes over distance. Looking at Fig.1, we perceive merely a texture impression for the twigs and leaves at far distance, but as they appear in larger size or near distance, we start to notice the individual twigs and then see the shapes of branches and trunks. In this paper, we adopt an artist's notion by calling the image portion with distinguishable elements as *sketchable*, e.g., represented by primitive shapes, and the portion without distinguishable elements is said to be *non-sketchable*. When we walk in such a scene, the resolutions and scales may evolve continuously in raw images (at retina) as a Gaussian pyramid representation could account for. However, abrupt model switching must be happening in our inner perception (at visual cortex), which are quantum jumps between the percepts of texture and shape/geometry.

Now we come to a long standing puzzle in vision: how do we divide texture and geometry? In other words, can we define a mathematical quantity for "sketchability"? Furthermore, as the sketchable and non-sketchable portions are intimately blended in generic images, what is a generic image model that can account for both patterns seamlessly? These questions are fundamental to vision, and the answers to these questions have significant implications to a range of other important vision problems raised at three levels of studies: mathematics, neuroscience, and psychophysics.

Figure 1: Scaling and sketchability illustration.

1. Markov random field theory vs. wavelet/sparse coding theory. There are two prevailing mathematical theories for image modeling. One is the descriptive Markov random field (MRF) theory originated from statistical mechanics. It represents a visual pattern by pooling the responses of a bank of filters over space and the statistics of the responses define a so-called Julesz ensemble[11] – a perceptual equivalence class, which is in turn equivalent to the MRF models [12]. The second theory for image modeling is generative wavelet/sparse coding theory originated from harmonic analysis. It represents images by elements selected from a dictionary of image bases (primitives or token) [10] like wavelets [3], ridgelets [1] etc. However, the two theories are almost disjoint at present. By defining sketchability, we shall quantify the regimes that the two theories work the best and integrate them seamlessly.

2. Filters vs. bases in V1. It is well known that V1 cells in primate visual cortex have Gabor like functions, but it is puzzling what roles the cells play in visual representation, as a Gabor function can be used as *filters* for pooling in-

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

a. Input b. primal sketch c. sketchable $K = 230$ d. synthesis e. primitives

Figure 2: **Sketching pursuit**. (a) Input image. (b) Primal sketch with each element represented by bar or circle. (c) Sketchable pixels (25%) generated by placing primitives at the elements. (d) Synthesized image with non-sketchable pixels filled in by texture using sketchable pixels as boundary condition. (e) Collection of primitives or textons.

formation to form the texture perception [7] or be used as a linear *base* for representing the image primitive [10]. We believe that the V1 cells can switch between the two roles when a critical sketchability condition occurs.

3. Marr's primal sketch. In his monumental book [9], Marr proposed primal sketch by extending a number of psychological theories, such as texture/texton theory [6] and Gestalt psychology. The primal sketch was supposed to be a first level inner representation of generic images, in terms of image primitives, such as bars, edges, terminators etc. However, despite many inspiring observations, Marr provided neither an explicit mathematical model nor a rigorous definition for the dictionary of textons and primitives. Thus a solid model for primal sketch still eludes us.

Image Size	300*240
Sketch Pixels	18,185 $\approx$ 25%
Primitive Number	230
Primitive Width	7
Primitive Parameters	2,350 $\approx$ 3.5%
MRF parameters	5*7*13 = 455

Table 1: The parameters in our primal sketch model for the horse riding image in Fig.2.

In this paper, we propose a mathematical theory of primal sketch and define sketchability. The theory consists of four components.

(1). A primal sketch model for natural images, which integrates the MRF and wavelet theories. An example is illustrated in Fig.2 and Table 1. The input image is of 300×240 pixels, of which $18,185$ pixels (or 25%) are considered by our model as "sketchable" and are shown in Fig.2.c, which is modeled by $2,350$ parameters. They are mostly located at the object boundaries as well as distinguishable features on objects. Each element is represented by a primitive selected from a dictionary (Fig.2.e).

Then the non-sketchable pixels (75%) are modeled by MRF models which condition on the sketchable pixels — so it is seamless. As all distinguished features are explained by the image primitives, the texture can be modeled and synthesized easily with a number of 455 parameters or less. The parameters are for 7 filters for 5 image patches and each pools a 1D histogram of filter responses in 13 bins. The synthesized image in Fig.2.d shows vivid texture, e.g., the grass under the fence. This model is parsimonious and the sketch in Fig.2.b capture the essence of perception. Similar results have been obtained for a large set of generic images.

(2). A definition of sketchability is given based on the log likelihood ratio between the MRF model and linear sparse coding model. Intuitively, a position with strong filter response is selected as sketchable primitive, whereas an area with many weak filter responses are pooled to represent texture.

(3). A sketching pursuit process, which combines the matching pursuit procedure (Mallat and Zhang, 1993) for the sketchable part by adding one base at a time, and the filter pursuit procedure (Zhu, Wu, and Mumford 1997) for the non-sketchable part by adding one filter at a time. The matching pursuit and filter pursuit compete to explain the image in the sense of maximizing the log-likelihood under the guidance of the above sketching model.

(4). Learning a dictionary of primitives (or textons) in image sketch. Traditional image bases, such as Gabor, DoG, LoG, are found to be ineffective as the matching pursuit example shows in Fig.3. Our model can be used to learn a dictionary of primitives and textons (see Fig.6) from a large number (> 100) of images.

