# ADAPTIVE, NONPARAMETRIC MARKOV MODELS AND INFORMATION-THEORETIC METHODS FOR IMAGE RESTORATION AND SEGMENTATION

by

Suyash P. Awate

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

School of Computing

The University of Utah

December 2006

THE UNIVERSITY OF UTAH GRADUATE SCHOOL

# SUPERVISORY COMMITTEE APPROVAL

of a dissertation submitted by

Suyash P. Awate

This dissertation has been read by each member of the following supervisory committee and by majority vote has been found to be satisfactory.

| | |
|---|---|
| | Chair: Ross T. Whitaker |
| | Christopher R. Johnson |
| | Tolga Tasdizen |
| | Sarang Joshi |
| | Gil Shamir |

THE UNIVERSITY OF UTAH GRADUATE SCHOOL

# FINAL READING APPROVAL

To the Graduate Council of the University of Utah:

I have read the dissertation of Suyash P. Awate in its final form and have found that (1) its format, citations, and bibliographic style are consistent and acceptable; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the Supervisory Committee and is ready for submission to The Graduate School.

Date

Ross T. Whitaker
Chair: Supervisory Committee

Approved for the Major Department

Martin Berzins
Chair/Director

Approved for the Graduate Council

David S. Chapman
Dean of The Graduate School

# ABSTRACT

The regularity in data fundamentally distinguishes itself from random noise. Describing this regularity in generic, yet powerful, ways is one of the key problems in signal processing. One way of capturing image regularity is by incorporating *a priori* information into the image model itself. Approaches extracting such prior information from *training* data have limited utility because of the lack of effective training sets for most applications. *Unsupervised* approaches that, typically, encode prior information via parametric models work best only when the data conforms to that model. Certain kinds of problems do not adhere to strict models, entailing unsupervised approaches to be *adaptive*. Statistical-inference methodologies that allow us to learn the underlying structure and variability in the data form important tools in adaptive signal processing.

This dissertation presents an adaptive Markov-random-field (MRF) image model that automatically learns the local statistical dependencies via data-driven nonparametric techniques. We use this model to create adaptive algorithms for processing images. We incorporate prior information, when available, through optimal Bayesian frameworks. We enforce optimality criteria based on fundamental information-theoretic concepts that capture the functional dependence and information content in the data.

We employ this adaptive-MRF framework for effectively solving several classic problems in image processing, computer vision, and medical image analysis. Inferring the statistical structure underlying corrupted images enables us to restore images without enforcing strong models on the signal. The restoration iteratively improves the predictability of pixel intensities from their neighborhoods, by decreasing their joint entropy. When the nature of noise is known, we present an effective empirical-Bayesian reconstruction strategy. We also present a method to optimally estimate the uncorrupted-signal statistics from the observed corrupted-signal statistics by minimizing a KL-divergence measure. We apply this adaptive-MRF framework to classify tissues in magnetic resonance (MR)

images of the human brain by maximizing the mutual information between the classification labels and image data, capturing their mutual dependency. The generic formulation enables the method to adapt to different MR modalities, noise, inhomogeneities, and partial-voluming. We incorporate *a priori* information via probabilistic brain-tissue atlases. We use a similar strategy for texture segmentation, using fast threshold-dynamics-based level-set techniques for regularization.

# CONTENTS

# CHAPTER 1

# INTRODUCTION

This dissertation is about processing digital images. An image is, essentially, data that are acquired to measure some physical properties of a natural process. Image processing, broadly speaking, deals with the transformation and representation of the information contained in image data. We use the term *image* to mean any scalar or vector-valued function defined on an $n$-dimensional ($n$D) domain. Digital images consist of discrete samples on dense Cartesian grids. We can find several examples of digital images in our day-to-day lives such as digital photographs and videos. Black-and-white photographs consist of scalar data on a 2D grid, while color photographs contain 3D data (the RGB color) on a 2D grid. Color videos are 3D data on a 3D grid where the third grid dimension constitutes time. In the field of medical imaging, magnetic resonance (MR) images can contain scalar, vector, or tensor data on 3D grids. Image processing subsumes a gamut of domains and applications ranging from the low-level tasks of image modeling, restoration, segmentation, registration, and compression to the high-level tasks of recognition and interpretation [65, 81, 25]. Image processing has applications in many fields including computer vision, robotics, and medicine.

The information contained in images manifests itself, virtually always, in some *patterns* evident in the image data. We refer to these patterns as the *regularity* in the data. Describing this regularity in a way that is both general and powerful is one of the key problems in image processing. Typically, we capture this regularity in geometric or statistical terms. We refer to the process of describing regularity in images as *image modeling*. Indeed, the use of the term *modeling* is synonymous with its colloquial meaning of a schematic description of a system that accounts for its known/inferred properties and is used for further study of its characteristics, e.g., an atomic model, an

economic model, etc. In this dissertation, we use the term in the statistical sense of a generative model. Thus, given an image model, we can generate image data that conform to, or are derived from, the model.

Typical image-modeling and processing techniques rely on a wide variety of mathematical principles in the fields of linear systems, variational calculus, probability and statistics, information theory, etc. In this dissertation, we desire algorithms that learn the physical model that generated the data through *statistical inference* methodologies. Observing that the image data always lie on a discrete Cartesian grid, we can model the regularity or the local statistical dependencies in the data through an underlying grid of *random variables* or a *Markov random field* (MRF). Theoretical and applied research over the last few decades has firmly established MRFs as powerful tools for statistical image modeling and processing.

This dissertation deals with several classic problems concerning *restoration* and *segmentation*. Image restoration deals with processing corrupted or *degraded* image data in order to obtain the uncorrupted image. This is typically performed by assuming certain models of the uncorrupted images or the degradation. For instance, image models try to capture the regularity in uncorrupted images. The literature presents different kinds of image models that suit best for different kinds of data. In practice, virtually all image data are degraded to an extent and many image-processing algorithms explicitly account for such degradations. Image segmentation is the process of dividing an image into partitions, or *segments*, where some semantics are associated with each segment.

Many image-processing strategies, including those for restoration and segmentation, make strong statistical or geometric assumptions about the properties of the signal or degradation. As a result, they break down when images exhibit properties that do not adhere to the underlying assumptions and lack the generality to be easily applied to diverse image collections. Strategies incorporating specific models work best when the data conform to that model and poorer otherwise. Models imposing stronger constraints (more restrictive) typically give better results with data conforming to those constraints as compared with weaker more-general models. However, schemes with restrictive models also fare much poorer when the data do not satisfy the model. As we shall see,

many image-processing applications are not inherently conducive to strict models and, therefore, there is a need for generic image models and the associated image-processing algorithms. This dissertation presents a very general image model that adapts its specifications based on the observed data. Subsequently, the dissertation presents effective algorithms for image restoration and segmentation that easily apply to a wide spectrum of images.

One way of capturing image regularity is by incorporating *a priori* information in the image model itself. Some approaches rely on training data to extract prior information that is, in turn, transfused into the model specification. This allows us to learn complex models to which the data truly conform. Effective training sets, however, are not readily available for most applications and, therefore, this calls for *unsupervised* approaches [74]. Unsupervised approaches do not use training exemplars for learning properties about the data. However, they typically encode prior information via *parametric* statistical or geometric models that define the model structure. To refrain from imposing ill-fitting models on the data, unsupervised approaches need to learn the optimal parameter values from the data. As an alternative, unsupervised approaches can also rely on *nonparametric* modeling approaches where even the model structure, together with the associated internal parameters, is determined from the data. In these ways, unsupervised approaches need to be *adaptive* [74]. Adaptive methods automatically adjust their behavior in accordance with the perceived environment by adjusting their internal parameters. They do not impose *a priori* models but rather adapt their behavior, as well as the underlying model, to the data. Therefore, adaptive methods have the potential for being easily applicable to a wide spectrum of image data.

This dissertation uses a statistical MRF model to build adaptive algorithms for image processing. Broadly speaking, a statistical *model* is a set of probability density functions (PDFs) on the sample space associated with the data. Parametric statistical modeling parameterizes this set using a few control variables. An inherent difficulty with this approach is to find suitable parameter values such that the model is well-suited for the data. For instance, most parametric PDFs are unimodal whereas typical practical problems involve multimodal PDFs. *Nonparametric* statistical modeling [48, 171, 156]

fundamentally differs from this approach by not imposing strong parametric models on the data. It provides the power to model and learn arbitrary (smooth) PDFs via data-driven strategies. As we shall see in this dissertation, such nonparametric schemes—that adapt the model to best capture the characteristics of the data and then process the data based on that model—can form powerful tools in formulating unsupervised adaptive image-processing methods.

We exploit the adaptive-MRF model to tackle several classic problems in image processing, medical image analysis, and computer vision. We enforce optimality criteria based on fundamental information-theoretic concepts that help us analyze the functional dependence, information content, and uncertainty in the data. In this way, information theory forms an important statistical tool in the design of unsupervised adaptive algorithms. The adaptive-MRF model allows us to statistically infer the structure underlying corrupted data. Learning this structure allows us to restore images without enforcing strong models on the signal. The restoration proceeds by improving the predictability of pixel intensities from their neighborhoods, by decreasing their joint entropy. When the noise model is known, e.g., MR images exhibit Rician noise, Bayesian reconstruction strategies coupled with MRFs can prove effective. We employ this model for optimal brain tissue classification in MR images. The method relies on maximizing the mutual information between the classification labels and image data, to capture their mutual dependency. This general formulation enables the method to easily adapt to various kinds of MR images, implicitly handling the noise, partial-voluming effects, and inhomogeneity. We use a similar strategy for unsupervised texture segmentation, observing that textures are precisely defined by the regularity in their Markov statistics.

## 1.1 Thesis Overview

The rest of the thesis is organized as follows. Chapter 2 presents a tutorial on general probability theory and statistical inference. It describes the important mathematical concepts, and the notation, concerning nonparametric statistics, information theory, and MRFs that form the foundation of many of the key ideas in this dissertation. The next five chapters give the new ideas and algorithms in this dissertation for several applications.

We present the related work from literature concerning each of these approaches as a part of each of those chapters. Chapter 3 presents the theoretical and engineering aspects of the adaptive-MRF image model. All subsequent chapters present adaptive image-processing methods that rely on this image model. The next two chapters, i.e., Chapters 4 and 5, present algorithms for image restoration in the absence and presence of the knowledge of the degradation process, respectively. Chapter 5 specifically concerns denoising MR images. Chapter 6 presents a method for classifying brain tissues in MR images. The optimality criteria for segmentation in this chapter are applied to texture segmentation in Chapter 7. Chapter 8 summarizes the dissertation and discusses a few directions for extending the work.

# CHAPTER 2

# TECHNICAL BACKGROUND

The ideas in this dissertation rely on fundamental principles in probability, statistics and information theory. This chapter reviews the relevant concepts and establishes the mathematical notation that we will use in the rest of the dissertation.

## 2.1 Probability Theory

Probability theory is concerned with the analysis of random, or chance, phenomena. Such random phenomena, or processes, occur all the time in nature in one form or the other. Pierre Simon de Laplace established the theory of probability in the year 1812, after publishing the *Theorie Analytique des Probabilites*. The theory now pervades a wide spectrum of scientific domains including thermodynamics, statistical mechanics, quantum physics, economics, information theory, machine learning, and signal processing.

Probability theory deals with *random experiments*, i.e., experiments whose outcomes are not certain. The set of all possible outcomes of an experiment is referred to as the *sample space*, denoted by $\Omega$, for that experiment. For instance, let us consider the experiment of picking up a random pixel from an $N \times N$ pixels digital image. The sample space is all possible coordinates of the grid image domain, i.e., $\Omega = \{\{0, 1, 2, \ldots, N-1\} \times \{0, 1, 2, \ldots, N-1\}\}$.

An *event* is a collection of the outcomes in the sample space, or a subset of the sample space. Consider an event $A$ in the sample space $\Omega$. The *probability* $P(A)$ of the event $A$ is the chance that the event will occur when we perform the random experiment. The probability is actually a function $P(\cdot)$ that satisfies the following properties:

$$\forall A, P(A) \quad \geq \quad 0, \tag{2.1}$$

$$P(\Omega) = 1, \tag{2.2}$$

$$P(A \cup B) = P(A) + P(B), \forall A \text{ and } B \text{ such that } A \cap B = \phi, \tag{2.3}$$

where $\phi$ is the empty set.

## 2.2 Random Variables

There are situations where one does not want the information concerning each and every outcome of an experiment. Instead, one is more interested in high-level information. For instance, given a grayscale digital image where each pixel takes one of the 256 values or *intensities*, $\{0, 1, 2, \ldots, 255\}$, one may want to know how many pixels had a particular intensity, rather than which particular pixels had that intensity. The notion of *random variables* helps us extract such information.

The term *random variable* can be a little misleading [167]. A random variable (RV), denoted by $X$, is a mapping, or a *function*, that assigns some real number to each element in the sample space $\Omega$. Thus, an RV is a function, $X : \Omega \rightarrow \Re$, whose domain is the sample space and the range is the set of real numbers [167]. The set of values actually taken by $X$ is typically a subset of $\Re$. When the sample space $\Omega$ is uncountable, or nondenumerable, not every subset of $\Omega$ constitutes an event to which we could assign a probability. This entails the definition of a class $\mathcal{F}$ denoting the class of measurable subsets of $\Omega$. Furthermore, we require that the set $\{\omega \in \Omega : X(\omega) \leq x\}$ be an event, and a member of $\mathcal{F}$, so that we can define probabilities such as $P(X \leq x)$. The collection of entities $(\Omega, \mathcal{F}, P)$ is called the *probability space* associated with the RV $X$. In this dissertation, uppercase letters, e.g., $X$, denote RVs and lowercase letters, e.g., $x$, denotes the value assigned by the RVs.

The *cumulative distribution function* (CDF) $F_X(\cdot)$ of an RV $X$ is

$$F_X(x) = P(X \leq x). \tag{2.4}$$

The CDF satisfies the following properties

$$\forall x \in (-\infty, +\infty), 0 \leq F_X(x) \leq 1, \tag{2.5}$$

$$F_X(x) \text{ is a nondecreasing function of } x, \tag{2.6}$$

$$\lim_{x \to -\infty} F_X(x) = 0, \tag{2.7}$$

$$\lim_{x \to +\infty} F_X(x) = 1. \tag{2.8}$$

The *joint* CDF $F_{X,Y}(\cdot)$ of two RVs $X$ and $Y$ is

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y). \tag{2.9}$$

A *continuous* RV is one whose CDF is a continuous function. A *discrete RV* has a piecewise-constant CDF. Most situations in image processing, and so also in this dissertation, entail the use of continuous RVs. Hence, from now on we focus on continuous RVs and, unless explicitly mentioned, we use to the term RV to refer to a continuous RV.

The *probability density function* (PDF) $P_X(\cdot)$ of an RV $X$ is

$$P_X(x) = \frac{dF_X(x)}{dx}. \tag{2.10}$$

The PDF $P_X(\cdot)$ satisfies the following properties

$$\forall x, P_X(x) \geq 0, \tag{2.11}$$

$$\int_{\mathcal{S}_X} P_X(x)dx = 1, \tag{2.12}$$

where $\mathcal{S}_X = \{x \in \Re : P_X(x) > 0\}$ is the *support* of $P_X(X)$.

The PDF of a discrete RV is a set of impulse functions located at the values taken by the RV. In this way, a discrete RV creates a mutually-exclusive and collectively-exhaustive partitioning of the sample space—each partition being $\Omega_x = \{\omega \in \Omega : X(\omega) = x\}$. For instance, assuming that the intensity takes only integer values in $[0, 255]$, we can define a discrete RV which maps each pixel in the image to its grayscale intensity. Then each partition corresponds to the event of a particular intensity $x$ being assigned to any pixel.

Here, we denote the PDF of an RV $X$ by $P_X(\cdot)$ that uses a subscript to signify the associated RV. In the future, for simplicity of notation, we may drop this subscript when

it is clear which RV we are referring to. The joint PDF $P_{X,Y}(\cdot)$ of two RVs $X$ and $Y$ is [123]

$$P_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}. \tag{2.13}$$

The *conditional distribution* $F_{X|M}(\cdot)$ of an RV $X$ assuming event $M$ is

$$F_{X|M}(x|M) = \frac{P(X \leq x, M)}{P(M)}, \tag{2.14}$$

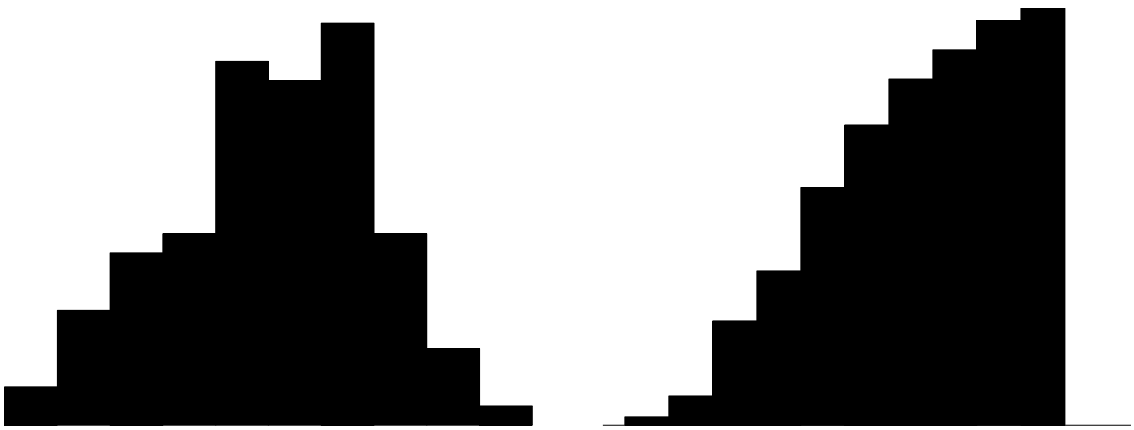
when $P(M) \neq 0$. The *conditional PDF* $P_{X|M}(\cdot)$ of an RV $X$ assuming event $M$ is

$$P_{X|M}(x|M) = \frac{dF_{X|M}(x|M)}{dx}. \tag{2.15}$$

Let us now consider examples of a few important PDFs, many of which we will encounter in the subsequent chapters in this dissertation. Figure 2.1 shows the PDF and CDF for a discrete RV. A continuous PDF, on the other hand, is the $d$D Gaussian PDF [123], also known as the Normal PDF:

$$G(x) = \frac{1}{(\sigma\sqrt{2\pi})^d} \exp\left( - \frac{(x-\mu)^2}{2\sigma^2} \right), \tag{2.16}$$

where $\mu$ and $\sigma$ are the associated parameters. Figure 2.2 shows the PDF and CDF of a Gaussian RV. One example of a PDF derived from Gaussian PDFs is the Rician



(a) (b)

**Figure 2.1**. Discrete RVs: (a) The PDF and (b) the CDF for a discrete RV.

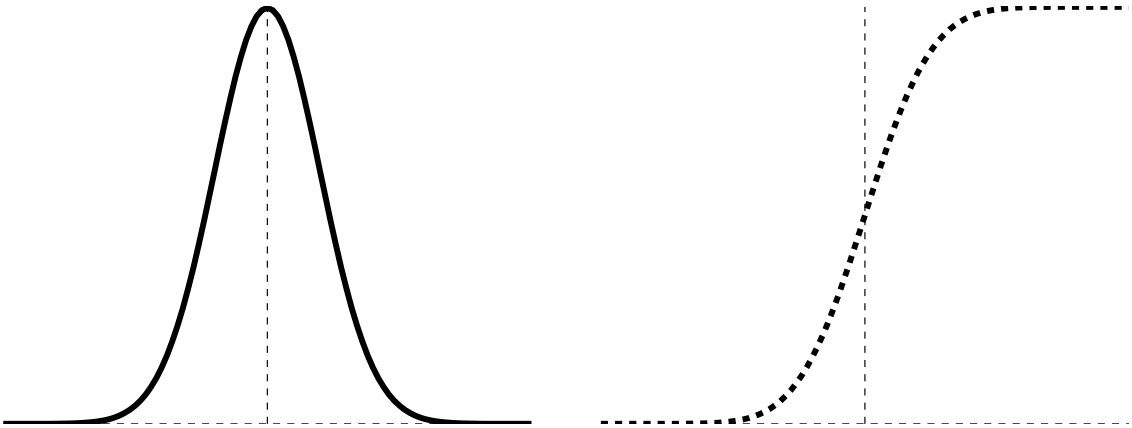



**Figure 2.2**. Continuous RVs: (a) The PDF and (b) the CDF for a continuous (Gaussian) RV with $\mu = 0$ and $\sigma = 1$.