2. The Primal Sketch Model

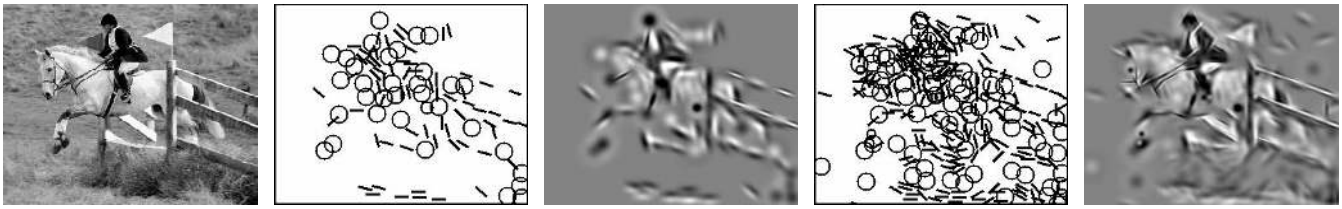
2.1. Two mathematical schemes for modeling

The objective of image modeling can be formulated as seeking a model $p(\mathbf{I})$ that approaches an underlying frequency $f(\mathbf{I})$ of the natural image ensemble, in the sense of minimizing the Kullback-Leibler divergence or MLE from an observed image $\mathbf{I}$ on sufficiently large lattice $\Lambda \to \mathbf{Z}^2$

$$p^* = \arg\min_{\Omega} KL(f||p) = \arg\max_{\Omega} \log p(\mathbf{I}), \;\; \mathbf{I} \sim f(\mathbf{I}).$$

p is pursued in a series of nested probability families which are general enough to approximate f to any precision.

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE



a. input b. tokens $K = 100$ c. image $K = 100$ d. tokens $K = 300$ e. image $K = 300$.

Figure 3: **Matching pursuit** (Mallat and Zhang, 1993): (b) and (d) are symbolic sketch using bars and circles for the K pursued Gabor and LoG bases from dictionary $\mathcal{D}_B$. (c) and (e) are the reconstructed images with $K = 100$ and $K = 300$ bases respectively. These bases not only miss the texture and blur object boundaries, but also do not summarize our percept.

$$\Omega_0 \subset \Omega_1 \subset \cdots \subset \Omega_K \to \Omega_f.$$

There are two mathematical theories/schemes for augmenting the probability family Ω, which we shall combine into our model.

1. The generative scheme and wavelet theory. Generative models interpret images by hidden "causes" of familiar structures, such as sparse coding [10], or wavelet [3]. They adopt a dictionary of base functions, $\mathcal{D}_B = \{\mathbf{b}_{x,y,l,r} : \forall x, y, l, r\}$.Each base $\mathbf{b}$ represents a known image structure and has x, y for its center, l for its type, such as Gabor, LoG, DoG, and r for attributes like orientation and scale.

Then the model interprets an image as a linear superposition of K base functions selected from the dictionary,

$$\mathbf{I} = \sum_{i=1}^{K} \alpha_i \mathbf{b}_i + \mathbf{n}, \;\; \mathbf{b}_i \in \mathcal{D}_B, \tag{1}$$

where α_i is the coefficient of $\mathbf{b}_i$ and $\mathbf{n}$ is the residue image.

Let $\mathbf{B} = \{\mathbf{b}_1, ..., \mathbf{b}_K\}$ and $\boldsymbol{\alpha} = \{\alpha_1, ..., \alpha_K\}$ denote the selected bases and coefficients.

With $\mathbf{B}$ and $\boldsymbol{\alpha}$, we write the model of $\mathbf{I}$ as

$$p(\mathbf{I}; \mathbf{B}, \boldsymbol{\alpha}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-||\mathbf{I} - \sum_{i=1}^{K} \alpha_i \mathbf{b}_i||^2 / 2\sigma^2\}. \tag{2}$$

We define the penalized log-likelihood as

$$l_{\text{sk}}(\mathbf{B}, \boldsymbol{\alpha}, \sigma^2) = \log p(\mathbf{I}; \mathbf{B}, \boldsymbol{\alpha}, \sigma^2) - \epsilon K. \tag{3}$$

The penalty term can be interpreted from the AIC, MDL, and Bayesian perspectives. From MDL perspective, ϵ can be viewed as the coding cost for each base. Then $\mathbf{B}, \boldsymbol{\alpha}, \sigma^2$ can be estimated by maximizing $l_{\text{sk}}(\mathbf{B}, \boldsymbol{\alpha}, \sigma^2)$. We define $\hat{l}_{\text{sk}}$ be the maximum. $-\hat{l}_{\text{sk}}$ can be interpreted as the minimum coding length using this generative wavelet/sparse coding regime.

This scheme augments the model family Ω by increasing the number of variables,i.e., K. A greedy, but powerful, algorithm is matching pursuit[8]. For each step, it selects a base from the dictionary $\mathcal{D}_B$ so that the reconstruction error decreases the most, or equivalently, the likelihood $l(\mathbf{B}, \boldsymbol{\alpha})$ increases the most. It starts from the white noise model.

At step $K + 1$, it chooses a new base $\mathbf{b}_{K+1} \in \mathcal{D}_B$, and thus augments the set from $\mathbf{B}$ and $\boldsymbol{\alpha}$ to

$$\mathbf{B}_+ = \mathbf{B} \cup \{\mathbf{b}_{K+1}\}, \;\text{ and }\; \boldsymbol{\alpha}_+ = \boldsymbol{\alpha} \cup \{\alpha_{K+1}\}.$$

The increase of the log-likelihood is

$$\Delta_B(K+1) = l(\mathbf{B}_+, \boldsymbol{\alpha}_+) - l(\mathbf{I}; \mathbf{B}, \boldsymbol{\alpha}). \tag{4}$$

In fact,

$$\Delta_B(K+1) = \alpha_{K+1}^2/2\sigma^2 - \epsilon = < \mathbf{b}_{K+1}, \mathbf{n} >^2 /2\sigma^2 - \epsilon. \tag{5}$$

It stops when $\Delta_B(K+1) < 0$. Therefore, **Matching pursuit criterion** is to choose

$$(\mathbf{b}_{K+1}, \alpha_{K+1})^* = \arg\max \Delta_B(K+1). \tag{6}$$

Fig.3 displays an example of matching pursuit. Two drawbacks are obvious. (1) The bases miss the texture and blur object boundaries. (2) The symbolic sketches do not represent the objects well, compared with Fig.2.b

2. The descriptive scheme and MRF theory. Descriptive image models interpret images at the signal level, e.g., raw pixel intensities, by pooling statistics using features, e.g., filters. A typical work is the FRAME model for texture[12], which pools feature statistics by a dictionary of filters $\mathcal{D}_F = \{F_{x,y,l,r} : \forall x, y, l, r\}$. Interestingly these filters have exactly the same form as the image bases in dictionary $\mathcal{D}_B$, but they play a different role.