PDF [123]. If independent RVs $X_1$ and $X_2$ have Gaussians PDFs with means $\mu_1, \mu_2$ and variance $\sigma^2$, then the RV $X = \sqrt{X_1^2 + X_2^2}$ has the Rician PDF:

$$P(x|\mu) = \frac{x}{\sigma^2} \exp\left( - \frac{x^2 + \mu^2}{2\sigma^2} \right) I_0\left( \frac{x\mu}{\sigma^2} \right), \tag{2.17}$$

where $\mu = \sqrt{\mu_1^2 + \mu_2^2}$. In practice, the Rician PDF results from independent additive Gaussian noise components in the real and imaginary parts of the complex MR data—the magnitude of the complex number produces a Rician PDF. The Rician PDF has close relationships with two other well-known PDFs: (a) the RV $((X_1/\sigma)^2 + (X_2/\sigma)^2)$ has a *noncentral chi-square* PDF [123] and (b) the Rician PDF reduces to a *Rayleigh* PDF [123] when $\mu = 0$. Figure 2.3 shows two Rician PDFs with different $\mu$ values and $\sigma = 1$. We can show that the Rician PDF approaches a Gaussian PDF as the ratio of $\mu/\sigma$ tends to infinity [123].

Two RVs are *independent* if their joint PDF is the product of the marginal PDFs, i.e.,

$$P_{X,Y}(X, Y) = P_X(X) P_Y(Y) \tag{2.18}$$

This is to say that knowing the value of one RV does not give us any information about the value of the other RV. In other words, the occurrence of some event corresponding to RV $X$ does not affect, in any way, the occurrence of events corresponding to RV $Y$, and

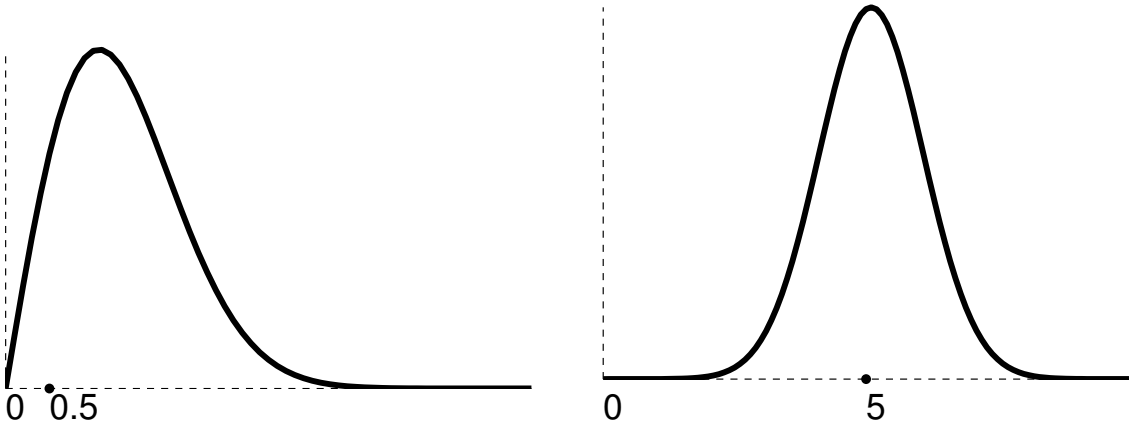



**Figure 2.3**. Rician PDFs with parameter values (a) $\mu = 0.5, \sigma = 1$, and (b) $\mu = 5, \sigma = 1$. Note the similarity between the Rician PDF in (b) and the Gaussian PDF in Figure 2.2(a).

vice versa. A set of RVs are *mutually independent* if their joint PDF is the product of the marginal PDFs, i.e.,

$$P_{X_1,X_2,\ldots,X_n}(X_1, X_2, \ldots, X_n) = P_{X_1}(X_1)P_{X_2}(X_2)\ldots P_{X_n}(X_n) \tag{2.19}$$

It is possible that each pair of RVs in a set be *pairwise independent* without the entire set being mutually independent [167].

Often, we deal with measures that characterize of certain properties of PDFs. One such quantity is the *expectation* or *mean* of an RV $X$:

$$E[X] = \int_{\mathcal{S}_X} xP(x)dx. \tag{2.20}$$

The expectation represents the average observed value $x$, if a sample is derived from the PDF $P(X)$. It also represents the *center of gravity* of the PDF $P(X)$. For example, the mean of a Gaussian PDF is $\mu$. The expectation is a linear operator, i.e., given two RVs $X$ and $Y$ and constants $a$ and $b$

$$E[aX + bY] = aE[X] + bE[Y]. \tag{2.21}$$

Deterministic functions $f(X)$ of an RV $X$ are also RVs [167]. The expected value of $Y = f(X)$ when the observations are derived from $P(X)$ is

$$E_{P(X)}[Y] = \int_{\mathcal{S}_X} f(x)P(x)dx. \tag{2.22}$$

The *variance* gives the variability or spread of the observations around the expectation:

$$\text{Var}(X) = \int_{\mathcal{S}_X} (x - E[X])^2 P(x) dx. \tag{2.23}$$

For example, the variance of a Gaussian PDF is $\sigma^2$.

## 2.3 Statistical Inference

In practice, we only have access to the data that a physical process generates rather than the underlying RVs or PDFs. Statistical inference refers to the process of using observed data to estimate the forms of the PDFs of the RVs, along with any associated *parameters*, that *model* the physical processes fairly accurately. The foundations of modern statistical analysis were laid down by Sir Ronald A. Fisher in the early 1900s.

In the statistical-inference terminology, a *population* is the set of elements about which we want to infer. A *sample* is a subset of the population that is actually observed. Thus, the goal is to learn about the statistical characteristics of the population from the sample data. Let us consider an RV $X$, with the associated PDF $P(X)$, that models some physical process and produces a set of $n$ independent observations $\{x_1, x_2, \ldots, x_n\}$. The goal is to infer some properties of $X$ from its observations. For instance, knowing that $P(X)$ was of a Gaussian form, we may want to determine the exact value for its mean and variance parameters such that the observed data best conform with the specific Gaussian model. We can consider each observation $x_i$ as the value of an RV $X_i$. Such a set of RVs $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ constitutes a *random sample*, and comprises a set of mutually independent RVs that are identically distributed:

$$\forall i, F_{X_i}(x) = F_X(x). \tag{2.24}$$

Suppose we want to estimate a particular parameter $\theta$ associated with the PDF of $X$. Here we assume that the data were derived from the PDF $P(X; \theta^*)$. A *statistic* $\hat{\Theta}$ is any deterministic function of the random sample and, hence, an RV itself. An *estimator* is a statistic $\hat{\Theta}(X_1, X_2, \ldots, X_n)$ that is used to estimate the value of some parameter $\theta$. Some properties of an estimator are highly desirable, e.g.,:

- *Unbiasedness*: we want the estimator to give the correct parameter value $\theta^*$, on an average, irrespective of the sample size—defined by

$$\forall n, E[\hat{\Theta}(X_1, X_2, \ldots, X_n)] = \theta^* \tag{2.25}$$

- *Consistency*: we want larger sample sizes to give progressively better estimates of the correct parameter value $\theta^*$ and asymptotically converge to $\theta^*$ *in probability*—defined by

$$\lim_{n\to\infty} P(|\hat{\Theta} - \theta^*| \geq \epsilon) = 0, \forall \epsilon > 0. \tag{2.26}$$

  If the estimator $\hat{\Theta}$ is unbiased, it is consistent when its variance $\text{Var}(\hat{\Theta})$ tends to zero asymptotically. This follows from the Chebyshev's inequality [167] that implies

$$P(|\hat{\Theta} - \theta^*| \geq \epsilon) \leq \frac{\text{Var}(\hat{\Theta})}{\epsilon^2}. \tag{2.27}$$

- *Efficiency*: we want the unbiased estimator to have the lowest possible variance—as determined by the Cramer-Rao bound [123]. Efficient estimators, however, need not exist in all situations.

As an example, for an RV $X$, an unbiased and consistent estimator of its mean, or expectation, is the sample mean [167],

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{2.28}$$

Another interesting example is that of the empirical CDF of a discrete RV, which is a consistent estimator of the true CDF $F_X(x)$ [167]. The empirical CDF for a discrete RV is

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \Big(1 - H(x_i - x)\Big), \tag{2.29}$$

where $H(x)$ is the Heaviside step (unit step) function.

### 2.3.1 Maximum-Likelihood (ML) Estimation

An important class of estimators is the *maximum-likelihood* (ML) estimators. The ML parameter estimate is the one that makes the set of mutually-independent observations $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ (which is an instance of the random sample $\{X_1, X_2, \ldots, X_n\}$) most likely to occur. The random sample comprises mutually independent RVs, thereby making the joint PDF equivalent to the product of the marginal PDFs. This defines the *likelihood function* for the parameter $\theta$ as

$$\begin{aligned} L(\theta|\mathbf{x}) &= P(\mathbf{x}|\theta) && (2.30)\\ &= P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n|\theta) && (2.31)\\ &= \prod_{i=1}^{n} P_{X_i}(x_i|\theta), && (2.32) \end{aligned}$$

The ML parameter estimate is

$$\theta^* = \underset{\theta}{\text{argmax}}\ L(\theta|\mathbf{x}). \tag{2.33}$$

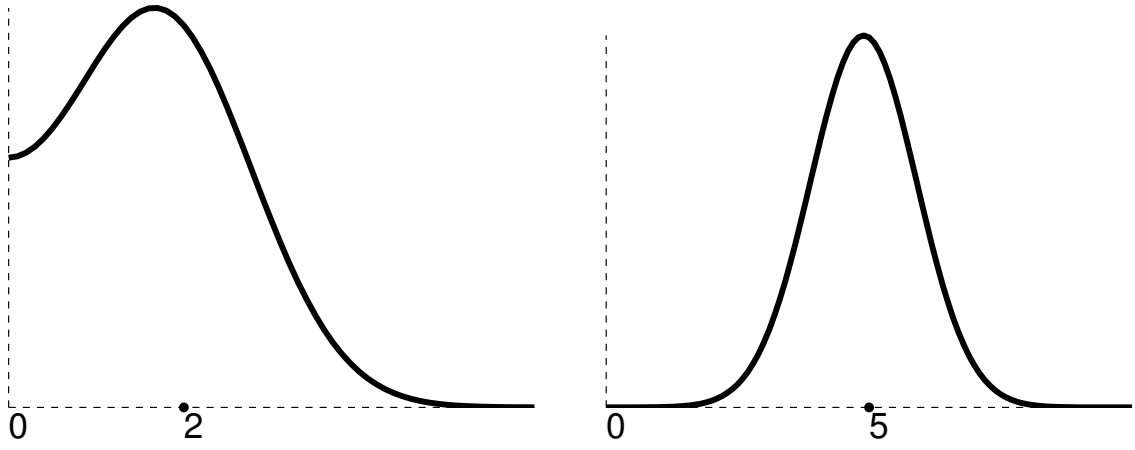
An interesting, and useful, property about ML estimators is that all efficient estimators are necessarily ML estimators [123]. As an example, consider a Rician PDF, with $\sigma = 1$ and unknown $\mu$, that generates a sample comprising just a single observation $x$. Then, the likelihood function $L(\mu|x)$ would be:

$$L(\mu|x) = \frac{1}{\eta}\frac{x}{\sigma^2}\exp\left(-\frac{x^2+\mu^2}{2\sigma^2}\right) I_0\left(\frac{x\mu}{\sigma^2}\right), \tag{2.34}$$

where $x$ and $\sigma$ are known constants, and $\eta$ is the normalization factor. Figure 2.4 shows the Rician-likelihood function for two different values of the observation $x$.

### 2.3.2 Maximum-a-Posteriori (MAP) Estimation

Sometimes we have *a priori* information about the physical process whose parameters we want to estimate. Such information can come either from the correct scientific knowledge of the physical process or from previous empirical evidence. We can encode such prior information in terms of a PDF on the parameter to be estimated. Essentially, we treat the parameter $\theta$ as the value of an RV. The associated probabilities $P(\theta)$ are called the *prior* probabilities. We refer to the inference based on such priors as *Bayesian*




(a) (b)

**Figure 2.4**. Rician likelihood functions with (a) $x = 2, \sigma = 1$, and (b) $x = 5, \sigma = 1$.

*inference*. Bayes' theorem shows the way for incorporating prior information in the estimation process:

$$P(\theta|\mathbf{x}) = \frac{P(\mathbf{x}|\theta)P(\theta)}{P(\mathbf{x})} \tag{2.35}$$

The term on the left hand side of the equation is called the *posterior*. On the right hand side, the numerator is the product of the likelihood term and the prior term. The denominator serves as a normalization term so that the posterior PDF integrates to unity. Thus, Bayesian inference produces the *maximum a posteriori* (MAP) estimate

$$\underset{\theta}{\text{argmax}}\ P(\theta|\mathbf{x}) = \underset{\theta}{\text{argmax}}\ P(\mathbf{x}|\theta)P(\theta). \tag{2.36}$$

### 2.3.3 Expectation-Maximization (EM) Algorithm

There are times when we want to apply the ML or MAP estimation technique, but the data $\mathbf{x}$ is *incomplete*. This implies that the model consists of two parts: (a) the observed part: $\mathbf{x}$ and (b) the *hidden* part: $\mathbf{y}$. We can associate RVs $X$ and $Y$ with the observed and hidden parts, respectively. We can still apply ML or MAP estimation techniques if we assume a certain joint PDF $P(X, Y)$ between the observed and hidden RVs, and then *marginalize* over the hidden RVs $Y$. Marginalization of an RV $Y$ chosen from a set of

RVs refers to the process of integration of the joint PDF over the values $y$ of the chosen RV. This is the key idea behind the EM algorithm.

Considering ML estimation, for example, we compute the optimal parameter as

$$\begin{aligned} \theta^* &= \underset{\theta}{\text{argmax}} \left( \log L(\theta|\mathbf{x}) \right) \\ &= \underset{\theta}{\text{argmax}} \left( \log \left( \int_{\mathcal{S}_Y} P(\mathbf{x}, y|\theta) dy \right) \right), \end{aligned} \tag{2.37}$$

where $L(\cdot)$ is the likelihood function described previously in Section 2.3.1. This key idea is formalized in the *expectation-maximization* (EM) algorithm [43, 104].

Herman O. Hartley [71] pioneered the research on the EM algorithm in the late 1950s. The first concrete mathematical foundation, however, was laid by Dempster, Laird, and Rubin [43] in the late 1970s. Neal and Hinton [111, 112, 108] presented the EM algorithm from a new perspective of lower-bound maximization. Over the years, the EM algorithm has found many applications in various domains and has become a powerful estimation tool [104, 48].

The EM algorithm is an iterative optimization procedure. Starting with an initial parameter estimate $\theta^0$, it is guaranteed to converge to the local maximum of the likelihood function $L(\theta|\mathbf{x})$. The EM algorithm consists of two steps: (a) the *E* step or the *expectation* step and (b) the *M* step or the *maximization* step.

- The E step constructs an *optimal lower bound* $B(\theta)$ to the log-likelihood function $\log L(\theta|\mathbf{x})$. This optimal lower bound is a function of $\theta$ that touches the log-likelihood function at the current parameter estimate $\theta^i$, i.e.,

$$B(\theta^i) = \log L(\theta^i|\mathbf{x}), \tag{2.38}$$

  and never exceeds the objective function at any $\theta$, i.e.,

$$\forall \theta \in (\infty, \infty) : B(\theta) \leq \log L(\theta|\mathbf{x}). \tag{2.39}$$

  Intuitively, maximizing this optimal lower bound $B(\theta)$ (in the M step) will surely take us closer to the maximum of the log-likelihood function $\log L(\theta|\mathbf{x})$, i.e., the ML estimate. We compute this optimal lower bound as follows [108, 42, 104].

Let us rewrite the log-likelihood function as

$$\begin{aligned}\log L(\theta|\mathbf{x}) &= \log P(\mathbf{x}|\theta) \\ &= \log \int_y P(\mathbf{x}, y|\theta) dy \\ &= \log E_{f(Y)}\left[\frac{P(\mathbf{x}, Y|\theta)}{f(Y)}\right], \end{aligned} \tag{2.40}$$

where $f(Y)$ is any arbitrary PDF. Applying Jensen's inequality [34], and using the concavity of the $\log(\cdot)$ function, gives:

$$\begin{aligned}\log E_{f(Y)}\left[\frac{P(\mathbf{x}, Y|\theta)}{f(Y)}\right] &\geq E_{f(Y)}\left[\log \frac{P(\mathbf{x}, Y|\theta)}{f(Y)}\right] \\ &\equiv B(\theta). \end{aligned} \tag{2.41}$$

Our goal is to try to find the particular PDF $f(Y)$ such that $B(\theta)$ is the *optimal* lower bound that touches the log-likelihood function at the current parameter estimate $\theta^i$. We can achieve this goal by solving the following constrained-optimization [137] problem:

$$\begin{aligned}&\text{Maximize } B(\theta^i) \\ &\text{with respect to } f(Y) \\ &\text{under the constraint } \int_y f(y)dy = 1. \end{aligned} \tag{2.42}$$

Using the Lagrange-multiplier [137] approach, the objective function to be maximized is

$$J(f(Y)) = B(\theta^i) + \lambda\left(1 - \int_y f(y)dy\right). \tag{2.43}$$

The derivative of the objective function $J(f(Y))$ with respect to $f(y)$ is

$$\frac{\partial J}{\partial f(y)} = -\lambda + \int_y \log P(\mathbf{x}, y|\theta^i)dy - \Big(1 + \log f(y)\Big). \tag{2.44}$$

The derivative of the objective function $J(f(Y))$ with respect to $\lambda$ is

$$\frac{\partial J}{\partial \lambda} = \int_y f(y)dy - 1. \tag{2.45}$$

The objective function achieves its maximum value when both the aforementioned derivatives in (2.44) and (2.45) are zero. Using these conditions, and after some simplification, we get

$$\begin{aligned} f(y) &= \frac{P(\mathbf{x}, y|\theta^i)}{P(\mathbf{x}|\theta^i)} \\ &= P(y|\mathbf{x}, \theta^i). \end{aligned} \tag{2.46}$$

This gives our optimal lower bound as

$$B(\theta) = \int_y P(y|\mathbf{x}, \theta^i) \log \frac{P(\mathbf{x}, y|\theta)}{P(y|\mathbf{x}, \theta^i)} dy \tag{2.47}$$

We can confirm that $B(\theta^i)$ indeed equals $\Big(\log P(\mathbf{x}|\theta^i)\Big)$, which indicates that $B(\theta)$ touches the log-likelihood function $\Big(\log P(\mathbf{x}|\theta)\Big)$ at $\theta^i$ and is an optimal lower bound.

- The M step performs the maximization of the function $B(\theta)$ with respect to the variable $\theta$.

$$\begin{aligned} \underset{\theta}{\text{argmax}}\ B(\theta) &= \underset{\theta}{\text{argmax}} \int_y P(y|\mathbf{x}, \theta^i) \log \frac{P(\mathbf{x}, y|\theta)}{P(y|\mathbf{x}, \theta^i)} dy \\ &= \underset{\theta}{\text{argmax}} \int_y P(y|\mathbf{x}, \theta^i) \log P(\mathbf{x}, y|\theta) dy \\ &= \underset{\theta}{\text{argmax}} \int_y P(y|\mathbf{x}, \theta^i) \log P(\mathbf{x}, y|\theta) dy \\ &= \underset{\theta}{\text{argmax}}\ Q(\theta) \end{aligned} \tag{2.48}$$

where the *Q function* is

$$Q(\theta) = E_{P(Y|\mathbf{x},\theta^i)}\Big[\log P(\mathbf{x}, Y|\theta)\Big] \tag{2.49}$$

$$= \int_{\mathcal{S}_Y} P(y|\mathbf{x}, \theta^i) \log P(\mathbf{x}, y|\theta) dy. \tag{2.50}$$

Observe that $Q(\theta)$ also depends on the current parameter estimate $\theta^i$ that is considered a constant. The M step assigns the new parameter estimate $\theta^{i+1}$ as the one that maximizes $Q(\theta)$, i.e.,

$$\theta^{i+1} = \underset{\theta}{\text{argmax}}\ Q(\theta). \tag{2.51}$$

The iterations proceed until convergence to a local maximum of $L(\theta|\mathbf{x})$. Actually, the M step need not find that $\theta^{i+1}$ corresponding to the maximum value of the Q function, but rather it is sufficient to find any $\theta^{i+1}$ such that

$$Q(\theta^{i+1}) \geq Q(\theta^i). \tag{2.52}$$

This modified strategy is referred to as the *generalized-EM* (GEM) algorithm and is also guaranteed to converge [43].

## 2.4 Nonparametric Density Estimation

Parametric modeling of PDFs assumes that the forms of the PDFs are known. Such knowledge typically comes from either a scientific analysis of the physical process or from empirical analysis of the observed data, e.g., a popular parametric PDF model for the noise in the k-space MRI data is the independent and identically distributed (i.i.d.) additive Gaussian. Then what remains, in statistical inference, is to estimate the parameters associated with the PDF. In many practical situations, however, simple parametric models do not accurately explain the physical processes. One reason for this is that virtually all the parametric PDF models are unimodal, but many practical situations exhibit multimodal PDFs. Attempts at modeling high-dimensional multimodal PDFs as products of 1D parametric PDFs do not succeed well in practice either. Therefore, one needs to employ the more sophisticated *nonparametric* density-estimation techniques that do not make any assumptions about the forms of the PDFs—except the mild assumption that PDFs are smooth functions [171, 156]—and can represent arbitrary PDFs given sufficient data. One such technique is the *Parzen-window density estimation*.

### 2.4.1 Parzen-Window Density Estimation

Emanuel Parzen [125] invented this approach in the early 1960s, providing a rigorous mathematical analysis. Since then, it has found utility in a wide spectrum of areas and applications such as pattern recognition [48], classification [48], image registration [170], tracking, image segmentation [32], and image restoration [9].

Parzen-window density estimation is essentially a data-interpolation technique [48, 171, 156]. Given an instance of the random sample, $\mathbf{x}$, Parzen-windowing estimates

the PDF $P(X)$ from which the sample was derived. It essentially superposes *kernel* functions placed at each observation or datum. In this way, each observation $x_i$ contributes to the PDF estimate. There is another way to look at the estimation process, and this is where it derives its name from. Suppose that we want to estimate the value of the PDF $P(X)$ at point $x$. Then, we can place a *window function* at $x$ and determine how many observations $x_i$ fall within our window or, rather, what is the contribution of each observation $x_i$ to this window. The PDF value $P(x)$ is then the sum total of the contributions from the observations to this window. The Parzen-window estimate is defined as

$$P(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n^d}K\left(\frac{x - x_i}{h_n}\right), \tag{2.53}$$

where $K(x)$ is the window function or kernel in the $d$-dimensional space such that

$$\int_{\Re^d} K(x)dx = 1, \tag{2.54}$$

and $h_n > 0$ is the *window width* or *bandwidth* parameter that corresponds to the width of the kernel. The bandwidth $h_n$ is typically chosen based on the number of available observations $n$. Typically, the kernel function $K(\cdot)$ is unimodal. It is also itself a PDF, making it simple to guarantee that the estimated function $P(\cdot)$ satisfies the properties of a PDF. The Gaussian PDF is a popular kernel for Parzen-window density estimation, being infinitely differentiable and thereby lending the same property to the Parzen-window PDF estimate $P(X)$. Using (2.53), the Parzen-window estimate with the Gaussian kernel becomes

$$P(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{(h\sqrt{2\pi})^d}\exp\left(-\frac{1}{2}\left(\frac{x - x_i}{h}\right)^2\right), \tag{2.55}$$

where $h$ is the standard deviation of the Gaussian PDF along each dimension. Figure 2.5 shows the Parzen-window PDF estimate, for a zero-mean unit-variance Gaussian PDF, with a Gaussian kernel of $\sigma = 0.25$ and increasing sample sizes. Observe that with a large sample size, the Parzen-window estimate comes quite close to the Gaussian PDF.

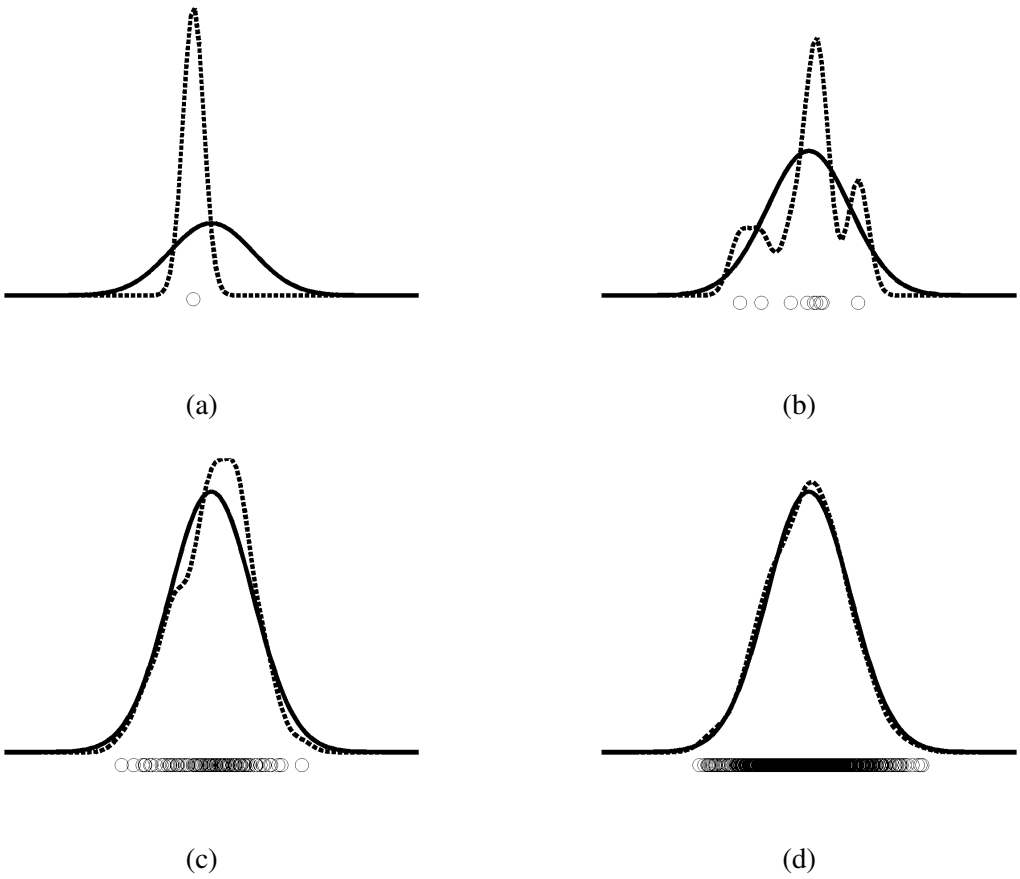



**Figure 2.5**. The Parzen-window PDF estimate (dotted curve), for a Gaussian PDF (solid curve) with zero mean and unit variance, with a Gaussian kernel of $\sigma = 0.25$ and a sample size of (a) 1, (b) 10, (c) 100, and (d) 1000. The circles indicate the observations in the sample.

### 2.4.2 Parzen-Window Convergence

We see in (2.53) that the kernel-bandwidth parameter $h_n$ can strongly affect the PDF estimate $P(X)$, especially when the number of observations $n$ is finite. Very small $h$ values will produce an irregular spiky $P(X)$, while very large values will excessively smooth out the structure of $P(X)$. For the case of finite data, i.e., finite $n$, the best possible strategy is to aim at a compromise between these two effects. Indeed, in this case, finding optimal values of $h_n$ entails additional constrains or strategies. For instance, the ML estimate yields an optimal $h_n$ value, and this is what we do in practice.

The case of an infinite number of observations, i.e., $n \rightarrow \infty$, is theoretically very interesting. In this case, Parzen proved that it is possible to have the PDF estimate

converge to the actual PDF [125, 48]. Let us consider $P_n(x)$ to be the estimator of the PDF at a point $x$ derived from a random sample of size $n$. This estimator has a mean $\bar{P}_n(x)$ and variance Var$(P_n(x))$. The estimator $P_n(x)$ *converges in mean square* to the true value $P(x)$, i.e.,

$$\begin{aligned} \lim_{n\to\infty} \bar{P}_n(x) &= P(x), \\ \lim_{n\to\infty} \text{Var}(P_n(x)) &= 0, \end{aligned} \tag{2.56}$$

when all the following conditions hold:

$$\begin{aligned} \sup_x K(x) &< \infty, \\ \lim_{|x|\to\infty} xK(x) &= 0, \\ \lim_{n\to\infty} h_n^d &= 0, \text{ and} \\ \lim_{n\to\infty} nh_n^d &= \infty. \end{aligned} \tag{2.57}$$

Figure 2.6 shows the process of convergence of the Parzen-window PDF, using a Gaussian kernel, to an arbitrary simulated PDF.

### 2.4.3 High-Dimensional Density Estimation

Some key ideas in this dissertation entail nonparametric PDF estimation where the observations lie in high-dimensional spaces. With a sufficiently large sample size, the Parzen-window estimate can converge to an arbitrarily-complex PDF. Alas, for guaranteeing convergence, the theory dictates that the sample size must increase exponentially with the dimensionality of the space. In practice, such a large number of samples are not normally available. Indeed, estimation in high-dimensional spaces is notoriously challenging because the available data populates such spaces very sparsely—regarded as *the curse of dimensionality* [155, 150, 156]. One reason behind this phenomenon is that high-dimensional PDFs can be, *potentially*, much more complex than low-dimensional ones, thereby demanding large amounts of data for a faithful estimation. There exists, however, inherent regularity in virtually all image data that we need to process [188, 79, 91, 40]. This makes the high-dimensional data lie on locally low-dimensional manifolds and, having some information about this locality, the PDF estimation becomes much simpler. Figure 2.7 depicts this phenomenon. Despite theoretical

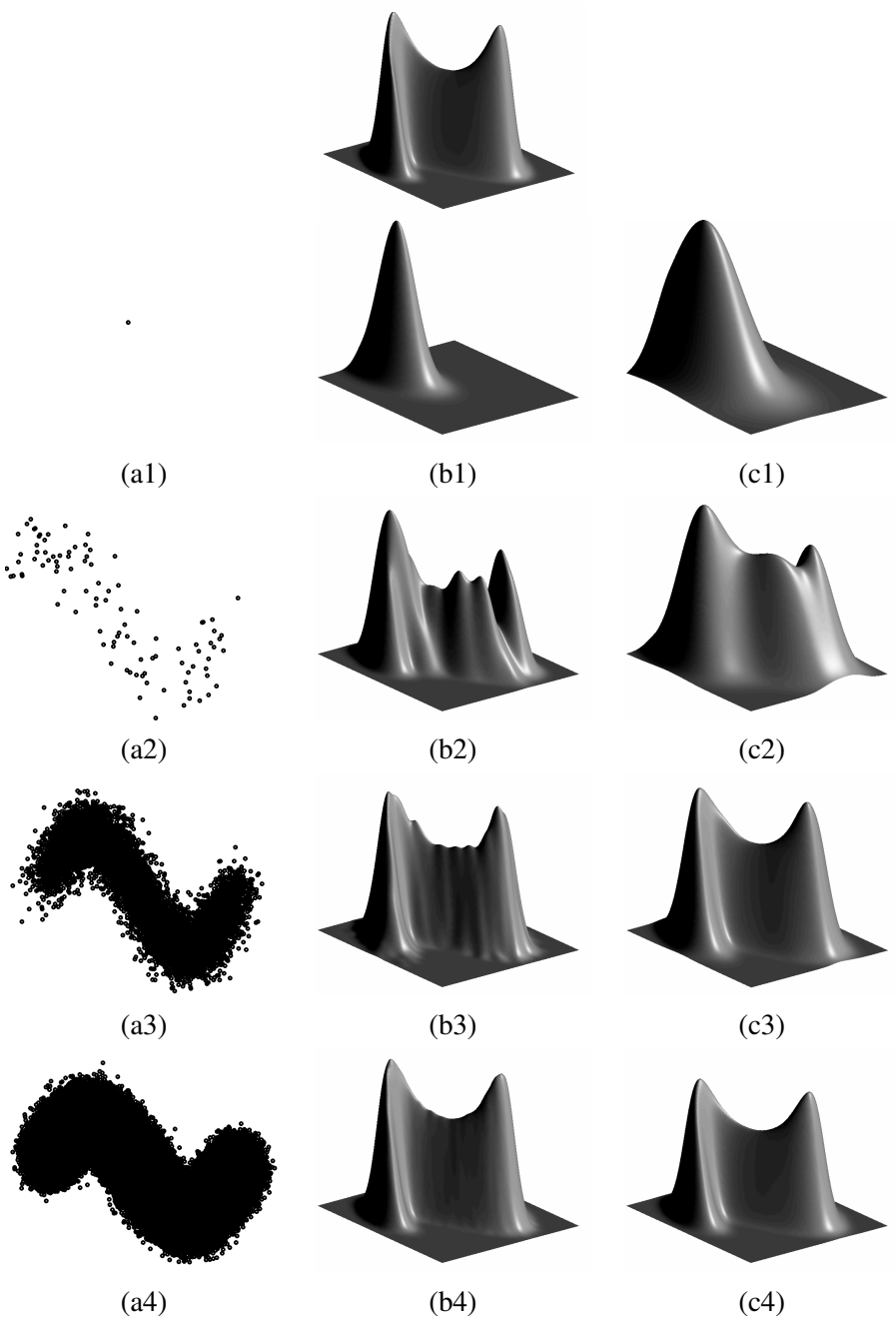



**Figure 2.6**. Convergence of the Parzen-window density estimate. The first row gives the true PDF. (a1)-(a4) show random samples derived from the true PDF: sample sizes progressively increasing by a factor of 100, starting with a sample size of one. (b1)-(b4) and (c1)-(c4) give the Parzen-window PDF estimate (2D Gaussian kernel) with progressively decreasing $\sigma$, starting with $\sigma = 2$ and $\sigma = 4$, respectively. Observe that *both* sequences of the estimated PDFs in (b1)-(b4) and (c1)-(c4) are converging towards the true PDF.

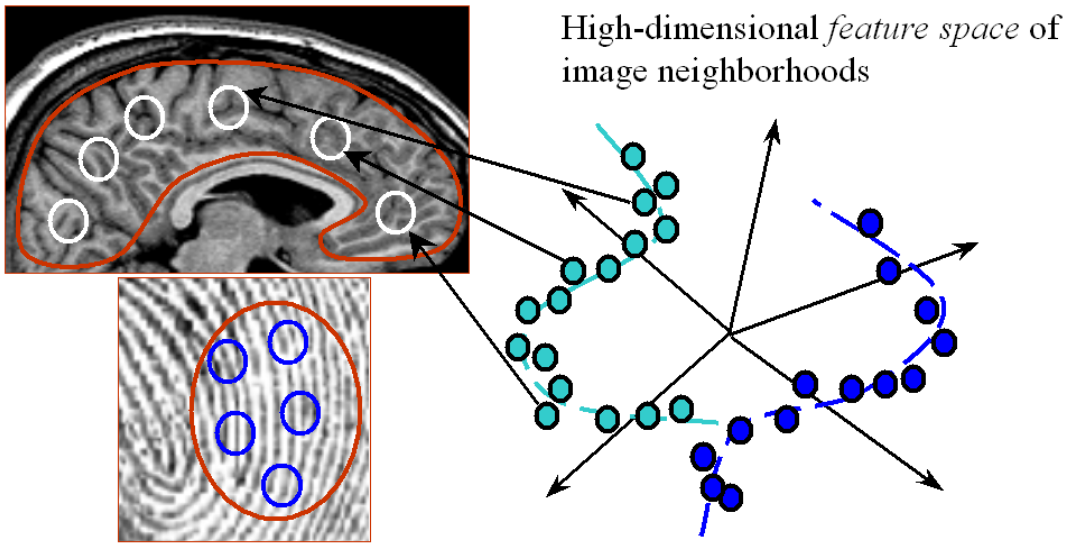



**Figure 2.7**. Neighborhoods (circles) in images and their locations (circles) on manifolds (dashed line) in the high-dimensional space. Different patterns in images, expectedly, produce neighborhoods lying on different manifolds.

arguments suggesting that density estimation beyond a few dimensions is impractical due to the unavailability of sufficient data, the empirical evidence from the literature is more optimistic [150, 131, 189, 50, 172]. The results in this dissertation confirm that observation.

## 2.5 Information Theory

Several algorithms in this dissertation enforce optimality criteria based on fundamental information-theoretic concepts that help us analyze the functional dependence, information content, and uncertainty in the data. In this way, information theory forms an important statistical tool in the design of unsupervised adaptive algorithms. This section presents a brief review of the relevant key information-theoretic concepts.

In the 1920s, Bell Labs researchers Harry Nyquist [116] and Ralph Hartley [72] pioneered the mathematical analysis of the transmission of messages, or information, over telegraph. Hartley was the first to define a quantitative measure of information associated with the transmission of a set of messages over a communication channel. Building on some of their ideas, another Bell Labs researcher Claude E. Shannon first presented [154], in the year 1948, a concrete mathematical model of communication

from a statistical viewpoint. This heralded the birth of the field of information theory. The principles underpinning the statistical theory have a universal appeal—virtually all practical systems process information in one way or the other—with information theory finding applications in a wide spectrum of areas such as statistical mechanics, business and finance, pattern recognition, data compression, and queuing theory [34, 85].

Information theory deals with the problem of quantifying the information content associated with events. If an event has a probability of occurrence $p$, then the *uncertainty* or *self-information* associated with the occurrence of that event is $\log\left(\frac{1}{p}\right)$ [154]. Thus, the occurrence of a less-certain event ($p \ll 1$) conveys more information. The occurrence of events that are absolutely certain ($p = 1$), on the other hand, conveys no information.

### 2.5.1 Entropy

The concept of entropy was prevalent, before Shannon, in the thermodynamics and statistical mechanics literature. In classical thermodynamics, the important second law states that the total entropy of any isolated thermodynamic system tends to increase with time. Ludwig Boltzmann and Josiah W. Gibbs, in the late 1800s, statistically analyzed the randomness associated with an ensemble of gas particles. They called this measure entropy and defined it to be proportional to the logarithm of the number of microstates such a gas could occupy. Their mathematical formulation of entropy, albeit in a different context, was equivalent to the definition by Shannon.

Shannon defined a measure of uncertainty or randomness associated with an RV, calling it *entropy* [154]. Thus, entropy is the average uncertainty associated with each possible value of the RV:

$$h(X) = \int_{\mathcal{S}_X} P(x) \log\left(\frac{1}{P(x)}\right) dx \tag{2.58}$$

$$= -\int_{\mathcal{S}_X} P(x) \log P(x) dx, \tag{2.59}$$

where $\mathcal{S}_X = \{x : P(x) > 0\}$ is the support set of $P(X)$.

Alfred Renyi [138] generalized Shannon's measure of entropy by presenting a family of entropy functions parameterized by a continuous parameter $\alpha$:

$$h_\alpha(X) = \frac{1}{1-\alpha} \log\left(\int_{\mathcal{S}_X} \Big(P(x)\Big)^\alpha dx\right). \tag{2.60}$$

He showed that the Renyi entropy converges to the Shannon entropy in the limit as $\alpha \rightarrow 1$. Many other measures of entropy exist such as the Havrda-Chavrat entropy [84], Hartley entropy [72], and Kapur's measures of entropy [85, 84]. This dissertation utilizes the Shannon measure for all purposes and, hence, we will restrict our focus to that measure.