Suppose we select a set of K filters $\mathbf{F} = \{F_1, ..., F_K\} \subset \mathcal{D}_F$. For each filter, we compute a 1D histogram by pooling the filter responses around a local neighborhood $\partial(x, y)$ on image $\mathbf{I}$, and denote it by

$$h_{x,y,l,r}(z) = \sum_{(\xi,\eta)\in\partial(x,y)} \omega_{\xi,\eta} \delta(z - < F_{\xi,\eta,l,r}, \mathbf{I} >).$$

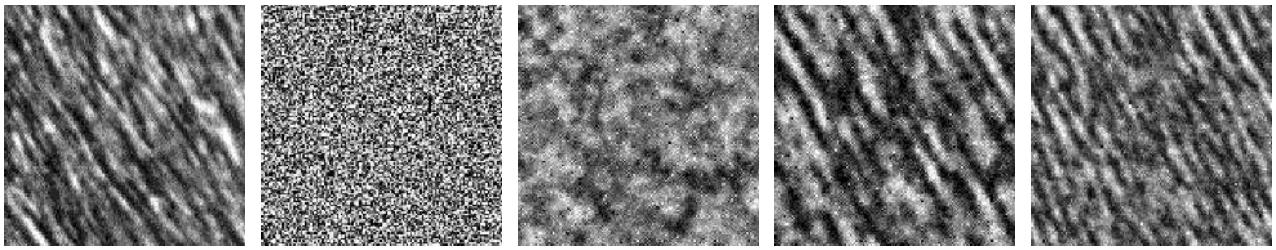
$\delta()$ is the Dirac delta function, and $\sum \omega_{\xi,\eta} = 1$ are the weights. We keep the location index x, y for inhomogeneous patterns. In practice, we may divide the image into several regions by clustering the histograms. We denote the histograms by vectors (after discretization)

$$\mathbf{h}_{x,y,l,r} = h_{x,y,l,r}(z), \;\; \mathbf{H} = \{\mathbf{h}_{x,y,l,r} : \forall F_{x,y,l,r} \in \mathbf{F}\}.$$

The statistics $\mathbf{H}$ define a Julesz ensemble[11]

$$\Omega(\mathbf{H}) = \{\mathbf{I} : h(\mathbf{I}; x, y, l, r) = \mathbf{h}_{x,y,l,r}, \forall F_{x,y,l,r} \in \mathbf{F}\}.$$

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE



a. input b. $K = 0$ c. $K = 2$ d. $K = 3$ e. $K = 7$

Figure 4: **Filter pursuit** (Zhu, Wu, Mumford,1997): adding K filters in a FRAME model by minimax entropy. The filters pool the statistical information (histograms) to yield a texture impression – we call it perceptually equivalent reconstruction.

When K is large or $\mathbf{H}$ is sufficient, any image $\mathbf{I} \in \Omega(\mathbf{H})$ from the ensemble is said to be perceptually equivalent reconstruction to the observed image. The Julesz ensemble is shown to be equivalent to the FRAME model of Gibbs form[11]

$$p(\mathbf{I}; \mathbf{F}, \boldsymbol{\beta}) = \frac{1}{Z} \exp\{ \sum_{F_{x,y,l,r} \in \mathbf{F}} \beta_{x,y,l,r}(< F_{x,y,l,r}, \mathbf{I} >)\}. \quad (7)$$

The potential $\beta_{x,y,l,r}(z)$ is a 1D function on the filter response $z =< F_{x,y,l,r}, \mathbf{I} >$ and it depends on location for inhomogeneous patterns. We denote them by vector $\boldsymbol{\beta}_{x,y,l,r}$ and collect them in $\boldsymbol{\beta} = \{\boldsymbol{\beta}_{x,y,l,r} : \forall F_{x,y,l,r} \in \mathbf{F}\}$. $\boldsymbol{\beta}$ is a dual representation of the pooled statistics $\mathbf{H}$.

Similar to generative model, we define the penalized likelihood

$$l_{\rm nsk}(\mathbf{F}, \boldsymbol{\beta}) = \log p(\mathbf{I}; \mathbf{F}, \boldsymbol{\beta}) - \epsilon K. \quad (8)$$

The ϵ can be interpreted as the coding cost for each filter. $\mathbf{F}$ and $\boldsymbol{\beta}$ can be estimated by maximizing $l_{\rm nsk}(\mathbf{F}, \boldsymbol{\beta})$. We define $\hat{l}_{\rm nsk}$ as the maximum.

This descriptive scheme augments the model family Ω by increasing the number of filters K and statistics . It starts from the white noise model. For each step it selects a filter from the dictionary $\mathcal{D}_F$ so that $l(\mathbf{F}, \boldsymbol{\beta})$ increases the most.

At step $K + 1$, it chooses a new filter $F_{K+1} \in \mathcal{D}_F$, and thus augments the sets from $\mathbf{F}$ and $\boldsymbol{\beta}$ to

$$\mathbf{F}_+ = \mathbf{F} \cup \{F_{K+1}\}, \text{ and } \boldsymbol{\beta}_+ = \{\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K, \boldsymbol{\beta}_{K+1}\}.$$

The increase of the log-likelihood is

$$\Delta_F(K + 1) = \log p(\mathbf{I}; \mathbf{F}_+, \boldsymbol{\beta}_+) - \log p(\mathbf{I}; \mathbf{F}, \boldsymbol{\beta}). \quad (9)$$

Let $\mathbf{J} \in \Omega(\mathbf{H})$ be a "reconstruction" image with K filters, and $\mathbf{h}_{K+1}(\mathbf{J})$ is the local histogram of $\mathbf{J}$ pooled by the new filter F_{K+1}. Let $\mathbf{h}_{K+1}(\mathbf{I})$ be the local histogram of F_{K+1} on the original image $\mathbf{I}$. Then approximately,

$$\Delta_F(K+1) \approx \frac{1}{2}(\mathbf{h}_{K+1}(\mathbf{J}) - \mathbf{h}_{K+1}(\mathbf{I}))^T V^{-1} (\mathbf{h}_{K+1}(\mathbf{J}) - \mathbf{h}_{K+1}(\mathbf{I}))$$

is the "reconstruction" error measured by F_{K+1}, where V is the variance of $\mathbf{h}_{K+1}$ (conditioning on histograms of previous filters). V can be estimated from $\mathbf{J}$.