We can also interpret Shannon entropy as the expectation of the RV $\Big( -\log P(X)\Big)$, i.e.,

$$h(X) = E_{P(X)}\Big[ -\log P(X)\Big]. \tag{2.61}$$

We saw previously that, given a random sample, an unbiased and consistent estimator of the expectation of the RV is the sample mean. Thus, given a random sample derived from an RV $X$, an estimate for the entropy of $X$ as

$$\begin{aligned} h(X) &\approx \frac{1}{n}\sum_{i=1}^{n}\Big( -\log P(x_i)\Big) \\ &= -\frac{1}{n}\log\left(\prod_{i=1}^{n} P(x_i)\right). \end{aligned} \tag{2.62}$$

We can observe that the expression on the right involves the product of the probabilities of occurrence of the observations. This product is, in fact, the likelihood function associated with the observations. Recall that the ML estimate selects that parameter value that maximizes the likelihood function—where each term is the probability conditioned on the parameter value. Indeed, we can prove that the ML parameter estimates are the same as the minimum-entropy parameter estimates when dealing with Shannon's entropy measure:

$$\begin{aligned} \underset{\theta}{\text{argmax}} \prod_{i=1}^{n} P(x_i|\theta) &= \underset{\theta}{\text{argmin}} \frac{-1}{n}\log\left(\prod_{i=1}^{n} P(x_i|\theta)\right) \\ &= \underset{\theta}{\text{argmin}} \frac{1}{n}\sum_{i=1}^{n}\Big( -\log P(x_i|\theta)\Big) \\ &\approx \underset{\theta}{\text{argmin}}\ h(X). \end{aligned} \tag{2.63}$$

The *joint entropy* of two RVs $X$ and $Y$ is

$$h(X,Y) = \int_{\mathcal{S}_X}\int_{\mathcal{S}_Y} -P(x,y)\log P(x,y)dxdy, \tag{2.64}$$

analogous to the definition of the entropy of a single RV [154].

### 2.5.2 Conditional Entropy

The *conditional entropy* of an RV $X$ given RV $Y$ is a measure of the uncertainty remaining in $X$ after $Y$ is observed [154]. It is defined as the weighted average of the entropies of the conditional PDFs of $X$ given the value of $Y$, i.e.,

$$h(X|Y) = \int_{\mathcal{S}_Y} P(y)h(X|y)dy. \tag{2.65}$$

Thus, functionally-dependent RVs will have minimal conditional entropy, i.e., $-\infty$. This is because, for a given $y$, the value $x$ is exactly known thereby causing $h(X|y) = 0, \forall y$. For independent RVs, however,

$$\begin{aligned} h(X|Y) &= \int_{\mathcal{S}_Y} P(y)h(X|y)dy \\ &= \int_{\mathcal{S}_Y} P(y)h(X)dy \\ &= h(X). \end{aligned} \tag{2.66}$$

### 2.5.3 Kullback-Leibler (KL) Divergence

The *Kullback-Leibler* (KL) divergence or *relative entropy* is a measure of mismatch between two PDFs $P(X)$ and $Q(X)$:

$$\text{KL}\left(P \parallel Q\right) = E_{P(X)}\left[\log \frac{P(X)}{Q(X)}\right]. \tag{2.67}$$

The KL divergence is always nonnegative. It is zero if and only if $P(X)$ and $Q(X)$ are exactly the same. It is not symmetric and does not follow the triangle inequality. Hence, it is not a true distance measure.

### 2.5.4 Mutual Information

The *mutual information* between two RVs $X$ and $Y$ is a measure of the information contained in one RV about another [154]:

$$I(X,Y) = \int_{\mathcal{S}_X}\int_{\mathcal{S}_Y} P(x)P(y)\log\frac{P(x,y)}{P(x)P(y)}dxdy. \tag{2.68}$$

Rewriting $I(X,Y) = h(X) - h(X|Y) = h(Y) - h(Y|X)$ allows us to interpret mutual information as the amount of uncertainty reduction in $h(X)$ when $Y$ is known, or vice

versa. Statistically-independent RVs have zero mutual information. We can see mutual information as the KL divergence between the joint PDF $P(X,Y)$ and the individual PDFs $P(X)$ and $P(Y)$. For independent RVs, i.e., when $P(X,Y) = P(X)P(Y)$, the mutual information is zero. The notion of mutual information extends to $N$ RVs and is termed *multi information* [162]:

$$\begin{aligned} I(X_1,\ldots,X_N) &= \int_{\mathcal{S}_{X_1}} \ldots \int_{\mathcal{S}_{X_N}} P(x_1,x_2,\ldots,x_N) \log \frac{P(x_1,\ldots,x_N)}{P(x_1)\ldots P(x_N)} dx_1 \ldots dx_N \\ &= \sum_{i=1}^{N} h(X_i) - h(X_1,\ldots,X_n). \end{aligned} \tag{2.69}$$

## 2.6 Markov Random Fields

Markov random fields (MRFs) are stochastic models that characterize the local spatial interactions in data. The last 40 years have seen significant advances in the mathematical analysis of MRFs as well as numerous application areas for MRFs ranging from physics, pattern recognition, machine learning, artificial intelligence, image processing, and computer vision. This has firmly established MRFs as powerful statistical tools for data analysis. This dissertation proposes an adaptive MRF image model and builds processes images relying on this model. This section gives a brief review of theory behind MRFs and some relevant MRF-based algorithms.

The first concept of the MRF theory came from the physicist Ernst Ising in the 1920s. Ising was trying to devise a mathematical model to explain the experimental results concerning properties of ferromagnetic materials. This dealt with local interactions between a collection of dipoles associated with such materials. He published the model in his doctoral thesis, which later became popular as the *Ising model*. The name Markov, however, is dedicated in the memory of the mathematician Andrei Markov who pioneered the work on *Markov chains*, i.e., ordered sequences of RVs where the conditional PDF of an RV given all previous RVs is exactly the same as the conditional PDF of the RV given only its preceeding RV. In other words, the next RV, given the present RV, is conditionally independent of all other previous RVs. This notion of conditional independence concerning chains of RVs generalizes to grids of RVs or random fields. Such random fields are called MRFs.

A random field [47, 161] is a family of RVs $\mathbf{X} = \{X_t\}_{t\in\mathcal{T}}$, for some index set $\mathcal{T}$. For each index $t$, the RV $X_t$ is defined on some sample-space $\Omega$. If we let $\mathcal{T}$ be a set of points defined on a discrete Cartesian grid and fix $\Omega = \omega$, we have a *realization* or an *instance* of the random field, $\mathbf{X}(\omega) = \mathbf{x}$, called the *digital image*. In this case, $\mathcal{T}$ is the set of grid points in the image. For vector-valued images $X_t$ becomes a vector RV.

In the early 1970s, Spitzer, Preston, Hammersely, Clifford, and Besag were among the pioneers who rigorously analyzed the theory behind the stochastic models for systems of spatially-interacting RVs. The joint PDF $P(\mathbf{X})$ of all the RVs in the random field dictates the image-formation process. However, modeling this joint PDF is intractable because of the enormous dimensionality $|\mathcal{T}|$ that equals the number of pixels in the image. Early researchers advocated the use of the lower-dimensional conditional PDFs, one associated with each RV $X_t$, to model the statistical dependencies between RVs. Such PDFs were conditioned only on the values of a few RVs in the spatial proximity of the RV in concern, thereby making the analysis tractable. These ideas rely on the notion of a neighborhood, which we define next.

We can associate with the index set $\mathcal{T}$, a family of *neighborhoods*

$$\begin{aligned} \mathcal{N} &= \{\mathcal{N}_t\}_{t\in\mathcal{T}} \text{ suchthat} \\ \mathcal{N}_t &\subset \mathcal{T}, \\ t &\notin \mathcal{N}_t, \text{ and} \\ \Big(u \in \mathcal{N}_t\Big) &\Leftrightarrow \Big(t \in \mathcal{N}_u\Big). \end{aligned} \tag{2.70}$$

Then $\mathcal{N}$ is called a *neighborhood system* for the set $\mathcal{T}$. Indices in $\mathcal{N}_t$ constitute the neighborhood of index $t$. $\mathcal{N}_t$ is also referred to as the *Markov blanket* or *Markov cover* for index $t$. We define a random vector $\mathbf{Y}_t = \{X_t\}_{t\in\mathcal{N}_t}$ to denote image neighborhoods. Figure 2.8 shows a 3-pixel $\times$ 3-pixel square neighborhood.

Based on this general notion of a neighborhood, $X(\Omega, T)$ is a MRF if and only if

$$\Big(P(x_t) > 0, \forall t\Big) \Rightarrow P(x_1, x_2, \ldots, x_{|\mathcal{T}|}) > 0, \text{ and} \tag{2.71}$$

$$\forall t, P\Big(X_t | \{x_u\}_{u\in\{\mathcal{T}\setminus\{t\}\}}\Big) = P(X_t|\mathbf{y}_t). \tag{2.72}$$

The first condition above is the *positivity condition*. The second one is the *Markovity* condition that implies the *conditional independence* of any RV ($X_t$), with respect to all

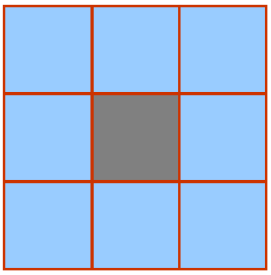


**Figure 2.8**. A 3-pixel $\times$ 3-pixel square neighborhood. The center pixel is shaded different from its neighbors.

other RVs not in its Markov cover ($\mathcal{T} - \mathcal{N}_t$), given the values of RVs in its Markov cover ($\mathcal{N}_t$). This means that, given the the Markov cover of an RV, the remaining RVs carry no extra information about the RV. We define a random vector $\mathbf{Z}_t = (X_t, \mathbf{Y}_t)$. We refer to the PDFs $P(X_t, \mathbf{Y}_t) = P(\mathbf{Z}_t)$ as *Markov PDFs* defined on the *feature space* $< \mathbf{z} >$.

### 2.6.1 Markov Consistency

The luxury of employing local conditional PDFs—locality is defined by the neighborhood system $\mathcal{N}$—to make the statistical analysis tractable, demands a price. Besag's seminal paper [14] states that Hammersely and Clifford, in their unpublished work of 1971, found that these conditional PDFs must conform to specific functional forms, namely the Gibbs PDFs, in order to give a *consistent* structure to entire system; a consistent system is one where we can obtain each conditional PDF, $P(X_t|\mathbf{y}_t)$ ($\forall t \in \mathcal{T}, \forall \mathbf{y}_t \in \Re^{|\mathcal{N}_t|}$) via rules of probabilistic inference from the joint PDF $P(\mathbf{X})$ of all the RVs in the system. Besag, later in 1974 [14], published the theorem and gave an elegant mathematical proof of the equivalence between the consistent Markov PDFs and Gibbs PDFs [14]. The consistency theorem is known as the *Hammersely-Clifford theorem*, or the *MRF-Gibbs equivalence theorem*. It states that every MRF is equivalent to a *Gibbs random field* (GRF) [14, 99]. We define the GRF next.

The definition of a GRF requires the notion of a *clique*. A clique $c$, associated with a neighborhood system $\mathcal{N}$, is a subset of the index set $\mathcal{T}$ such that it either comprises a single index $c = \{t\}$ or a set of indices where each each index is a neighbor of every other index. Let us call $\mathcal{C}_m$ as the set of all cliques comprising $m$ indexes. Then,

$$\mathcal{C}_1 = \{\{t\}|t \in \mathcal{T}\}, \tag{2.73}$$

$$\mathcal{C}_2 = \{\{t_1, t_2\}|t_1 \in \mathcal{T}, t_2 \in \mathcal{T}, t_2 \in \mathcal{N}_{t_1}\}, \tag{2.74}$$

$$\mathcal{C}_3 = \{\{t_1, t_2, t_3\}|t_1 \in \mathcal{T}, t_2 \in \mathcal{T}, t_3 \in \mathcal{T}, t_2 \in \mathcal{N}_{t_1}, t_3 \in \mathcal{N}_{t_1}, t_3 \in \mathcal{N}_{t_2}\}, \tag{2.75}$$

and so on. The collection of all cliques for the neighborhood system $\mathcal{N}$ is

$$\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3 \cup \ldots \cup \mathcal{C}_{|\mathcal{T}|}, \tag{2.76}$$

where the $|\cdot|$ operator gives the cardinality of sets. Figure 2.9 shows all possible clique types for a 3-pixel $\times$ 3-pixel square neighborhood system depicted in Figure 2.8.

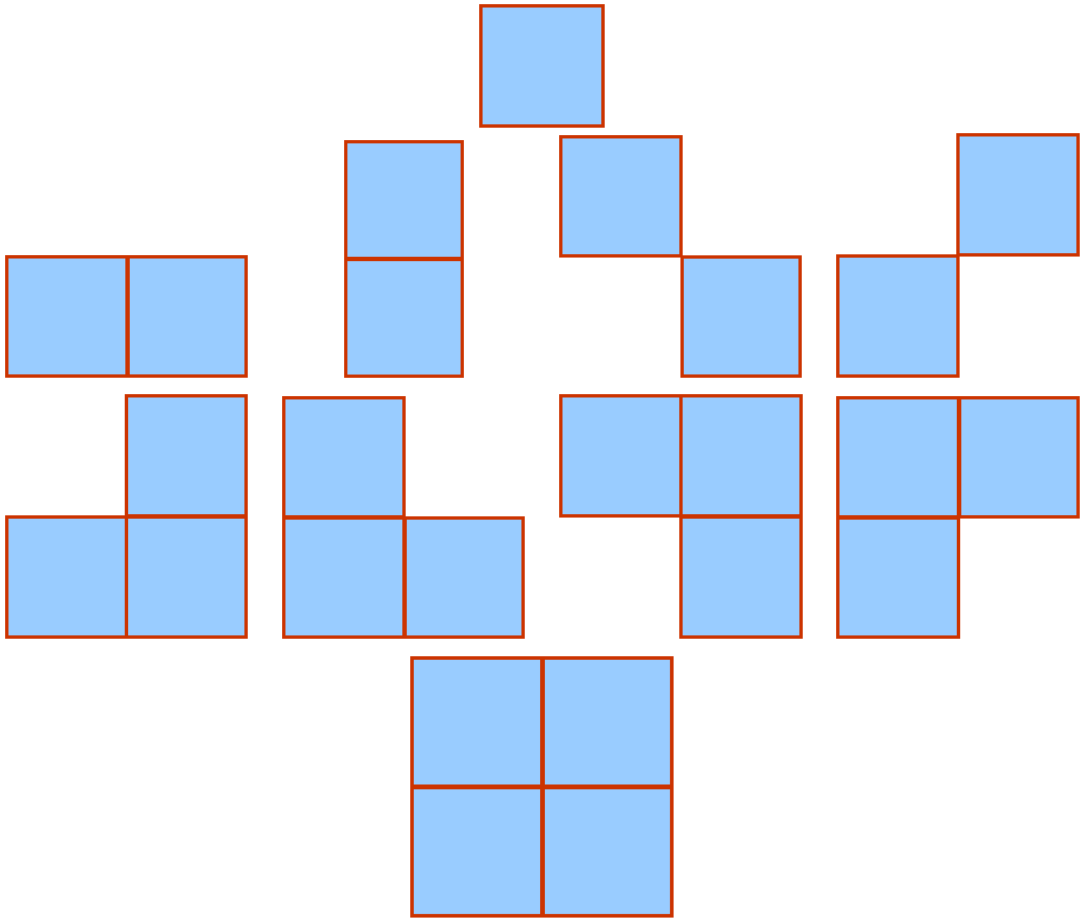


**Figure 2.9**. All possible clique types for a 3-pixel $\times$ 3-pixel square neighborhood system in Figure 2.8. The four rows (top to bottom) show cliques of types $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$, and $\mathcal{C}_4$, respectively.

A GRF is a random field whose joint PDF is

$$P(\mathbf{x}) = \frac{1}{\eta} \exp\left( - \frac{U(\mathbf{x})}{\tau} \right), \tag{2.77}$$

where $\tau$ is the *temperature*,

$$U(\mathbf{x}) = \sum_{c \in \mathcal{C}} V_c(\mathbf{x}) \tag{2.78}$$

is the *energy function*, $V_c(\cdot)$ is an arbitrary *clique-potential* function, and

$$\eta = \int_{\mathbf{x}} \exp\left( - \frac{U(\mathbf{x})}{\tau} \right) d\mathbf{x} \tag{2.79}$$

is the *partition function*. The temperature $\tau$ controls the probabilities—at high $\tau$ every instance $\mathbf{x}$ is almost equally probable, but at low values of $\tau$ it is the clique potentials that dictate the probabilities.

### 2.6.2 Parameter Estimation

Modeling the Markov PDFs parametrically entails data-driven optimal estimation of the parameters associated with the GRF potential functions or the Markov PDFs, lest we enforce an ill-fitted model on the data. Even nonparametric schemes are not free of internal parameters and one would want to learn these parameters in a data-driven manner. Standard estimation schemes, e.g., maximum likelihood, are not applicable in a straightforward manner for this task. Consider that we want to estimate some parameter $\theta$ in the MRF model. A ML-estimation scheme needs to evaluate the joint PDF of all the RVs in the MRF, i.e., $P(\mathbf{x}|\theta)$, which is a function of $\theta$. We can compute the potential functions $V_c(\mathbf{x}, \theta)$, as functions of $\theta$, in a simple way. The partition function $\eta(\theta)$, however, involves a $\theta$-dependent integral over the entire $|\mathcal{T}|$-dimensional space of possible realizations of the MRF. This is virtually intractable for any practical dataset, or image, comprising a reasonable number of indices $|\mathcal{T}|$. For instance, a $256 \times 256$ pixels image results in a $65536$D space.

Besag [14, 15] devised one way to bypass this problem in the following way. Based on his idea, we first choose a set of indices $\mathcal{T}_\alpha$ such that the neighborhoods for the indices in $\mathcal{T}_\alpha$ do not overlap, i.e.,

$$\mathcal{T}_\alpha \subset \mathcal{T}, \tag{2.80}$$

$$\mathcal{N}_t \cap \mathcal{N}_u \;\; = \;\; \phi, \forall t, u \in \mathcal{T}_\alpha. \tag{2.81}$$

This makes the set of random vectors corresponding to these neighborhoods mutually independent and identically distributed and, hence, a random sample. Besag referred to this partitioning process as the *coding scheme*. Then, the likelihood function is

$$L(\theta) = \prod_{t \in \mathcal{T}_\alpha} P(x_t | \mathbf{y}_t, \theta) \tag{2.82}$$

and the optimal parameter estimate is

$$\underset{\theta}{\text{argmax}} \;\; L(\theta). \tag{2.83}$$

This does not involve evaluation of the unwieldy partition function and standard numerical optimization techniques, e.g., the Newton-Raphson method, can produce the optimal estimate.

A major drawback of the coding-based parameter estimation is the wastage of data [14, 15] because it utilizes only a small part $\mathcal{T}_\alpha$ ($|\mathcal{T}_\alpha| \ll |\mathcal{T}|$) of the entire data. Another drawback is that the partition $\mathcal{T}_\alpha$ is not unique, and different partitions produce potentially different parameter estimates. There appears no clear way of reconciliation between these different estimates [99].

To alleviate the drawbacks of the coding scheme, Besag [14, 15] invented a simple approximate scheme called the *pseudo-likelihood* estimation. This eliminated any coding strategies and used all the data at hand. The pseudo-likelihood function $L_{\text{pseudo}}(\theta)$ is simply the product of the conditional likelihoods at each index $t \in \mathcal{T}$, i.e.,

$$L_{\text{pseudo}}(\theta) = \prod_{t \in \mathcal{T}} P(X_t | \mathbf{y}_t, \theta). \tag{2.84}$$

The optimal parameter estimate is

$$\underset{\theta}{\text{argmax}} \;\; L_{\text{pseudo}}(\theta). \tag{2.85}$$

The overlapping neighborhoods of indices $t$ in the product do not produce independent observations, and the resulting function is not the true likelihood function—hence the name. Geman and Graffigne [62], later proved that the pseudo-likelihood estimate converges, with probability one, to the true ML estimate asymptotically with infinite data ($|\mathcal{T}| \to \infty$).

The literature also presents other methods of MRF-parameter estimation such as those based on mean-field approximations and least-squares fitting [99].

### 2.6.3 Bayesian Image Restoration

We can use MRF models together with fundamental principles from statistical decision theory to formulate optimal image-processing algorithms. One such optimality criterion is based on the MAP estimate. Let us consider the uncorrupted image $\mathbf{x}$ as a realization of a MRF $\mathbf{X}$, and the observed degraded image $\tilde{\mathbf{x}}$ as a realization of a MRF $\tilde{\mathbf{X}}$. Given the true image $\mathbf{x}$, let us assume, for simplicity, that the RVs in the MRF $\tilde{\mathbf{X}}$ are *conditionally independent*. This is equivalent to saying that the noise affects each image location independently of any other location. Given the stochastic model $P(\tilde{x}_t|x_t)$ for the degradation process, conditional independence implies that the conditional probability of the observed image given the true image is

$$P(\tilde{\mathbf{x}}|\mathbf{x}) = \prod_{t\in\mathcal{T}} P(\tilde{x}_t|x_t). \tag{2.86}$$

Our goal is to find the MAP estimate $\hat{\mathbf{x}}^*$ of the true image $\mathbf{x}$

$$\hat{\mathbf{x}}^* = \underset{\mathbf{x}}{\text{argmax}}\ P(\mathbf{x}|\tilde{\mathbf{x}}) \tag{2.87}$$

This MAP-estimation problem is an optimization problem that, like many other optimization problems, suffers from the existence of many local maxima. Two classes of optimization algorithms exist to solve this problem: (a) methods that guarantee to find the unique global maximum and (b) methods that converge only to local maxima. Typically, the former class of methods are significantly slower. Here we face a trade-off between finding the global maximum at a great expense and finding local maxima with significantly less cost.

### 2.6.4 Stochastic Restoration Algorithms

Optimization methods that find the global maximum of the objective function $P(\mathbf{X}|\tilde{\mathbf{x}})$ include *annealing*-based methods [99]. These methods optimize iteratively, starting from

an arbitrary initial estimate. Recalling the discussion in Section 2.6.1, where $\tau$ is the temperature parameter of the GRF, consider the parametric family of functions

$$P_\tau(\mathbf{X}|\tilde{\mathbf{x}}) = \left(\frac{P(\mathbf{X})P(\tilde{\mathbf{x}}|\mathbf{X})}{P(\tilde{\mathbf{x}})}\right)^{1/\tau}. \tag{2.88}$$

- As $\tau \to \infty$, $P_\tau(\mathbf{X}|\tilde{\mathbf{x}})$ is a uniform PDF.
- For $\tau = 1$, $P_\tau(\mathbf{X}|\tilde{\mathbf{x}})$ is exactly the same as our objective function $P(\mathbf{X}|\tilde{\mathbf{x}})$.
- At the other extreme, as $\tau \to 0$, $P_\tau(\mathbf{X}|\tilde{\mathbf{x}})$ is concentrated on the peaks of our objective function $P(\mathbf{X}|\tilde{\mathbf{x}})$.

The key idea behind annealing-based method is to decrease the temperature parameter $\tau$, starting from a very high value, via a *cooling schedule*. At sufficiently high temperatures $\tau \gg 1$, the objective-function landscape is smooth with a unique local maximum. Annealing first tries to find this maximum and then, as the temperature $\tau$ reduces, continuously tracks the evolving maximum. Annealing-based methods mimic the physical annealing procedure, based on principles in thermodynamics and material science, where a molten substance is gradually cooled so as to reach the lowest energy state.

The literature presents two kinds of annealing strategies:

- *Stochastic* strategies such as *simulated annealing* by Kirkpatrik *et al.* [89] that typically rely on the sampling procedures including the *Metropolis-Hastings* algorithm [106, 73] and the *Gibbs sampler* [61]. Direct sampling from the PDFs of all RVs in the random field is intractable. The sampling algorithms can generate samples from any PDF by generating a Markov chain that has the desired PDF as the stationary (steady-state) distribution. Once in the steady state, samples from the Markov chain can be used as samples from the desired PDF. Gibbs sampling entails that all the conditional Markov PDFs associated with the random field are known and can be sampled exactly. Simulated annealing is extremely slow in practice and significantly sensitive to the cooling schedule [99].
- *Deterministic* strategies include *graduated nonconvexity*, by Blake and Zisserman [20], that is much faster than simulated annealing. The graduated nonconvexity, however, gives no guarantees for convergence to the exact global maximum [99].

### 2.6.5 Deterministic Restoration Algorithms

The MAP optimization problem can be dealt with much faster if we give up the need to converge to a global maximum and be satisfied on finding local maxima. Indeed, using a smart choice for an initial estimate, one can obtain local-maximum solutions that serve the purpose just as well as the global-maximum solution. Besag suggested *deterministic* algorithms for the optimization, guaranteeing convergence to local maxima. Writing the posterior as

$$P(\mathbf{x}|\tilde{\mathbf{x}}) = P(x_t|\{x_u\}_{u\in\mathcal{T}\setminus\{t\}}, \tilde{\mathbf{x}})P(\{x_u\}_{u\in\mathcal{T}\setminus\{t\}}|\tilde{\mathbf{x}}) \tag{2.89}$$

motivates us to employ an iterative restoration scheme where, starting from some initial image estimate $\hat{\mathbf{x}}^0$, we can always update the current estimate $\hat{\mathbf{x}}^i$, at iteration $i$, so that the posterior never decreases. The algorithm computes the next estimate $(i+1)$ by cycling through all indices as follows:

1. Label the indices in $\mathcal{T}$ as $t_1, t_2, \ldots, t_{|\mathcal{T}|}$. Set $i \leftarrow 1$.
2. Set $t \leftarrow t_i$.
3. Update value at index $t$:

$$x_t \leftarrow \underset{x_t}{\text{argmax}}\ P(x_t|\{x_u\}_{u\in\mathcal{T}\setminus\{t\}}, \tilde{\mathbf{x}}). \tag{2.90}$$

4. Increment index: $i \leftarrow i+1$.
5. If $i > |\mathcal{T}|$ stop, otherwise go to Step 2.

This algorithm is the *iterated conditional modes* (ICM) algorithm [14], because it repeatedly updates image values based on modes of the conditional PDFs in Step 3. We can compute the mode of such conditional PDF by using Bayes rule, Markovity, and (2.86), as follows:

$$\begin{aligned} \underset{x_t}{\text{argmax}}\ P(x_t|\{x_u\}_{u\in\mathcal{T}\setminus\{t\}}, \tilde{\mathbf{x}}) &= \underset{x_t}{\text{argmax}}\ P(x_t|\{x_u\}_{u\in\mathcal{T}\setminus\{t\}})P(\tilde{\mathbf{x}}|\mathbf{x}) \\ &= \underset{x_t}{\text{argmax}}\ P(x_t|\mathbf{y}_t)P(\tilde{x}_t|x_t), \end{aligned} \tag{2.91}$$

where $P(x_t|\mathbf{y}_t)$ is the prior and $P(\tilde{x}_t|x_t)$ is the likelihood determined from the statistical noise model.

The ICM algorithm guarantees convergence to a local maximum provided that no two neighboring indices are simultaneously updated. Updating all sites at once, namely *synchronous updating* that is typically observed in image-processing algorithms [99], may cause small oscillations. On the other hand, synchronous-updating schemes are easily parallelizable. A *partially-synchronous* updating scheme offers a compromise. Such a scheme relies on codings, as described before in Section 2.6.2, to partition the index set $\mathcal{T}$ into mutually-exclusive and collectively-exhaustive sets $\mathcal{T}_\alpha$ such that no two indices in the same set are neighbors. Then, we can simultaneously update the values at all indices in a set $\mathcal{T}_\alpha$, cycle through the sets to update all index values, and guarantee convergence as well. Such schemes, however, typically result in artifacts related to the order in which index values are updated and, hence, it is helpful to vary the coding scheme randomly after each iteration.

Owen introduced the *iterated conditional expectation* (ICE) [119, 120, 186] algorithm as a variation of the ICM procedure. The only difference between ICE and ICM is that ICE updates each intensity $x_t$ as the expectation of the posterior—the ICM updates rely on the posterior mode. The ICE update is the optimal choice, based on Bayesian decision theory, for a squared-error penalty associated with the posterior PDF [48]. In the same sense, the ICM update is optimal for a zero-one penalty [48]. The ICE algorithm modifies the update rule in Step 3 of the ICM algorithm to

$$x_t \leftarrow E\Big[P(x_t|\{x_u\}_{u\in\mathcal{T}\setminus\{t\}}, \tilde{\mathbf{x}})\Big]. \tag{2.92}$$

The ICE algorithm also possesses good convergence properties [119, 120, 186]. The ICE steady state relates to the *mean-field* approximation [186] of the MRF where the spatial interactions between RVs are approximated by the interactions between their means.

### 2.6.6 Stationarity and Ergodicity

The adaptive modeling strategy in this dissertation relies on certain assumptions on the MRF. These are, namely, the *stationarity* and *ergodicity* properties.

A *strictly stationary* [161] random field on an index set $\mathcal{T}$, defined on a Cartesian grid, is a random field where all the joint CDFs are shift-invariant, i.e.,

$$F(X_{t_1},\ldots,X_{t_n}) = F(X_{t_1+S},\ldots,X_{t_n+S}); \forall n, \forall t_1,\ldots,t_n, \forall S. \tag{2.93}$$

If the CDFs are differentiable, then it implies that all the joint PDFs are also shift invariant, i.e.,

$$P(X_{t_1},\ldots,X_{t_n}) = P(X_{t_1+S},\ldots,X_{t_n+S}); \forall S, \forall n, \forall t_1,\ldots,t_n. \tag{2.94}$$

A strictly-stationary MRF implies that the Markov statistics are shift invariant, i.e.,

$$\forall t \in \mathcal{T}, P(\mathbf{Z}_t) = P(\mathbf{Z}). \tag{2.95}$$

Such a MRF is also referred to as a *homogenous* MRF. In this dissertation, all references to stationarity imply strict stationarity.

In this dissertation, we also refer to a *piecewise-stationary* random fields, similar to the references in [175]. Through this terminology, we actually mean that the image comprises a mutually-exclusive and collectively-exhaustive decomposition into $K$ regions $\{\mathcal{T}_k\}_{k=1}^{K}$, where the data in each $\mathcal{T}_k$ are cut out from a different stationary random field.

Ergodicity allows us to learn ensemble properties of a stationary random field solely based on one instance of the random field. We use this property to be able to estimate the stationary Markov PDF $P(Z)$ from an observed image. A strictly-stationary random field $\mathbf{X}$, defined on an $m$D Cartesian grid, is *mean ergodic* [161] if the time average of $X_t$, over $t$, converges to the ensemble average $E[X_t] = \mu_X$ asymptotically, i.e.,

$$\lim_{S\to\infty} \frac{1}{(2S)^m} \int_{-S}^{S} \ldots \int_{-S}^{S} X_t dt = \mu_X. \tag{2.96}$$

A strictly-stationary random field $\mathbf{X}$ is *distribution ergodic* [161] if the indicator process $\mathbf{Y}$ defined by

$$Y_{x,t} = H(x - X_t) \tag{2.97}$$

is mean ergodic for every value of $x$. This implies that RVs in the random field are asymptotically independent as the distance between them approaches infinity [161]. This behavior is also captured in the notion of a *mixing* random field. A random field $\mathbf{X}$ on

an index set $\mathcal{T}$ is *strongly mixing* if two RVs become independent with as the distance between them tends to infinity, i.e.,

$$\lim_{\|u-v\|\to\infty} |P(X_u, X_v) - P(X_u)P(X_v)| = 0; \forall X_u, X_v \in \mathbf{X}. \tag{2.98}$$

In this dissertation, all references to ergodicity imply distribution ergodicity.

# CHAPTER 3

# ADAPTIVE MARKOV IMAGE MODELING

In many situations involving Markov modeling, the Markov PDFs or the associated Gibbs PDFs are described parametrically. This means that the functional forms for the PDFs must be known *a priori*. These forms, typically, correspond to a parameterized family of PDFs, e.g., Gaussian. Fixing the parameter values chooses one particular member of this family. The parameters for these Markov PDFs, however, are unknown. In order to choose a suitable model for the data, we need to optimally estimate the parameters from the data.

Typically, these parameterized families of PDFs are relatively simple and have limited expressive power to accurately capture the structure and variability in image data [188, 79, 91]. As a result, in many instances, the data do not comply well with such parametric MRF models. This chapter proposes a method [9, 5] of modeling the Markov PDFs *nonparametrically* and using *data-driven* strategies, in order to capture the properties underlying the data more accurately. In this way, the model is able to adapt to the data. As we saw in the previous chapter, with sufficient data, the nonparametric estimates can come very close to the underlying models. This chapter introduces the mathematics and engineering underpinning the proposed data-driven nonparametric MRF modeling scheme. The following chapters exploit this model for solving many classic image-processing problems dealing with image restoration and segmentation. The results demonstrate the success of this adaptive-MRF model, confirming that the model indeed adaptively captures the regularity in a wide-spectrum of images for a variety of applications.

## 3.1 Overview of Image Modeling

Researchers have taken different kinds of image modeling approaches including those based on (a) geometry, (b) statistics, and (c) *wavelets*. We briefly describe the characteristic features of each of these models, next.

### 3.1.1 Geometric modeling

Geometric image modeling relies on the interpretation of an image as a function defined on a grid domain. Such models describe and analyze the local spatial relationships, or geometry, between the function values via tools relying on calculus. In this way, such models invariably connect to the fields of differential geometry and differential equations. Such models treat images as functions that can be considered as points in high-dimensional *Sobolev* spaces. A Sobolev space is a normed space of functions such that all the derivatives upto some order $k$, for some $k \geq 1$, have finite $L_p$ norms, given $p \geq 1$. Modeling image functions in such spaces, however, does not accommodate for the existence of discontinuities, or *edges*, in images. Edges are formed at the silhouettes of objects and are vital features in image analysis and processing. To accommodate edges in images, two popular models exist. Mumford and Shah [110] invented the *object-edge* model assuming that the grid image domains can be partitioned into mutually-exclusive and collectively-exhaustive sets such that the resulting functions on each partition belong to Sobolev spaces. Moreover, the partitions have *regular* boundaries, not fractals, with finite lengths or areas as characterized by the Hausdorff measure. In this way, the partition boundaries can coincide with the edges in the image, segmenting the image into continuous functions that belong to Sobolev spaces. Rudin, Osher, and Fatemi [145] proposed the *bounded-variation* image model where they assumed images to possess bounded variation. Both these image models, however, impose strong constraints on the data and do not apply well to textured images. To explicitly deal with textured images, researchers have proposed more sophisticated image models that decompose an image into the sum of a piecewise-constant part and an oscillatory texture part. Such models are known as *cartoon-texture models* [13].

### 3.1.2 Statistical modeling

Statistical models, on the other hand, aim to capture the variability and dependencies in the data via joint or conditional PDFs. Specifically, they treat image data as realizations of random fields. A prominent example of such models is the MRF model [99] that we discussed in Section 2.6. Such models are good at capturing the regularities in natural images that are rich in texture-like features.

### 3.1.3 Wavelet modeling

From yet another perspective, images are formed as a superposition of local responses from some kind of sensor elements. Moreover, they exhibit such phenomena at multiple scales [59]. These local dependencies at multiple scales are well captured, mathematically as well as empirically, by the *wavelet*-based models [45, 102]. Some limitations of these methods stem from the choice of the particular wavelet decomposition basis as well as the parametric models typically imposed on the wavelet coefficients.

Although these models may seem diverse, there exist many theoretical connections between them at a high level. For instance, some wavelet-based image processing techniques relate to regularity-based schemes in certain Besov spaces [26], and some statistical schemes relying on MRFs relate to variational schemes via the Gibbs formula in statistical mechanics [26].

The fundamental concept in this dissertation, the idea of nonparametric modeling of Markov PDFs, is not entirely new. In the past, however, such approaches involve supervision or training data where many observations from the unknown MRF are available *a priori* [131, 50, 172]. The novelty in this dissertation, though, is that we derive the MRF model unsupervisedly from the given input data itself and process the images based on this model. In this way, we are able to design unsupervised adaptive algorithms for many classic image-processing problems. Furthermore, we have applied these algorithms to many new relevant applications to produce results that compete with, and often further, the current state-of-the-art. During the process of applying the nonparametric MRF model for image processing, we have also tried to provide some new theoretical insights into statistical and information-theoretic image processing.

Popat and Picard [131] were the first to employ nonparametric MRF image models. They model the Markov PDFs via clustering-based nonparametric density estimation, unlike the kernel-based Parzen-window scheme underlying the proposed approach. They exploit their model for image restoration, image compression, and texture classification. Their learning approach, however, relies on training data, which limits its practical use. In contrast, the proposed method learns the Markov statistics of the image directly from the input data.

Learning Markov statistics nonparametrically entails estimation of PDFs in high-dimensional spaces. For instance, for a first-order local neighborhood having 6 voxels, i.e., 2 neighbors along each cardinal axis, we need to estimate PDFs on a 7D space (center voxel along with its 6 neighbors). Lee *et al.* [91] as well as deSilva and Carlsson [40] analyze the statistics of $3 \times 3$ pixel neighborhoods, in 2D images, in the corresponding 9D spaces, and find the data to be concentrated in clusters and low-dimensional manifolds exhibiting nontrivial topologies. If we consider the neighborhood intensities as observations derived from a MRF, then the inherent structure of their distribution closely relates to the regularity captured by the Markov PDFs.

The literature on texture modeling also sheds light on the proposed modeling scheme. Elfadel and Picard [52] demonstrate the explicit connection between co-occurrence matrices for image intensities and the Gibbs PDFs for MRFs. Specifically, the nonlinear Gibbs energy is equivalent to a linear combination of co-occurrence measures over the Markov neighborhood. The proposed modeling technique employs Parzen-window density estimation, a generalization of co-occurrences, to estimate the Markov PDFs. Some texture-synthesis algorithms rely on learning Markov statistics from a sample texture image to construct new images having the same Markov statistics as the input texture [41, 189, 50, 172]. Levina [98] proves that the empirically-learned Markov statistics converge asymptotically to the true texture statistics. This proof of convergence is also applicable towards the nonparametric learning of the Markov statistics in the proposed method. Paget [122] presents a nonparametric multiscale MRF framework to learn Markov statistics from a sample texture for synthesizing novel texture images.

## 3.2 Data-Driven Nonparametric Markov Statistics

In order to rely on image samples to produce nonparametric estimates of Markov statistics, we must assume that different neighborhood-intensities in the image are derived from the same PDF. Mathematically, this is the notion of *stationarity* associated with a random field. A stationary region $\mathcal{R} \subset \mathcal{T}$ is one where the Markov PDFs $P(\mathbf{Z}_t)$ are exactly the same for all voxels $t$ in that region [47, 161], i.e.,

$$\forall t \in \mathcal{T}, P(\mathbf{Z}_t) = P(\mathbf{Z}). \tag{3.1}$$

In other words, the Markov statistics are shift invariant. Stationarity provides many observations $\{\mathbf{z}_t\}_{t\in\mathcal{R}}$, all derived from $P(\mathbf{Z})$.

Stationarity alone, however, is not sufficient to provide accurate estimates of the Markov PDFs from a single observed image. To do this, we must rely on another statistical property, namely *ergodicity*. Essentially, ergodicity guarantees accurate estimation of certain *ensemble* properties of the random field, e.g., the Markov PDFs $P(\mathbf{Z})$, from observations $\{\mathbf{z}_t\}_{t\in\mathcal{R}}$ in a *single* realization of the stationary random field, i.e., the observed image. Mathematically, it guarantees that, for certain quantities associated with $P(\mathbf{Z})$, the spatial averages (i.e., over $\mathcal{R}$) converge to the ensemble averages (i.e., over $\mathbf{z}$) as the size of the image $|R|$ tends to infinity [161]. Ergodicity achieves this by ensuring that: (a) random variables become independent as the shift between them approaches infinity, and (b) the random variables in the MRF become progressively less dependent with increasing spatial distance at a sufficiently-rapidly rate. Therefore, spatial averages over sufficiently-large regions $\mathcal{R}$ appear as averages of nearly-independent random variables and, subsequently, the weak law of large numbers [161] ensures the convergence of such averages to the desired ensemble average.