Therefore, **Filter Pursuit criterion** is to choose

$$(F_{K+1}, \boldsymbol{\beta}_{K+1})^* = \arg\max \Delta_F(K + 1). \quad (10)$$

It stops when $\Delta_F(K + 1) < 0$.

Fig.4 displays an example of filter pursuit on homogeneous texture. With $K = 0$, the sample image is white noise. With $K = 7$ filters, the sampled image in Fig.4.e is perceptually equivalent to the input image. The main drawbacks of this model are (1) difficulties in synthesizing sharp features, like shapes and geometry –i.e., the "sketchable" features, and (2) computational complexity when filters of large window sizes are selected for sketchable features.

For computational efficiency, we can use the following approximation

$$\log p(\mathbf{I}; \mathbf{F}, \boldsymbol{\beta}) \approx l_0 + \frac{1}{2}(\mathbf{H} - \mathbf{H}_0)^T V^{-1} (\mathbf{H} - \mathbf{H}_0), \quad (11)$$

where l_0 is the log-likelihood of the featureless white noise model, $\mathbf{H}_0$ is the histograms computed from the white noise model, and V is the variance-covariance matrix of $\mathbf{H}$ that can be estimated from the observed image.

2.2. Two entropy regimes and sketchability

It comes to our attention that the two schemes are effective on two different entropy (complexity) regimes which are complementary to each other. Thus by integrating the two schemes, we can remove their "blind spots", and resolve the representational and computational difficulties in both schemes.

The FRAME/Julesz model targets the high entropy regime, whereas sparse coding targets the low entropy regime. This is revealed in the following two propositions.

Proposition 1: *Let $f(\mathbf{I})$ be the true distribution that generates* $\mathbf{I}$*, $p(\mathbf{I})$ the FRAME model approching $f(\mathbf{I})$ by minimizing $KL(f||p)$. Then*

$$KL(f \,||\, p) = \text{entropy}(p) - \text{entropy}(f) \geq 0.$$

That is, the entropy of the fitted FRAME model is always no less than the entropy of f.

We rewrite reconstruction eqn (1) in a matrix form $\mathbf{J} = BA$ and $\mathbf{I} = \mathbf{J} + \mathbf{n}$. The images $\mathbf{I}$ and $\mathbf{J}$ are $|\Lambda| \times 1$ vector and B is a $|\Lambda| \times |\mathcal{D}_B|$ constant matrix with each column being a base function in $\mathcal{D}_B$, and A is the $|\mathcal{D}_B| \times 1$ vector for the coefficients. Due to sparsity, elements in A are mostly close to zero except $|\boldsymbol{\alpha}| = O(|\Lambda|/100)$ elements are non-zero. Thus $p(A) = p(\boldsymbol{\alpha})$ has very low entropy, suppose we bound it by $\text{entropy}(p(A)) \leq C$.

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

Proposition 2: *In sparse coding model, with* $\mathbf{J} = BA$,

$$\text{entropy}(p(\mathbf{J})) + \text{entropy}(p(A \mid \mathbf{J})) = \text{entropy}(p(A)) + \frac{1}{2}\log\det(BB').$$

That is, the resulting $p(\mathbf{J})$ has low entropy bounded by $C + \text{const}$, and it cannot account for the images generated from $f(\mathbf{I})$ whose entropy is larger than this bound. In other words, the sparse coding model puts all the non-sketchable patterns into the residue $\mathbf{n}$. This is very obvious in Fig. 3 where the texture background are not captured by the bases at all.

Furthermore, if the sparse coding model is forced to reconstruct the texture $\mathbf{n}$, then it becomes non-sparse and the computation in that regime is NP-complete. So it is computationally impaired as well in this regime.

To summarize, the sparse coding (wavelet) model is effective in low entropy regime where images have order and structures, such as the shape and geometry. We call this regime as "sketchable". The FRAME model is effective in high entropy regime where images have less structures, such as stochastic texture. We call this regime as "non-sketchable".

The intrinsic connection between the two models are revealed by the following proposition. Here a filter is the same as a base $F_{x,y,l,r} = \mathbf{b}_{x,y,l,r}$.

Proposition 3: *The FRAME model* $p(\mathbf{I}; \mathbf{F}, \boldsymbol{\beta})$ *is the equilibrium distribution of the following PDE*

$$d\mathbf{I}(t) = \frac{1}{2}\sum_{\mathbf{b}\in\mathbf{F}} \beta'_{x,y,l,r}(< \mathbf{I}(t), \mathbf{b}_{x,y,l,r} >) dt \times \mathbf{b}_{x,y,l,r} + d\mathbf{n}(t),$$

where $\beta'()$ *is the derivative of the potential function.*

In this dynamics, each step is a linear superposition of bases $\mathbf{b}_{x,y,l,r}$, plus a small Brownian noise $d\mathbf{n}(t)$. This addtive form coincides with sparse coding model in eqn (1). The difference is that this dynamics is iterative and non-sparse.

The following theorem tells us the behavior of the entropy with the change of scale/distance.

Theorem: *Let* $\mathbf{I}$ *be an image defined on a lattice* Λ*, let* $p(\mathbf{I})$ *be the probability density of* $\mathbf{I}$*. Let* $\mathbf{I}_s$ *be a sub-sampled version of* $\mathbf{I}$ *defined on the smaller lattice* Λ_s*. Then 1). the overall entropy of* $\mathbf{I}_s$ *is smaller than or equal to the entropy of* $\mathbf{I}$*.*

$$\text{entropy}(p(\mathbf{I}_s)) \leq \text{entropy}(p(\mathbf{I})).$$

2). the entropy rate (i.e., entropy per pixel) of $\mathbf{I}_s$ *is larger than or equal to the entropy rate of* $\mathbf{I}$*.*

$$\text{entropy}(p(I_s))/|\Lambda_s| \geq \text{entropy}(p(\mathbf{I}))/|\Lambda|.$$

This suggests that when we look at an image at higher resolution, we see more things (i.e., overall entropy increases), but at the same time, things are getting sparser (i.e., entropy rate decreases). This suggests that we need an overcomplete dictionary to achieve more sparsity.