To represent the Markov PDFs $P(\mathbf{Z})$, we use the nonparametric *Parzen-window* technique [125, 48]. The Parzen-window probability estimate for $P(\mathbf{z})$ is defined as the ensemble average

$$P(\mathbf{z}) = \frac{1}{|\mathcal{S}'|} \sum_{\mathbf{z}'\in\mathcal{S}'} G_d(\mathbf{z} - \mathbf{z}', \Psi_d), \tag{3.2}$$

where $\mathcal{S}'$ is a *random sample* [47, 161] drawn from the PDF $P(\mathbf{Z})$, $d = |\mathcal{N}_t|$ is the neighborhood size, and $G_d(\mathbf{z}; \Psi_d)$ is the $d$-dimensional Gaussian kernel with zero mean

and covariance matrix $\Psi_d$. Having no *a priori* information on the structure of $P(\mathbf{Z})$, we choose an isotropic Gaussian kernel, i.e.,

$$\Psi_d = \sigma^2 I_d, \tag{3.3}$$

where $I_d$ is the $d \times d$ identity matrix and $\sigma$ is the standard deviation along each dimension. Ergodicity enables us to approximate the ensemble average as a spatial average:

$$P(\mathbf{z}) \approx \frac{1}{|\mathcal{A}|} \sum_{t \in \mathcal{A}} G_d(\mathbf{z} - \mathbf{z}_t; \Psi_d), \tag{3.4}$$

where the set $\mathcal{A}$ is a small subset of $\mathcal{R}$. Taking $\mathcal{A} = \mathcal{R}$ increases the algorithmic complexity of the scheme. Section 3.5.1 describes an effective technique of choosing this Parzen-window sample. As we saw in Section 2.4, the density estimate varies with the kernel-parameter $\sigma$ value and Section 3.4 describes a data-driven technique to estimate an optimal kernel-parameter $\sigma$ value.

## 3.3 Consistency of the Data-Driven Markov Model

The power of the Markov model on the random field and nonparametric density estimation comes with some additional theoretical constraints that warrant mention. In order for the Parzen-window estimation to converge [125, 48] the kernel parameter $\sigma$ must decrease with increasing number of samples. This relationship can be derived from the actual data, and several authors have proposed ML-based schemes for estimating $\sigma$ [15, 62]. Section 3.4 discusses this in more detail.

Another important issue is *consistency*. A consistent system is one where the joint PDF $P(\{X_t\}_{t \in \mathcal{T}})$ of all the random variables gives, using rules of probabilistic inference, each conditional PDF $P(X_t|\mathbf{y}_t)$ uniquely. Besag's proof of the Hammersely-Clifford theorem [14], also known as the Markov-Gibbs equivalence theorem, shows that the conditional Markov PDFs $P(X_t|\mathbf{y}_t)$ must be restricted to a specific form in order to give a consistent structure to the entire system.

The Markov PDFs that the proposed method learns empirically from the data do, indeed, yield a consistent system asymptotically, i.e., as the amount of data tends to infinity. This follows from the convergence of the Parzen-window density estimate to

the true Markov PDF. This convergence, however, holds only when the *observations* in the sample are independently generated from a single underlying PDF. The stationarity of the Markov random field implies that all observations are derived from a single PDF. However, in our case, these observations are the neighborhood-intensity vectors, which may share neighboring voxel values. Independence requires sampling from a subset $\mathcal{U}$ of the entire voxel-set $\mathcal{T}$, such that no two voxels in the subset have overlapping neighborhoods, i.e.,

$$\begin{aligned} \mathcal{U} &\subset \mathcal{R}, \\ \forall a, b \in \mathcal{U} &: \mathcal{N}_a \cap \mathcal{N}_b = \phi, \end{aligned} \tag{3.5}$$

The constraint of nonoverlapping neighborhoods leads to a wastage of a large amount of data ($\{\mathbf{z}_t\}_{t\in\mathcal{T}\setminus\mathcal{U}}$) [14], which would, in practice, lead to too few image samples. However, Levina [98] shows that ergodicity allows convergence even in the case of overlapping data, and thus it is appropriate to derive the sample $\mathcal{A}$ from the entire set of image neighborhoods in $\mathcal{R}$.

## 3.4 Optimal Parzen-Window Kernel Parameter

The nonparametric Parzen-window scheme for estimating Markov PDFs entails setting an appropriate value for the kernel-parameter $\sigma$. Section 3.3 described a ML-based estimate for this parameter and discussed the theoretical advantages of such a strategy. A maximum likelihood estimate for $\sigma$ is equivalent to the choice that minimizes the entropy of the Markov statistics of the stationary-ergodic random field. That is,

$$\begin{aligned} \sigma^* &= \operatorname*{argmax}_{\sigma} \prod_{t\in\mathcal{R}} P(\mathbf{z}_t;\sigma) \\ &= \operatorname*{argmax}_{\sigma} \sum_{t\in\mathcal{R}} \log P(\mathbf{z}_t;\sigma) \\ &\approx \operatorname*{argmin}_{\sigma} \sum_{\mathbf{z}'\in\mathcal{S}'_\sigma} \Big( -\log P(\mathbf{z}';\sigma)\Big) \\ &= \operatorname*{argmin}_{\sigma} E_{P(\mathbf{Z};\sigma)}\Big[ -\log P(\mathbf{Z};\sigma)\Big] \\ &= \operatorname*{argmin}_{\sigma} h(\mathbf{Z};\sigma), \end{aligned} \tag{3.6}$$

where $\mathcal{S}'_\sigma$ is a random sample derived from the PDF $P(\mathbf{Z};\sigma)$, and $h(\mathbf{Z};\sigma)$ is the $\sigma$-dependent entropy of the random variable $\mathbf{Z}$. Indeed, the relationship between log-

likelihood and entropy is well-documented in the literature [170]. We use the iterative Newton-Raphson optimization scheme [137] to find the optimal $\sigma$ value.

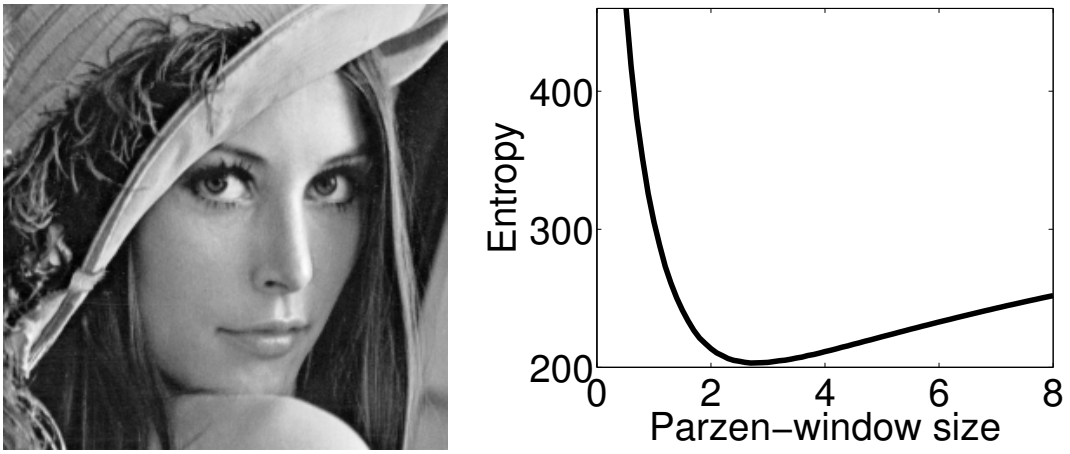
It is important to note that a naive application of ML estimation results in $\sigma = 0$ leading to a highly irregular PDF of little use. Careful observation shows that computing $P(\mathbf{z}_t)$ using a sample $\mathcal{A}$ that includes $\mathbf{z}_t$ produces an optimal kernel-parameter $\sigma$ estimate of zero [70, 29, 135]. This is because $\sigma = 0$ places impulse functions at each of the observations $\{\mathbf{z}_t\}_{t\in\mathcal{T}}$, thereby maximizing their each probability $P(\mathbf{z}_t)$. The resulting PDF estimate $P(\mathbf{Z})$, a superposition of impulse functions, is highly irregular/rough and has little practical utility. Therefore, in order to regularize the PDF estimate we ensure that, while computing $P(\mathbf{z}_t)$, the set $\mathcal{A}$ does not contain the observation $\mathbf{z}_t$, i.e.,

$$\begin{aligned} P(\mathbf{z}_t) &\approx \frac{1}{|\mathcal{A}_t|}\sum_{u\in\mathcal{A}_t} G_d(\mathbf{z}_t - \mathbf{z}_u; \Psi_d), \text{ where} \\ \mathcal{A}_t &\subset \mathcal{R}, \text{ and} \\ t &\notin \mathcal{A}_t. \end{aligned} \tag{3.7}$$

This method of regularization is called *cross validation* and we employ this scheme throughout this dissertation. It is known to be versatile, producing effective density estimates in a variety of situations [49, 151, 63, 70, 29]. Chow *et al.* [29] prove the consistency of the resulting nonparametric data-driven density estimator. The cross-validation-based PDF estimate, however, is also known to undersmooth the density estimate at times and is sensitive to outliers [151, 156].

Other schemes such as *plug-in* bandwidth estimators perform more smoothing, but at the risk of missing subtle features in the PDF [156]. This is an example of the classic tradeoff between robustness and sensitivity. As Simonoff [156] puts it: data-driven smoothing-parameter selection remains a controversial issue where no specific method is accepted as the *gold standard*. Figure 3.1 shows the variation of the entropy measure as a function of $\sigma$ for the standard *Lena* image.

Alternative strategies for regularization of the PDF estimate include spline-based methods [156] and incorporation of roughness penalties via the first/second derivatives of the logarithm or square-root of the PDF. For instance, Good and Gaskins [66, 67] derive such a derivative-based roughness penalty by penalizing the KL-divergence between




(a) (b)

**Figure 3.1**. Optimal kernel bandwidth. (a) The *Lena* image. (b) The entropy estimate for the *Lena* image as a function of Parzen-window kernel $\sigma$.

the estimated PDF and its shifted version. The resulting $\sigma$ estimates are known as *penalized-ML* estimates.

## 3.5 Engineering Enhancements

### 3.5.1 Parzen-Window Sampling Schemes

This section discusses practical, effective strategies for choosing the sample $\mathcal{A}$ during the Parzen-window density estimation.

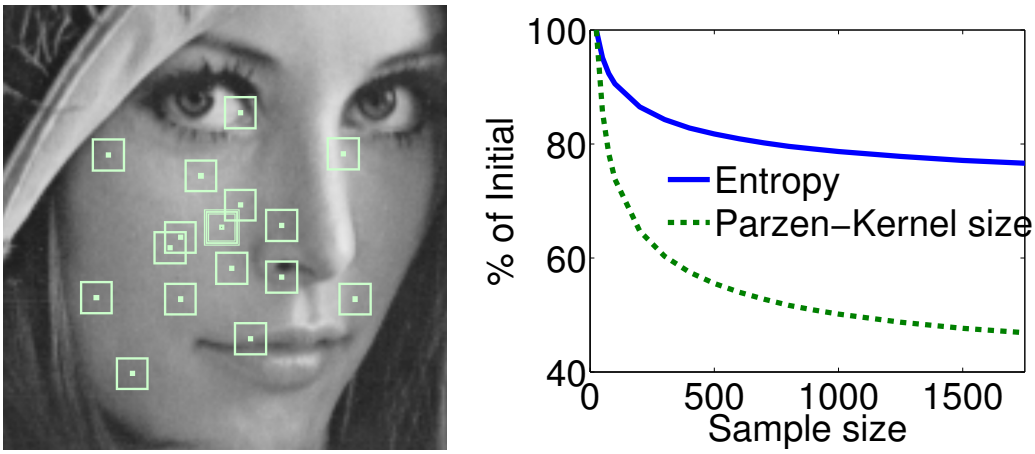
For images that conform very well to the stationarity assumption, we proposed the following strategy. To estimate the probability $P(\mathbf{z})$, we construct $\mathcal{A}$ as a random sample uniformly distributed over $\mathcal{R}$. We call this the *global-sampling* strategy. The random selection results in a stochastic approximation for the PDFs that alleviates the effects of spurious local maxima introduced in the finite-sample Parzen-window density estimate [170]. The uniform sampling works well for certain applications, e.g., while dealing with textured images which, by definition, are derived from stationary MRFs.

We have found that most image statistics are not stationary and, in practice, are more consistent in proximate regions in the image than between distant regions. In other words, images are better approximated as realizations of *piecewise* stationary-ergodic MRFs [175]. To account for this, we use a *local-sampling* strategy. In this local-sampling

framework, for each voxel $t$, we draw a unique random sample $\mathcal{A} = \mathcal{A}_t$ from an isotropic Gaussian PDF, defined on the image-coordinate space, with mean at the voxel $t$ and variance $\sigma^2_{\text{spatial}}$. Thus, the sample $\mathcal{A}_t$ is biased and contains more voxels near the voxel $t$ being processed. Experiments show that the method performs well for any choice of $\sigma_{\text{spatial}}$ that encompasses more than several hundred voxels. Figure 3.2(a) shows a local random sample for a particular pixel of the *Lena* image.

### 3.5.2 Parzen-Window Sample Size

Section 3.4 described that we chose the ML (or, equivalently, minimum entropy) value of the Gaussian-kernel standard-deviation $\sigma$ . We have found that for sufficiently large sample size $|\mathcal{A}_t|$, the choice of $\sigma$ is not sensitive to the value of $|\mathcal{A}_t|$, thereby enabling us to automatically set $|\mathcal{A}_t|$ to an appropriate value before the processing begins. Figure 3.2(b) depicts this behavior. Thus, given the Markov neighborhood and the local-sampling Gaussian variance, the method chooses the critical Parzen-window kernel parameters $\sigma$ and $|\mathcal{A}_t|$ automatically in a data-driven fashion using information-theoretic metrics.




(a) (b)

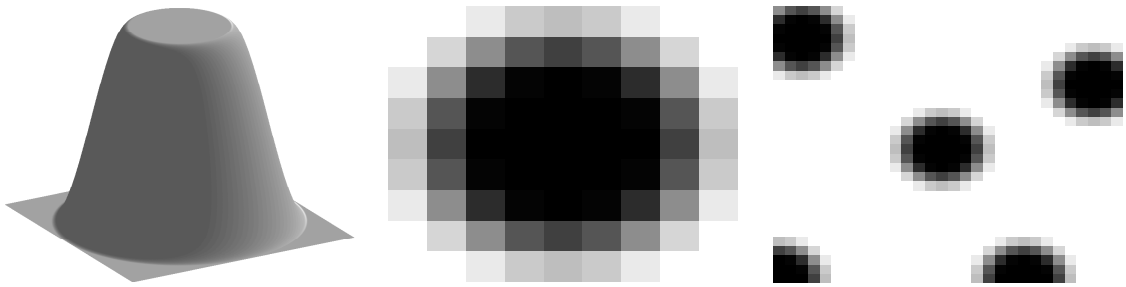
**Figure 3.2**. Parzen-window sampling. (a) Some pixels in $\mathcal{A}_t$ (black dots) along with the neighborhoods (squares around the dots) that form the Parzen-window sample for pixel $t$ (square with thickest edges). (b) The entropy of the Markov PDF and the optimal $\sigma$ are almost unaffected for $|\mathcal{A}_t| > 1000$. (To give smoother curves, each measurement, for a particular $|\mathcal{A}_t|$, is averaged over three different random sets $\mathcal{A}_t$).

### 3.5.3 Neighborhood Shape for Rotational Invariance

Hypercube-shaped neighborhoods $\mathcal{N}_t$ produce results with undesirable artifacts exhibiting preferences for grid-aligned features. A solution is to weight the intensities, making neighborhoods more isotropic. We incorporate such fuzzy weights by using an anisotropic feature-space distance metric, $\| \mathbf{z} \|_M = \sqrt{\mathbf{z}' M \mathbf{z}}$, where $\mathbf{z}'$ is the transpose of $\mathbf{z}$, and $M$ is a diagonal $d \times d$ matrix with the elements being the appropriate weights on the influence of the neighbors on the center pixel. Figure 3.3(a)-(b) shows the disk-shaped mask that we use in this dissertation. The intensities near the center are unchanged ($M(i,i) = 1$) while the intensities near the corners are weighted down $M(i,i) < 1$ gradually, via cubic-spline interpolation, to zero. The proposed isotropic mask is a grayscale version of the DUDE [175] strategy of using a binary disc-shaped mask for discrete (half-toned) images.

### 3.5.4 Neighborhood Shape for Handling Image Boundaries

Typical image boundary conditions, e.g., replicating pixels or toroidal topologies, can produce neighborhoods that distort the feature-space statistics. We handle boundary neighborhoods by collapsing the feature space along the dimensions corresponding to the neighbors falling outside the image. We crop the square regions crossing image boundaries and process them in the lower-dimensional subspace, as in Figure 3.3(c). This strategy results in important modifications in the image-processing algorithms. First, the



(a) (b) (c)

**Figure 3.3**. Neighborhood shapes. (a) Preserving rotational invariance via a neighborhood mask consisting of a flat central circular plateau with cubic splines on the sides. (b) The discrete sampling of the mask (black $\equiv$ 1, white $\equiv$ 0) for a 9 $\times$ 9 pixels neighborhood. (c) Anisotropic neighborhoods at boundaries.

cropped intensity vectors are processed based on the Markov PDFs only in the particular subspace where they reside. Second, we choose the optimal Parzen-window kernel parameter $\sigma$ based only on the observations $\mathbf{z}_t$ at indices where the neighborhoods are not cropped.

### 3.5.5 Neighborhood Size

The size of the neighborhood seems to be a modeling issue, where the mathematics may not give optimal solutions by itself. The choice of the neighborhood size typically stems from the prior knowledge, either scientific or empirical, about the physical process being modeled. For most applications in the dissertation, we have used a $9 \times 9$ pixel neighborhood. However, for certain applications, e.g., texture segmentation in Chapter 7, this neighborhood size may not work for some images. We can alleviate the sensitivity of the model to the neighborhood size by considering a multiscale adaptive-MRF model. Such a model relies on the assumption of MRFs at each level or scale of a specific multiscale image pyramid [122]. Even in such a case, some important engineering tasks persist including (a) which image decomposition to use, (b) how many levels to use in the pyramid, and (c) the size of the Markov neighborhood at each level. This dissertation does not focus on a multiscale-MRF model and such an advancement forms an important part of future work.

## 3.6 Discussion

As we will show in subsequent sections, the key ideas in this section apply well in practice. Nevertheless, we can further improve the proposed method via some engineering advances. For instance, the method of nonparametric density estimation with single-scale isotropic Parzen-window kernels is, perhaps, one of the simplest such schemes. Parzen-window density estimation can improve by choosing kernels adaptively to accommodate the signal or noise. This, however, introduces a risk of *overtraining*. If we try to learn the subtle features in the data when the amount of data is insufficient, then we could end up learning the local noise patterns in feature space. The PDFs learned will not *generalize* well to predict the structure underlying the image data.

An intrinsic limitation of the model is that its performance degrades for image regions not having stationary statistics, because that is the assumption underpinning the adaptive-MRF model. Nevertheless, one of the interesting empirical outcomes of this dissertation is that the model, and the performance of the algorithms based on the model, performs well even as these conditions are relaxed.

All algorithms in this dissertation entail computation of the Markov probabilities an $O(1)$ times for processing each pixel. This makes the algorithmic complexity of methods based on this adaptive-MRF image model as $O(|\mathcal{T}||\mathcal{A}_t|E^D)$ where $D$ is the image dimension and $E$ is the extent of the neighborhood along a dimension. This grows exponentially with increasing $D$ and, for many applications in this dissertation, the long computation times limit our experiments to 2D images. The literature suggests some improvements for faster density estimation, e.g., reduction in the computational complexity via the improved fast-gauss transform [185]. Such an approach entails approximating the PDFs in the feature-space by grouping or clustering important chunks of feature-space vectors offline. Thus, although this preprocessing phase is computationally extensive, subsequent density estimates can be computed very fast. Most of our applications, however, take few iterations of processing—around five on average—and, hence, the performance gains by a direct application of the improved fast-Gauss transform are significantly offset by the increase in preprocessing time. Another alternative for speedup is to exploit parallelism. All algorithms proposed in this dissertation are relatively straightforward to parallelize on shared-memory-multiprocessor machines (e.g., dual-processor Pentium workstations; not distributed-shared-memory supercomputers) and shared-memory-multicore machines (e.g., those using dual-core Intel/AMD processors). In general, speedup from the parallelization will depend significantly on the locality of the data references and the cache management. For shared-memory machines with two processors, we obtain a speedup close to two.