Definition of sketchability: *Let* $\mathbf{I}$ *be a local image patch. Let* $\hat{l}_{\text{nsk}}$ *be the penalized log-likelihood of the fitted FRAME model, and let* $\hat{l}_{\text{sk}}$ *be the penalized log-likelihood of the fitted linear sparse coding/wavelet model. Then the image patch* $\mathbf{I}$ *is sketchable if* $\hat{l}_{\text{sk}} > \hat{l}_{\text{nsk}}$*. The sketchability is defined as* $\hat{l}_{\text{sk}} - \hat{l}_{\text{nsk}}$*.*

Perhaps the most interesting situation is where the sketchability is around 0. For example, when we are getting close to a tree, there is a distance where the pattern of twigs starts to jump back and forth between being non-sketchable and being sketchable.

2.3. The primal sketch model

In light of the above analysis, we adopt the following primal sketch model for early vision. It integrates and improves the sparse coding and MRF in the following way.

1. The image lattice is divided, automatically, into two disjoint parts: the sketchable Λ_{sk} where structures occur and the non-sketchable Λ_{nsk}.

$$\Lambda = \Lambda_{\text{sk}} \cup \Lambda_{\text{nsk}}.$$

2. The sketchable part of $\mathbf{I}$ is explained by the sparse coding model with an unknown number of K_B bases,

$$\mathbf{I}_{\Lambda_{\text{sk}}} = \sum_{i=1}^{K_B} \alpha_i \mathbf{b}_i + \mathbf{n}, \;\; \mathbf{b}_i \in \mathcal{D}_{\text{txn}}.$$

Note that we replace the generic base dictionary $\mathcal{D}_B$ by a texton/primitive dictionary $\mathcal{D}_{\text{txn}}$. Fig. 6 shows some typical examples of the primitives in $\mathcal{D}_{\text{txn}}$. These primitives have much stronger contrast than the Gabors/LoG bases. They are "expert" and one location is explained by one primitive. Thus it will not have the blurring problem of $\mathcal{D}_B$ as Fig. 3 illustrated. In other words, $\mathcal{D}_{\text{txn}}$ gives much sparser representation than $\mathcal{D}_B$, and $\mathcal{D}_{\text{txn}}$ is much larger $\mathcal{D}_B$.

The image follows the generative model in eqn (2),

$$\mathbf{I}_{\text{sk}} \sim p(\mathbf{I}_{\Lambda_{\text{sk}}}; \mathbf{B}, \boldsymbol{\alpha}). \tag{12}$$

Again $\mathbf{B}, \boldsymbol{\alpha}$ denote the sets of bases and coefficients

$$\mathbf{B} = \{\mathbf{b}_1, ..., \mathbf{b}_{K_B}\}, \;\; \boldsymbol{\alpha} = \{\alpha_1, ..., \alpha_{K_B}\}.$$

3. The non-sketchable part of $\mathbf{I}$ is explained by the inhomogeneous Gibbs model (FRAME in eqn.7), using the sketchable part $\mathbf{I}_{\Lambda_{\text{sk}}}$ as boundary conditions

$$\mathbf{I}_{\Lambda_{\text{nsk}}} \sim p(\mathbf{I}_{\Lambda_{\text{nsk}}} | \mathbf{I}_{\Lambda_{\text{sk}}}; \mathbf{F}, \boldsymbol{\beta}). \tag{13}$$

It is specified by the set of K_F filters and potentials,

$$\mathbf{F} = \{F_1, ..., F_{K_F}\}, \;\; \boldsymbol{\beta} = \{\beta_1, ..., \beta_{K_F}\}.$$

The filters are selected from generic dictionary $\mathbf{F} \subset \mathcal{D}_F$ and vary over space for inhomogeneity.

4. In the original sparse coding model, the base locations and coefficients are assumed to be iid. Thus they don't line up well in space, see Fig. 3.b and d. To resolve this problem, we introduce a new concept called **Gestalt field** similar to the mixed random fields[4].

We organize the bases (like tokens) in a 2D attributed graph $G =< V, E >$. Any vertex $v \in V$ is a primitive with

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

two sets of variables: the attributes (α, x, y, l, r) and up to 5 address variables $(a_1, ..., a_5)$ which are either empty or point to neighbors of v and the type of connection, i.e, continuity, junction etc. Suppose we connect all the neighboring pairs (s, t) and their relationship ℓ in edge set

$$E = \{e_j = (\ell_j, s_j, t_j) : s_j, t_j \in V,\ j = 1, 2, ..., N\}.$$

Then we have the Gestalt field,

$$p(\mathbf{B}, \boldsymbol{\alpha}) \propto \exp\{-\lambda_o K_B - \sum_{e_j \in E} \psi_{\ell_j}(s_j, t_j)\}. \quad (14)$$

$\psi_\ell(s, t)$ is the potential function of relationship ℓ on the attributes of two bases, such as their relative distance and angle.

Summarizing the sketchable (for textons), non-sketchable (for texture), and Gestalt field in eqns.12, 13, and 14, we obtain the *primal sketch model*,

$$p(\mathbf{I}; \mathbf{B}, \boldsymbol{\alpha}, \mathbf{F}, \boldsymbol{\beta}) = p(\mathbf{B}, \boldsymbol{\alpha}) p(\mathbf{I}_{\Lambda_{\rm sk}}; \mathbf{B}, \boldsymbol{\alpha}) p(\mathbf{I}_{\Lambda_{\rm nsk}} | \mathbf{I}_{\Lambda_{\rm sk}}; \mathbf{F}, \boldsymbol{\beta}).$$

The above model may need MCMC method for global inference. Here we propose a greedy method — called sketching pursuit. Intuitively, the sketching pursuit process coordinates the competition between filter pursuit and matching pursuit, under the guidance of the primal sketch model.