The implications of the results in this dissertation are significant. They show that it is possible to construct nonparametric density estimates in the very high-dimensional spaces of image neighborhoods. These results also suggest that the statistical structure in these spaces captures important geometric properties of images. The adaptive-MRF

formulation also generalizes in several different ways. All of the mathematics, statistics, and engineering in the proposed adaptive-MRF modeling scheme are appropriate for any kind of densely-sampled data including data on higher-dimensional image domains and vector-valued data. Furthermore, the same scheme could easily apply to other image representations, such as image pyramids, wavelets, or local geometric features.

# CHAPTER 4

# IMAGE RESTORATION BY ENTROPY MINIMIZATION

This chapter describes a novel *unsupervised information-theoretic adaptive filter* (UINTA) for image restoration [6, 9]. UINTA restores pixels by comparing pixel values with other pixels in the image that have similar neighborhoods. The underlying formulation relies on an information-theoretic measure of goodness combined with a nonparametric model of image statistics. UINTA minimizes a penalty function that captures the entropy of the patterns of intensities in image regions. UINTA filtering, obtained as the derivation of the entropy, is nonlinear. UINTA operates without *a priori* knowledge of the geometric or statistical structure of the signal, but relies instead on some general observations about the entropy of natural images. It does not rely on labeled examples to shape its output, and is therefore *unsupervised*. UINTA automatically *learns* the true image statistics from the degraded input data and constructs a filtering strategy based on that model, making it *adaptive*. Moreover, UINTA adjusts virtually all its important internal parameters automatically using a data-driven approach and information-theoretic metrics. Because UINTA is nonlinear, nonparametric, adaptive, and unsupervised, it can restore a wide spectrum of images with very little parameter tuning.

## 4.1 Overview of Image Restoration

The literature on signal and image restoration is vast, and this chapter by no means aims at a comprehensive review. This section establishes the relationship of this work to several important, relevant areas of nonlinear image filtering. Nonlinear filtering approaches are typically based on either variational methods, leading to algorithms based on partial differential equations (PDEs), or statistical methods, leading to nonlinear estimation problems.

PDE-based image processing methods became widespread after the work of Perona and Malik [127], where they propose a modified version of the heat equation (calling it *anisotropic diffusion*) that adapted the diffusivity to image features. The anisotropic diffusion equation is also the first variation of an image energy [114, 159] that favors piecewise-constant solutions (in 1D—the situation is somewhat more complex in multiple dimensions). Because such variational approaches prefer certain image geometries, we refer to these local geometric configurations as *models*. A multitude of nonlinear PDE models have been developed for a wide variety of images and applications [143, 173], including the total variation model by Rudin *et al.* [145], PDE versions [26] of the Mumford and Shah [110] variational model, the cartoon-texture model by Vese and Osher [169], the coherence-enhancing flow by Weickert [174], and various algorithms based on level sets [118, 153, 164, 117, 26]. These nonlinear PDE models have proven to be very effective, but only for particular applications where the input data are well suited to the model's underlying geometric assumptions. Moreover, the parameter tuning is a challenge because it entails fuzzy thresholds that determine which image features are enhanced and which are smoothed away.

Statistical formulations have given rise to a wide variety of image filters. For instance, the median and other order-statistics on image neighborhoods can be quite effective [107]. Tomasi and Manduchi [166] describe a bilateral filter, which does a robust averaging in Gaussian-weighted image neighborhoods. A great deal of image processing work develops from a stochastic model of image structure given by Markov random fields (MRFs). Geman and Geman [61] exploit the equivalence between MRFs and Gibbs distributions to model images with Gibbs distributions, in which case the optimal image estimate is given as a fixed point of an iterative procedure that relies on neighborhood-dependent updates. Besag [16] and Owen [120] propose the ICM and ICE schemes, respectively, for Bayesian denoising of images in the light of *a priori* information. The conditional probabilities for image neighborhood configurations, namely *cliques*, play a similar role to the image energy in the variational approaches. The most widely-used models penalize intensity differences and simultaneously estimate hidden parameters that explicitly model intensity edges, which pushes the iterative process to-

ward piecewise-constant solutions. The cliques in the MRF approach encode a set of probabilistic assumptions (priors) about the geometric properties of the signal, and thus they are effective only when the signal conforms sufficiently well to the prior. UINTA also exploits the Markov property of the images, but in a different context. Rather than imposing a particular model on the image, UINTA *learns* the relevant conditional PDFs from the input data and updates pixel intensities to decrease the randomness of these conditional PDFs. Unlike ICM and ICE, UINTA does not employ any priors in the restoration process and is fully unsupervised.

Figure 4.1 demonstrates the effects of such strong models on image filtering [1] by showing the effects of some of the prevalent nonlinear techniques on the *Lena* image. Anisotropic diffusion (Figure 4.1(c)) restores the cheeks but introduces spurious edges near the nose and the lips. Bilateral filtering [166] tends to smooth away fine textures resulting in their elimination, e.g., on the lips in Figure 4.1(d). Both of these algorithms entail two free parameters, i.e., scale and contrast, and require significant tuning. The coherence-enhancing diffusion forces specific elongated shapes in images, as seen in the enlarged nostril and the lips' curves in Figure 4.1(e). On the other hand, Figure 4.1(f) shows the curvature flow [118, 153, 117], which is very similar to the total variation strategy of [145], tends to shrink features by rounding them off. The *Lena* image, which appears to be a very typical grayscale photograph, does not adhere very well to the basic geometric models underlying these algorithms.

An alternative to filtering with variational models is to construct nonlinear transforms in the frequency domain. In this context, the wavelet literature addresses image denoising extensively. The current state-of-the-art wavelet denoising methods [133, 152, 128, 160] treat the wavelet coefficients as random variables and model their *a priori* marginal/joint PDFs parametrically. They then estimate the coefficients of the noiseless image given the observed coefficients of the noisy image via various schemes such as Bayesian estimation. The limitations of these methods stem both from the choice of the particular wavelet decomposition basis and the parametric models imposed on the coefficients. A

[1] Please refer to the electronic copy of this dissertation to analyze subtle image features.

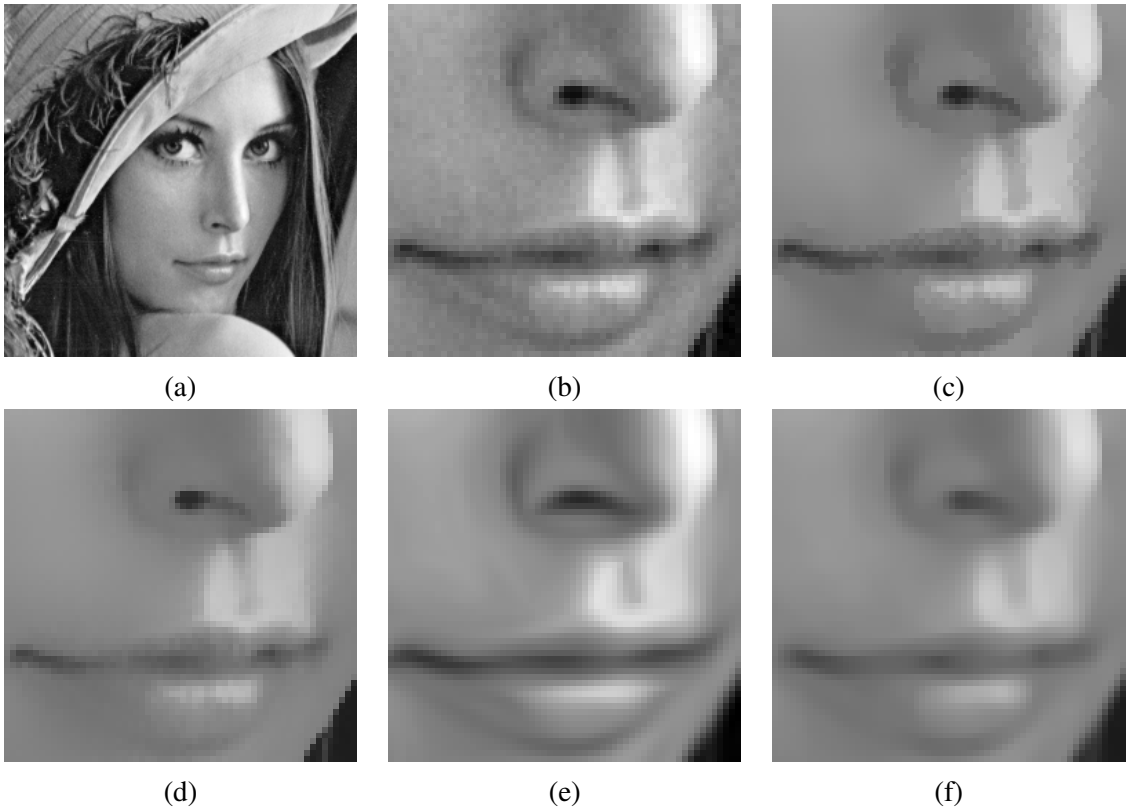



**Figure 4.1**. Comparison of UINTA with prevalent strategies. (a) Degraded *Lena* image: grayscale values range:0–100 grayscale unit (G.U.). Zoomed insets of: (b) the degraded image; (c) anisotropic diffusion: $K$=0.5 G.U.s, 20 iterations, (d) bilateral filtering: $\sigma_{\mathrm{domain}}$=3 pixels, $\sigma_{\mathrm{range}}$=12 G.U., (e) coherence-enhancing diffusion: $\sigma$=0.1 pixels, $\rho$=2 pixels, $\alpha$=0.0001, $C$=0.0001, 15 iterations, and (f) curvature flow: time step=0.2, 8 iterations.

very recent work [132] aims at the blind removal of correlated Gaussian noise using Gaussian-scale-mixture signal models in the wavelet domain. It adapts to the noise statistics by estimating the noise covariance from the input image. The sparse-code shrinkage strategy [80] chooses the transformation based on the statistical properties of the data, using noiseless training data, in order to concentrate the energy in only a few components and then shrinking the sparse component values in a manner similar to wavelet-based methods.

Weissman *et al.* [175] propose the DUDE algorithm that addresses the problem of denoising data sequences generated by a discrete source and received over a discrete, memoryless channel. It assumes no knowledge of the source statistics and yet performs

(asymptotically) as well as any denoiser (e.g., one that knows the source statistics), thereby making DUDE *universal*. DUDE assigns image values based on the similarity of neighborhoods gathered from image statistics, which resembles the construction of conditional probabilities in UINTA. However, the DUDE approach does not account for noise in the neighborhoods that are used to condition the probabilities for the reconstruction, and it is limited to discrete-valued signals. Motta *et al.* [109] extend DUDE to handle continuous-tone images, corresponding to large number of discrete intensity levels, with i.i.d. additive Gaussian noise.

The literature shows several statistically-based image processing algorithms that do rely on information theory. The *mean-shift* algorithm [60, 155, 27, 32, 10] modifies image intensities so that they move uphill on the PDF associated with the grayscale histogram of the image. At steady state (assuming appropriate windowing strategies) all samples converge to the nearest mode. The mean-shift procedure, thus, can be said to be a *mode seeking* process. However, the mean-shift algorithm operates only on image intensities (be they scalar or vector valued) and does not account for neighborhood structure in images. Thus, mean shift resembles a kind of data-driven thresholding process, particularly in the algorithm proposed by [32], in which the density estimate is static as the algorithm iterates. We show the mathematical relationship between the mean-shift procedure and entropy reduction, thereby establishing UINTA as a generalization of the mean-shift algorithm, which incorporates image neighborhoods to reduce the entropy of the associated conditional PDFs.

Buades *et al.* [22, 23], in their work that was developed simultaneously with this dissertation, propose a *nonlocal means* (NL means) algorithm for image denoising that computes the denoised image intensity as a weighted average of a sample of image intensities, where the weights are derived from the neighborhoods of the pixels in the sample. Empirical analysis of their method shows that it produces denoised images having a low degree of correlation in the difference image between the noisy image and the denoised image. The intensity updates in their method are based on the expectation of the conditional Markov PDF $P(X_t|\mathbf{y}_t)$ and closely resemble those in UINTA. However, their method contains a free parameter that defines the weights in the weighted-average

update, unlike UINTA which automatically tunes this parameter optimally in a data-driven manner. Furthermore, UINTA is iterative and arrives at the intensity updates via an entropy-reduction scheme coupled with a stochastic-relaxation approach using MRFs. The NL means algorithm could be considered a special case of the UINTA algorithm involving a single iteration and a user-defined Gaussian kernel width. Empirical comparisons show that UINTA typically produces better results than NL means, both quantitatively and qualitatively, at the cost of increased processing time.

## 4.2 Restoration via Entropy Reduction on Markov Statistics

UINTA models images as derived from stationary MRFs. Thus,

$$\forall t \in \mathcal{T}, P(\tilde{X}_t, \tilde{\mathbf{Y}}_t) = P(\tilde{X}, \tilde{\mathbf{Y}}) = P(\tilde{\mathbf{Z}}). \tag{4.1}$$

Degraded images, by definition, have less regularity in the Markov statistics as compared to their original nondegraded versions. This increases the randomness associated with the Markov PDF $P(\tilde{\mathbf{Z}})$ or the conditional Markov PDFs $P(\tilde{X}|\tilde{\mathbf{y}}_t)$ at each pixel $t$. In simpler words, degradations reduce the *predictability* of pixel values from the values in their neighborhoods. UINTA attempts to counter the degradations by increasing this regularity. One measure of randomness associated with a PDF is the entropy [34] and, hence, UINTA attempts to restore images by reducing the entropy of the stationary Markov PDF $P(\tilde{\mathbf{Z}})$.

The choice of entropy as the optimization measure is also consistent with several other observations. If we assume i.i.d. additive zero-mean noise, the addition of two independent random variables, i.e., the signal and additive noise, increases the entropy [154, 34]. Entropy reduction reduces the randomness in corrupted PDFs and tries to counteract noise. Of course, continued entropy reduction might also eliminate some of the normal variability in the signal (original image). However, we have found that non-degraded images tend to have very low entropy relative to their degraded counterparts. Therefore, entropy reduction first affects random degradations substantially more than the signal. Furthermore, the entropy reduction is limited by the entropy-based stopping criterion, as described in Section 4.5.

The UINTA strategy is to reduce the entropy $h(\tilde{\mathbf{Z}})$ of the Markov PDF by manipulating the pixel values $\{\tilde{x}_t\}_{t\in\mathcal{T}}$. This requires the entropy of the Markov PDFs $P(\tilde{\mathbf{Z}})$ to be expressed as a function of each pixel value $\tilde{x}_t$. This follows naturally from the Parzen-window density-estimation technique, based on the proposed adaptive-MRF image model. Thus, all pixel-neighborhood values $\tilde{\mathbf{z}}_t = (\tilde{x}_t, \tilde{\mathbf{y}}_t)$ in the image are observations that participate in defining the PDFs.

To update every pixel value in order to reduce the entropy, UINTA employs a gradient-descent strategy. Note that a gradient descent on $h(\tilde{\mathbf{Z}}) = h(\tilde{X}, \tilde{\mathbf{Y}})$ has components corresponding to both the center-pixel value $\tilde{x}_t$, and the neighborhood values $\tilde{\mathbf{y}}_t$. Thus, at each pixel $t$, a gradient-descent scheme can potentially update the entire region $(\tilde{x}_t, \tilde{\mathbf{y}}_t)$. In practice, however, we update only the center-pixel value $\tilde{x}_t$, i.e., we project the gradient onto the direction associated with the center pixel.

## 4.3 The UINTA Algorithm

The high-level algorithm for UINTA is as follows:

1. The input degraded image $\tilde{\mathbf{x}}$ comprises a set of intensities $\{\tilde{x}_t\}_{t\in\mathcal{T}}$, neighborhoods $\{\tilde{\mathbf{y}}_t\}_{t\in\mathcal{T}}$, and regions $\{\tilde{\mathbf{z}}_t\}_{t\in\mathcal{T}} = \{(\tilde{x}_t, \tilde{\mathbf{y}}_t)\}_{t\in\mathcal{T}}$. These values form the initial estimate $\hat{\mathbf{x}}^0 = \tilde{\mathbf{x}}$ of a sequence of images $\hat{\mathbf{x}}^0, \hat{\mathbf{x}}^1, \hat{\mathbf{x}}^2, \ldots$.
2. At iteration $m$, compute

$$\forall t \in \mathcal{T}, \frac{\partial h(\hat{X}_t^m, \hat{\mathbf{Y}}_t^m)}{\partial \hat{x}_t^m} \equiv \frac{\partial h(\hat{X}_t | \hat{\mathbf{y}}_t^m)}{\partial \hat{x}_t^m}. \tag{4.2}$$

Each $\tilde{x}_t$ undergoes a gradient descent based on the entropy of the Markov PDF estimated from $\mathcal{A}_t$. The gradient descent is

$$\begin{aligned} \frac{\partial \hat{x}_t}{\partial \tau} &= -\frac{\partial h(\hat{X}, \hat{\mathbf{Y}})}{\partial \hat{x}_t} \\ &\approx \frac{1}{|\mathcal{T}|} \frac{\partial \log P(\hat{x}_t, \hat{\mathbf{y}}_t)}{\partial \hat{x}_t} \\ &= \frac{1}{|\mathcal{T}|} \frac{\partial \log P(\hat{x}_t | \hat{\mathbf{y}}_t)}{\partial \hat{x}_t} \end{aligned}$$

$$= -\frac{1}{|\mathcal{T}|}\frac{\partial \hat{x}_t}{\partial \hat{\mathbf{z}}_t}\sum_{s\in\mathcal{A}_t}\left(\frac{G_d(\hat{\mathbf{z}}_t-\hat{\mathbf{z}}_s,\Psi_d)}{\sum_{u\in\mathcal{A}_t}G_d(\hat{\mathbf{z}}_t-\hat{\mathbf{z}}_u,\Psi_d)}\Psi_d^{-1}(\hat{\mathbf{z}}_t-\hat{\mathbf{z}}_s)\right) \quad (4.3)$$

where $\partial\hat{x}_t/\partial\hat{\mathbf{z}}_t$ is a projection operation that projects a $d$-dimensional vector $\hat{\mathbf{z}}_t$ onto the dimension associated with the center pixel intensity $\hat{x}_t$, and $\tau$ is a dummy evolution parameter. Figure 4.2 elucidates this process.

3. Construct the new image $\hat{\mathbf{x}}^{m+1}$, using gradient descent with first-order finite forward differences:

$$\forall t\in\mathcal{T}, \hat{x}_t^{m+1} = \hat{x}_t^m - \lambda\frac{\partial h(\hat{X}_t^m|\hat{\mathbf{y}}_t^m)}{\partial\hat{x}_t^m}, \quad (4.4)$$

where $\lambda$ is the time step associated with the gradient descent. Section 4.5 explains more about the choice of $\lambda$.

4. Check stopping criteria, as explained in Section 4.5. If not done, go to Step 2, otherwise the latest image estimate $\hat{\mathbf{x}}^{m+1}$ is the output.

## 4.4 Generalizing the Mean-Shift Procedure

The mean-shift procedure [60, 155, 27, 32, 57] moves each point in a feature space to a weighted average of other points using a weighting scheme that is similar to Parzen windowing. We can also view this as moving points uphill on a PDF defined by placing a Parzen-window kernel at the points. Comaniciu and Meer [32] propose an iterative mean-shift algorithm for image intensities, where the PDF does not change with iterations, for image segmentation. Each grayscale or vector pixel intensity is drawn toward a local maximum in the corresponding PDF.