The sketching pursuit process.

Given the current $\mathbf{B}, \boldsymbol{\alpha}$ and $\mathbf{F}, \boldsymbol{\beta}$,

1. Compute the log-likelihood increase for a primitive $\mathbf{b}^*$
$$\Delta_B = \max_{\mathbf{b} \in \mathcal{D}_{\rm txn}} \log p(\mathbf{I}; \mathbf{B}_+, \boldsymbol{\alpha}_+, \mathbf{F}, \boldsymbol{\beta}) - \log p(\mathbf{I}; \mathbf{B}, \boldsymbol{\alpha}, \mathbf{F}, \boldsymbol{\beta}).$$
2. Compute the log-likelihood increase for a filter F^*
$$\Delta_F = \max_{F \in \mathcal{D}_F} \log p(\mathbf{I}; \mathbf{B}, \boldsymbol{\alpha}, \mathbf{F}_+, \boldsymbol{\beta}_+) - \log p(\mathbf{I}; \mathbf{B}, \boldsymbol{\alpha}, \mathbf{F}, \boldsymbol{\beta}).$$
3. If $\Delta_F > \Delta_B$ and $\Delta_F > \epsilon$, then augment $\mathbf{F}$ and update $\boldsymbol{\beta}$.
$$(\mathbf{F}, \boldsymbol{\beta}) \leftarrow (\mathbf{F} \cup \{F^*\}, \boldsymbol{\beta} \cup \{\beta^*\}).$$
4. If $\Delta_B > \Delta_F$ and $\Delta_B > \epsilon$, then augment $\mathbf{B}$ and $\boldsymbol{\alpha}$.
$$(\mathbf{B}, \boldsymbol{\alpha}) \leftarrow (\mathbf{B} \cup \{\mathbf{b}^*\}, \boldsymbol{\alpha} \cup \{\alpha^*\}).$$
5. Stop if $\Delta_B < \epsilon$ and $\Delta_F < \epsilon$. Otherwise repeat 1-4.

3. Algorithms and experiments

3.1. Sketching pursuit and refinement

In the above sketching pursuit process, we expect that matching pursuit will win first, and after most of the sketchable parts get sketched, filter pursuit will start to win, and the non-sketchable parts will be filled in. In this paper, we adopt the following algorithms in our experiment.

First, we whiten the image, so that the texture backgrounds are closer to white noise, thus we can concentrate on the sketchable parts by a simplified and fast sketching pursuit process. Recall that in the matching pursuit algorithm, all candidate bases $\mathbf{b} \in \mathcal{D}_B$ are ranked from high to low according to their coefficient $\alpha =< \mathbf{n}, \mathbf{b} >$. The base with the highest coefficient is selected each time. In contrast, in the sketching pursuit, each existing primitive in the current set $\mathbf{B}$, i.e., $\mathbf{b}_{x,y,l,r} \in \mathbf{B}$, will erect some preferences in their neighborhood. For a candidate primitive $\mathbf{b}'$ in its neighborhood, we measure the possible type of neighboring relationship

$$\ell^* = \arg\max_{\ell} \psi_\ell(\mathbf{b}, \mathbf{b}').$$

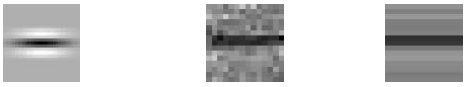
For example, $\mathbf{b}'$ and $\mathbf{b}$ form a junction or continuity etc. Then we rank a combination of $\psi_{\ell^*}(\mathbf{b}, \mathbf{b}')$ and the coefficient of $\mathbf{b}'$. Thus a primitive that fits to existing primitives are given priority. This is consistent with the primal sketch model $p(\mathbf{I}; \mathbf{B}, \boldsymbol{\alpha}, \mathbf{F}, \boldsymbol{\beta})$.

The above fast sketching process is refined by a second step. We replace the selected base functions by local segments of ridge functions estimated from the image patch under the base window. We compute the profiles of primitives by averaging along their main axis. The primitive windows can extend or shorten slightly for better alignment. Some gaps are filled and some isolated primitives are removed to further minimize the energy of the Gestalt field.

Fig.8 shows two examples of the refinement effects. We represent an elongated primitive by a line segment, Fig.8.b is the results of the fast sketching process, which are refined in Fig.8.c.

After the sketchable parts are identified, we then target the non-sketchable parts. We first cluster the non-sketchable pixels into a number of clusters or patches based on local histograms of filter responses. Then we model each patch by a Julesz ensemble or equivalently MRF model, and synthesize the non-sketchable parts by randomly sampling from the Julesz ensembles.

3.2. Dictionary of image primitives or textons



(a) Gabor Cosine (b) local image (c) primitive

Figure 5: The comparison of a Gabor Cosine base, the original local image which the Gabor base is supposed to fit, and a primitive in the texton dictionary.

Our model can also be used to learn a large dictionary $\mathcal{D}_{\rm txn}$ of image primitives from a larger number of natural images. Some are shown in Fig. 6. We use a simple algorithm along with the sketching refinement process. We start with $\mathcal{D}_{\rm txn} = \mathcal{D}_B$. After obtaining a raw sketch and refine the windows, we collect the windowed image primitive covered by a base $\mathbf{b} \in \mathcal{D}_B$. If $\mathbf{b}$ is an elongated base,

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

like Gabor cosine and Gabor sine, we average the intensities of the primitive along the main axis. If $\mathbf{b}$ is an isotropic base, like LoG or DoG, we average the intensities circularly around its center. Then we add the image primitive to $\mathcal{D}_{\text{txn}}$.

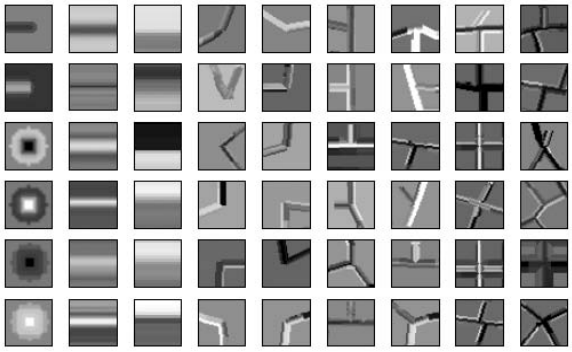
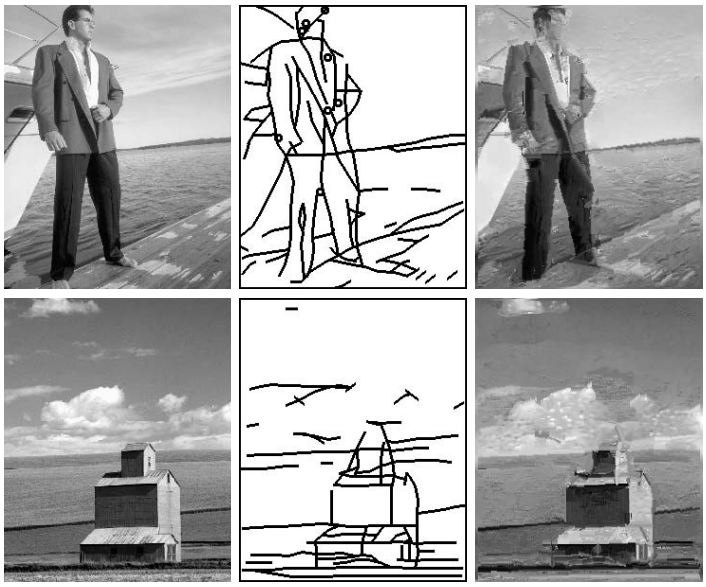


Figure 6: The primitives (or textons) collected across images. They are isotropic blobs, bars, step edges, (L, T, Y)-junctions, and crosses.

Fig.5 shows a typical example. The image structure – a bar is shown in Fig.5.b. It is captured by a Gcos base (see Fig.5.a) in matching pursuit. Fig.5.c is the image primitive with its horizontal pixel intensities averaged along the axis. In out experiment all primitives are 5 to 7 pixels in width. The junctions with degrees larger than 2 are obtained from the Gestalt field.



(a) (b) (c)

Figure 7: Results of the primal sketch model. (a) input image; (b) sketching refinement results; (c) synthesized image from the model $p(\mathbf{I}; \mathbf{B}, \boldsymbol{\alpha}, \mathbf{F}, \boldsymbol{\beta}))$.

We shall learn $\mathcal{D}_{\text{txn}}$ in a more rigorous setting under an EM-type learning strategy in future work.

4. Discussion

In this paper, we present a primal sketch model that integrates three components: a descriptive texture model (Julesz ensemble), a generative model with image primitives (textons), and a Gestalt field.

The integration of the two modeling schemes merged two powerful methodologies: MRF and wavelet theories. This naturally leads to the answer for sketchability as a model switching/selection problem in statistics.

Our work is interestingly related to the inpainting work (Chan and Shen, 01 and others) which adopts an PDE for filling in scrached pictures. The inpainting work is a variational method for minimizing the smoothness term. Our method is more general in the potential formulation and simulates the texture by sampling, instead of maximization.

References

[1] E. J. Cands and D. L. Donoho, "Ridgelets: a key to higher-dimen. intermittency?" *Phil. Trans. R. Soc. Lond. A.*, 357, 2495-509, 1999.

[2] T. Chan and J. Shen, "Local inpainting model and TV inpainting", *SIAM J. of Appl. Math*, 62:3, 1019-43, 2001.

[3] R.R. Coifman and M.V. Wickerhauser, "Entropy based algorithms for best basis selection." *IEEE Info. Theory.*, 38, pp713-18, 1992.

[4] A. Fridman, "Mixed Markov Models", *Ph.D. Thesis, Department of Mathematics, Brown University*, 2000.

[5] D. J. Heeger and J. R. Bergen, "Pyramid Based Texture Analysis/Synthesis", *Computer Graphics Proc.*, 1995.

[6] B. Julesz, "Textons, the elements of texture perception and their interactions", *Nature*, Vol. 290, pp. 91-97, 1981.

[7] J. Malik and R. Perona, "Preattentive Texture Discrimination with Early Vision Mechanisms", *J. Opt. Soc. AM* , Vol. 7, No. 5, pp. 923-932, 1990.

[8] S. Mallat and Z. Zhang, "Matching Pursuit in a Time-Frequency Dictionary", *IEEE Sig. Proc.*, 41, 3397-415, 1993.