This section shows how UINTA relates to the mean-shift procedure. Consider, as an example, a gradient descent on the entropy of the grayscale pixel intensities. This gives

$$\begin{aligned}\frac{\partial\tilde{x}_t}{\partial\tau} &= -\lambda\frac{\partial h(\tilde{X})}{\partial\tilde{x}_t}\\ &\approx -\frac{\lambda}{|\mathcal{T}|}\sum_{s\in\mathcal{A}_t}\left(\frac{G_1(\tilde{x}_t-\tilde{x}_s,\Psi_1)}{\sum_{u\in\mathcal{A}_t}G_1(\tilde{x}_t-\tilde{x}_u,\Psi_1)}\Psi_1^{-1}(\tilde{x}_t-\tilde{x}_s)\right),\end{aligned} \quad (4.5)$$

where $\tau$ denotes the time-evolution variable. Finite forward differences, i.e.,

$$\tilde{x}_t^{m+1} = \tilde{x}_t^m - \lambda\frac{\partial h(\tilde{X})}{\partial\tilde{x}_t^m}, \quad (4.6)$$

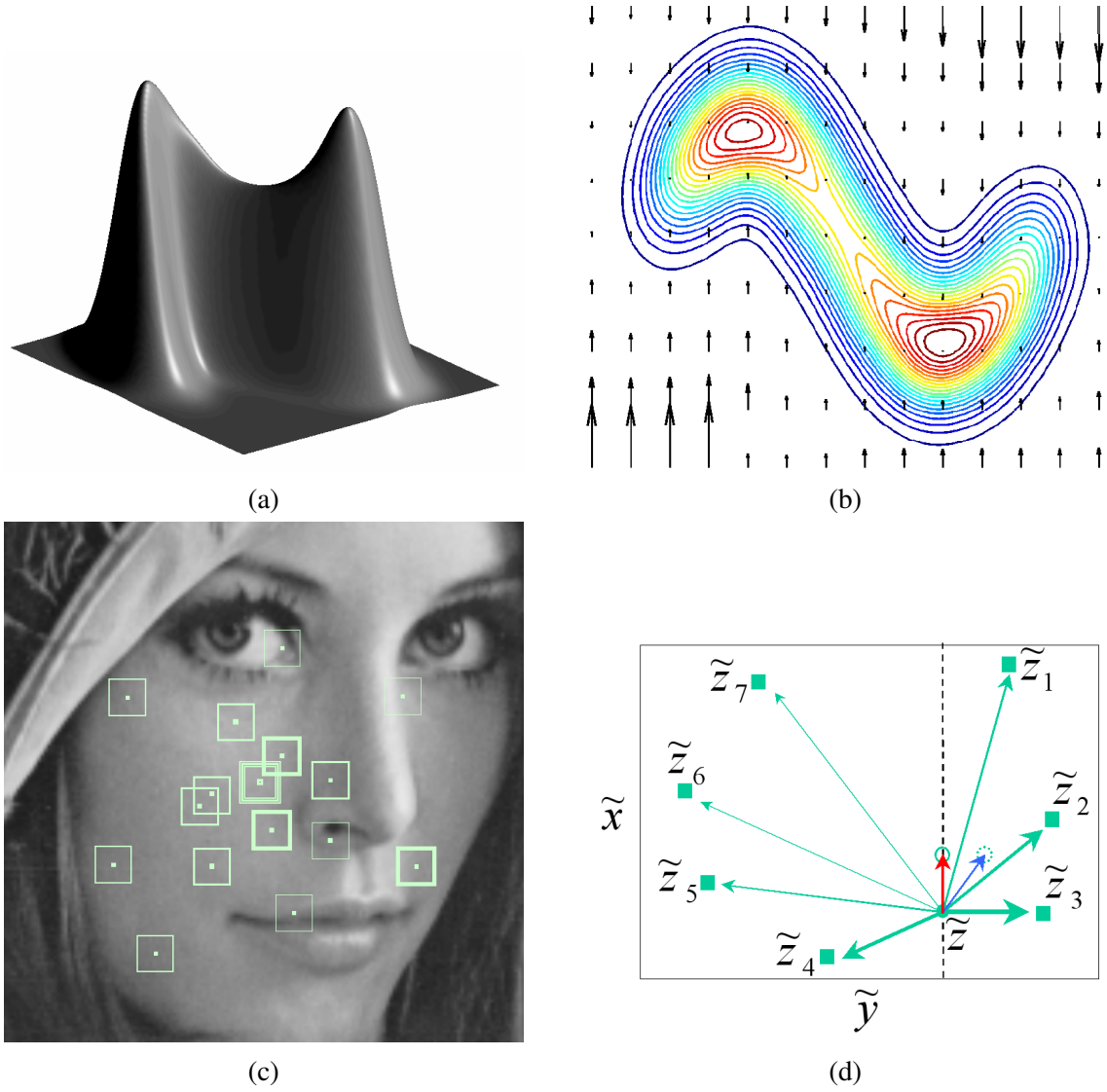



**Figure 4.2**. The mechanism for updating pixel intensities UINTA. (a) An example 2D PDF $P(\tilde{X}, \tilde{Y})$ on feature space $<\tilde{x}, \tilde{y}>$. (b) A contour plot of the PDF depicts the forces (vertical arrows) that reduce the entropy of the conditional PDFs $P(\tilde{X}|\tilde{y})$, as in (4.3). (c) Some pixels in $\mathcal{A}_t$ (black dots) along with their neighborhoods (squares around the dots) yielding feature-space observations $(\tilde{x}_t, \tilde{y}_t)$. The square thickness indicates the weights, as in (4.3), for the intensities of pixels in $\mathcal{A}_t$. The square with thickest edges denotes the neighborhood around the pixel being processed. (d) Attractive forces (arrow width $\equiv$ force magnitude) act on an observation ($(\tilde{x}, \tilde{y})$:circle) towards other observations ($(\tilde{x}_t, \tilde{y}_t)$:squares) in the set $\mathcal{A}_t$, as per (4.3). The resultant force acts towards the weighted mean (dotted circle), and the observation $(\tilde{x}, \tilde{y})$ moves based on its projection (vertical arrow).

with a time step $\lambda = |\mathcal{T}|\sigma^2$ give

$$\begin{aligned} \tilde{x}_t^{m+1} &= \tilde{x}_t^m + \left( \frac{\sum_{s\in\mathcal{A}_t} G_1(\tilde{x}_t^m - \tilde{x}_s^m, \Psi_1)\tilde{x}_s^m}{\sum_{u\in\mathcal{A}_t} G_1(\tilde{x}_t^m - \tilde{x}_u^m, \Psi_1)} - \tilde{x}_t^m \right) \\ &= \sum_{s\in\mathcal{A}_t} \tilde{x}_s^m W_s(\tilde{x}_t^m, \tilde{x}_s^m, \Psi_1) \end{aligned} \tag{4.7}$$

Each new pixel value $x_t^{m+1}$ is, therefore, a weighted average of a selection $\mathcal{A}_t$ of pixel values from the previous iteration $x_s^m$ with weights $W_s(\cdot) > 0$ such that

$$\forall t \in \mathcal{T}, \sum_s W_s(\tilde{x}_t^m, \tilde{x}_s^m, \Psi_1) = 1. \tag{4.8}$$

Taking $\mathcal{A}_t = \mathcal{T}$ gives exactly the mean-shift update proposed by Fukunaga [60]—note that UINTA updates the PDFs on which the samples climb every iteration. Thus the mean-shift algorithm is a gradient descent on the Shannon entropy [154, 34] associated with the grayscale intensities of an image. In the mean-shift algorithm each sample $\tilde{x}_t$ is being attracted towards every other sample in $\mathcal{T}$, with a weighting term that diminishes with the distance between the two samples. The UINTA updates have the same form, except that it influences the weights not only by the distances between intensities $\tilde{x}_s$, but also by the distances between the neighborhoods $\tilde{\mathbf{y}}_s$. That is, pixels in the image with similar neighborhoods have a relatively larger impact on the weighted mean that drives the updates of the center pixels.

## 4.5 Convergence

The UINTA updates are closely related to the ICM updates (described in Section 2.6.3) with a uniform prior. If the prior PDF $P(\tilde{X}_t|x_t)$ is uniform over the range of intensities in the image, then ICM guarantees convergence when every update for $x_t$ increases the probability $P_{\text{Markov}}(x_t|\mathbf{y}_t)$; where $P_{\text{Markov}}(\mathbf{Z})$ is the stationary Markov PDF. The analysis of ICM [16] shows that synchronous-update schemes, i.e., when all pixel intensities update at once as in UINTA, can cause small oscillations while other schemes can cause artifacts [16, 99].

We can show that with an appropriate update schedule as dictated by the analysis of ICM, UINTA guarantees convergence to a local mode of $P_{\text{Markov}}(\tilde{\mathbf{Z}})$. We do this by

relating the UINTA updates to the mean-shift algorithm. Cheng [27] analyzes a certain kind of mean-shift procedure, namely the *blurring process*, where the evolving sample of points redefines the nonparametric PDFs based on which it evolves. This is exactly the nature of the mean-shift procedure in UINTA where pixel-intensity updates redefine the nonparametric PDFs that are used for the next update. Based on Cheng's results (Theorem 5 in [27]), we deduce that points in the set $\mathcal{S}' = \{s'\}$ evolving based on Gaussian Parzen-window kernels converge to the local mode of a PDF $f(\cdot)$ that is estimated using Gaussian kernels on the initial set $\mathcal{S}'$. Every update brings the points $\{s'\}$ closer to the local mode of the PDF $f(\cdot)$ and, hence, increase their probability $f(s')$. In UINTA, the sample comprises image neighborhoods $\tilde{\mathbf{z}}_t$ and, hence, the nonparametrically estimated PDF $f(\cdot)$ converges to the stationary Markov PDF $P_{\text{Markov}}(\tilde{\mathbf{Z}})$. In every UINTA update, therefore, the pixel intensities $x_t$ must change such that the $P_{\text{Markov}}(x_t|\mathbf{y}_t)$ increases. Therefore, with an appropriate update schedule as dictated by the analysis of ICM, UINTA guarantees convergence to a local mode of $P_{\text{Markov}}(\tilde{\mathbf{Z}})$.

We have found that UINTA can produce small oscillations when using the mean-shift based time step $\lambda = |\mathcal{T}|\sigma^2$ together with a synchronous-updates scheme. Because other update schedules typically produce artifacts [16, 99] related to the order in which the pixels are updated, we prefer to use UINTA with the synchronous-update scheme with a smaller time step of $\lambda = 0.2|\mathcal{T}|\sigma^2$ that significantly reduces the oscillations.

An analysis of simple examples shows the existence of nontrivial steady states, e.g., an image which is a discrete sampling of a linear function such as a ramp or a binary image of a checkerboard. Empirical evidence shows that the filtering algorithm does sometimes converge to interesting results—Figure 4.3 gives two such examples where the UINTA iterations converges to a useful steady state. However, for most applications, convergence to a fixed point is not a useful goal. As with many other iterative filtering strategies, several iterations of the gradient descent are sufficient for acceptable restoration, but this requires either parameter tuning or the definition of suitable stopping criteria.

The choice of stopping criteria for this algorithm depends on a number of factors. For instance, in the absence of *any* knowledge of the signal, noise, or other types of

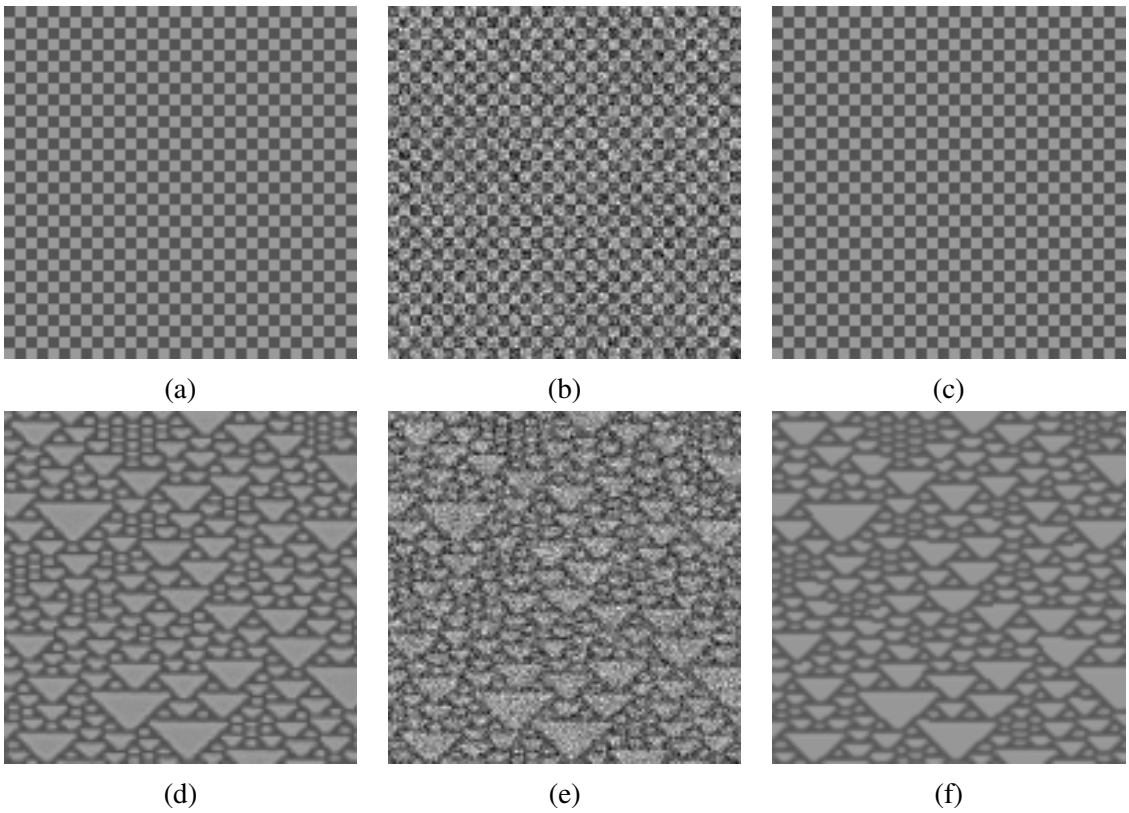



**Figure 4.3**. UINTA convergence. (a),(d) Uncorrupted images consisting of textures: a binary checkerboard image and a fractal image containing triangles. (b),(e) Images corrupted with i.i.d. additive Gaussian noise. (c),(f) Restored images that correspond to the *steady-state* of UINTA, i.e., the UINTA iterations converge to these images.

degradation, the algorithm will inevitably require some parameter tuning. We assume that noiseless images have conditional PDFs with low entropy, and degradations substantially increase this randomness. We have found empirically that entropy reduction via gradient descent starts by counteracting the randomness introduced by the noise much more than reducing the inherent randomness in the signal. Thus an effective strategy is to stop when the relative rate of change of entropy, from one iteration to the next, falls below some threshold.

When the level of additive noise is known, UINTA can iterate until the root-mean-square (RMS) difference (residual) between input and the processed image equals the noise level. We have found empirically that this method is quite effective (see Figure 4.4),

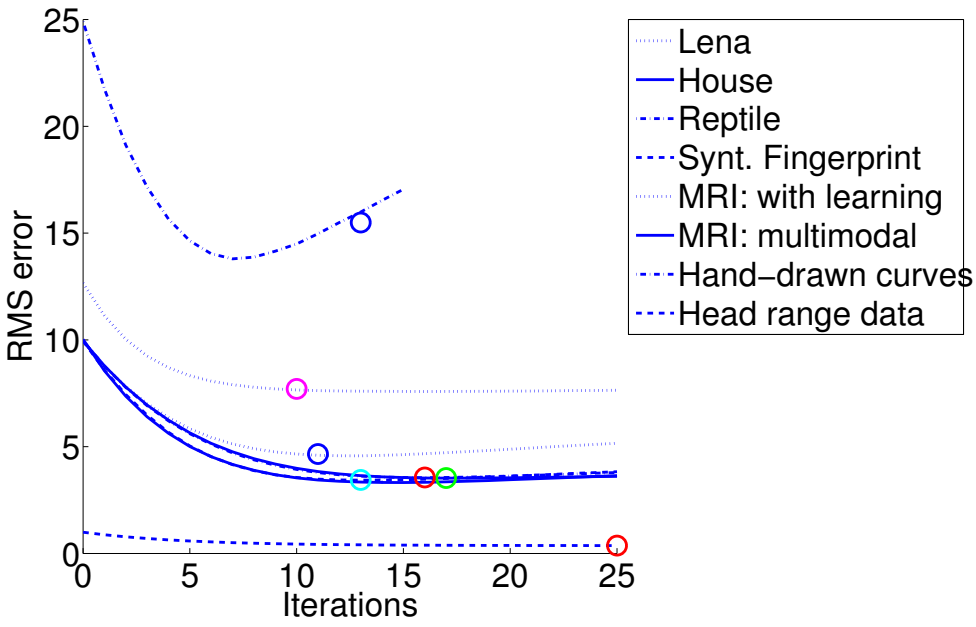



**Figure 4.4**. Root-mean-square (RMS) errors versus iterations for several images (see Section 4.6) with varying additive-noise levels. The circles represent the points where the residual equals the noise level.

and we have used this approach in all of the examples for which the Gaussian-noise levels are known.

## 4.6 Results

This section gives experimental results on numerous real and synthetic images along with the analysis of UINTA's behavior and qualitative and quantitative comparisons with the state-of-the-art wavelet methods. UINTA exposes only three parameters to the user: (i) the size $|\mathcal{N}_t|$ of the neighborhoods, (ii) the standard deviation $\sigma_{\text{spatial}}$ of the Gaussian PDF that defines the extent from which local samples are taken for density estimation (for stationary images such as textures, a global-sampling scheme will work best), and (iii) the number of iterations, or other parameters, related to the stopping criterion. Empirical results show that UINTA's performance degrades gracefully—no drastic effects as in typical PDE-based filtering schemes—for suboptimal values of these

parameters. We use masked, rotationally-symmetric, $9 \times 9$ pixels neighborhoods, as described in Section 3.5.3. Parzen windowing in all of the examples uses a local Gaussian random sampling ($\sigma_{\text{spatial}} = 40$ pixels) in the image domain with $1000$ samples (i.e., $|\mathcal{A}_t| = 1000$), as explained in Section 3.5.2. For certain experiments, we simulate i.i.d. additive Gaussian noise. We recompute the size of the Parzen window $\sigma$ after each iteration, as explained in Section 3.4. The computation for each iteration of UINTA is $O(|\mathcal{A}_t||\mathcal{T}||\mathcal{N}_t|)$. Typically, UINTA takes about $5$ iterations for the restoration. The implementation runs about twice as fast on a dual-processor shared-memory Pentium machine. For $|\mathcal{A}_t| = 1000$, it takes about $25$ seconds to process a $256 \times 256$ pixels image on a Pentium-IV $2.8$GHz dual-processor workstation. The implementation in this chapter relies on the Insight Toolkit [2].

All original (uncorrupted) images have intensities ranging from $0$ to $100$. As a visualization aid for comparing different images/results, the intensities of all images within a set have been consistently rescaled to span the available range of intensities.

Figure 4.5 shows the result of UINTA filtering on the *Lena* image. UINTA preserves and enhances fine structures, such as strands of hair or feathers in the hat, while removing random noise without imposing a piecewise-constant intensity profile. The results are noticeably better than any of those obtained using other methods shown in Figure 4.1. Figure 4.5 also shows the results of processing an MR image of a human head.

Figure 4.6 shows the result of UINTA processing on electron-microscopy data—Figure 4.7 shows the zoomed insets. These examples show UINTA's ability to adapt to a variety of grayscale features in real images approximated by piecewise-stationary models.

The fingerprint image in Figure 4.8 is an example where the degradation involves smudges (blurring), which is clearly not additive noise. UINTA enhances the light and dark lines without significant shrinkage. UINTA performs a kind of multidimensional classification of neighborhoods—therefore some features in the top-left are lost because they resemble the background more than the ridges. For the stopping criteria, we use the relative change in entropy as described in Section 4.5. Figure 4.8 also presents the results with other restoration strategies for visual comparison with UINTA. The