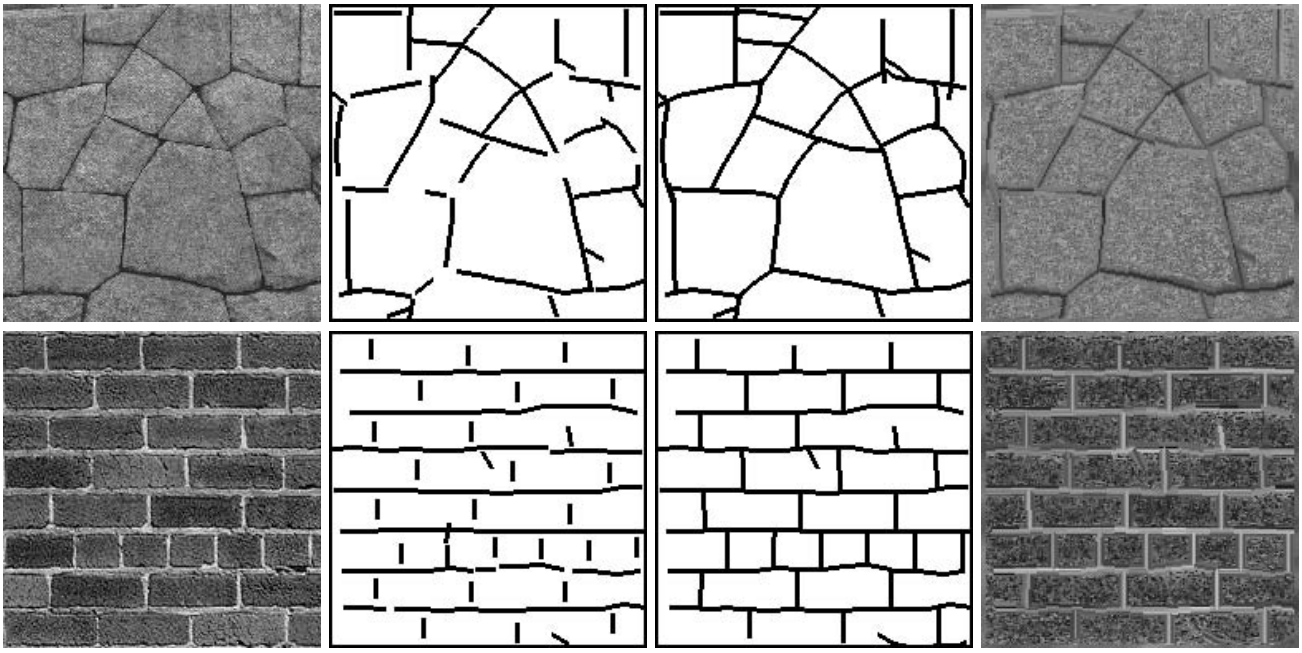
[9] D. Marr, *Vision,* W. H. Freeman and Company, 1982.

[10] B. A. Olshausen and D. J. Field, "Emergence of Simple-cell Receptive Field Properties by Learning a Sparse Code for Natural Images," *Nature*, Vol. 381, pp. 607-609, 1996.

[11] Y. N. Wu, S. C. Zhu, and X. Liu, "Equivalence of Julesz and Gibbs Ensembles", *Proc. of ICCV*, Corfu, Greece, 1999.

[12] S. C. Zhu, Y. N. Wu, and D. Mumford, "Minimax Entropy Principle and Its Applications in Texture Modeling", *Neural Computation*, 9(8), 1627-1660, 1997.

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE



(a) (b) (c) (d)

Figure 8: Results of the primal sketch model. (a) input image; (b) fast sketching pursuit result; (c) sketching refinement result; (d) synthesized image from the model $p(\mathbf{I}; \mathbf{B}, \boldsymbol{\alpha}, \mathbf{F}, \boldsymbol{\beta}))$.

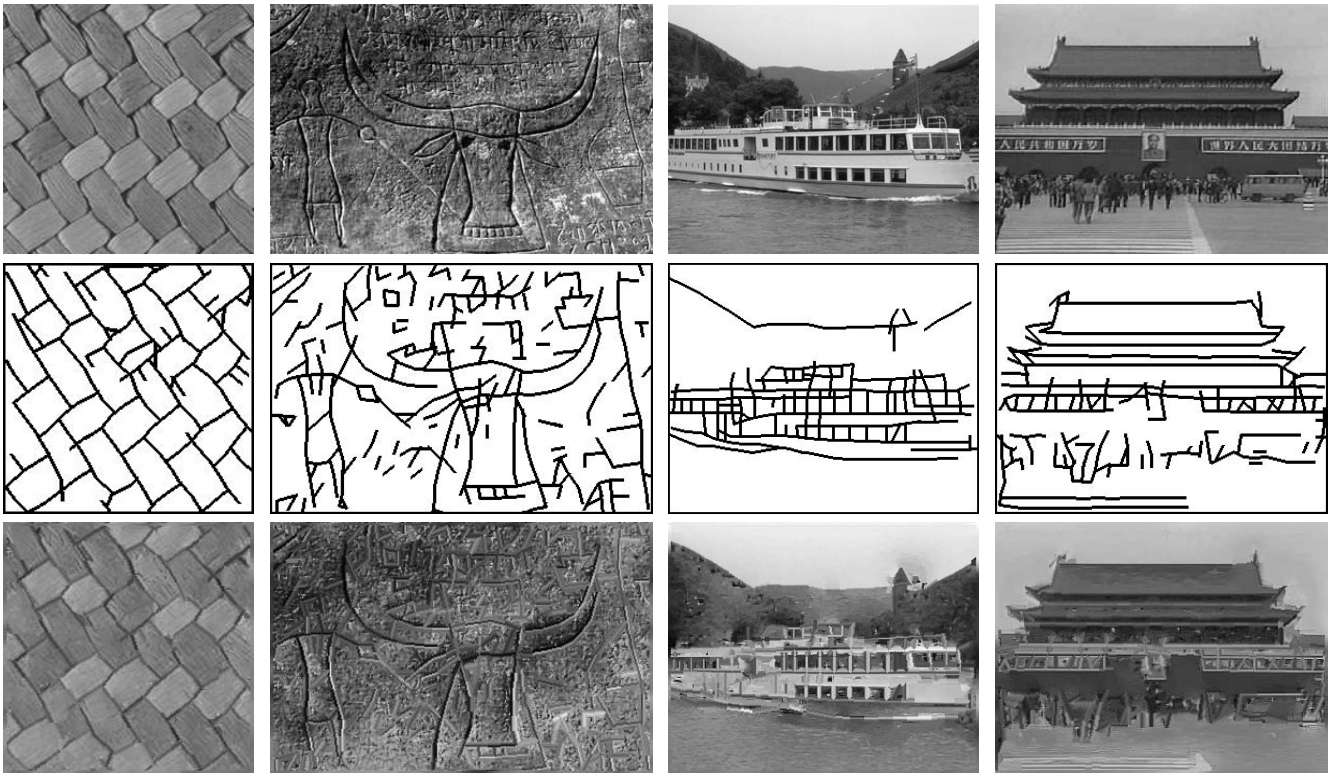


Figure 9: Results of the primal sketch model. First row: input image; Second row: sketching refinement result; Last row: synthesized image from the model $p(\mathbf{I}; \mathbf{B}, \boldsymbol{\alpha}, \mathbf{F}, \boldsymbol{\beta}))$.

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE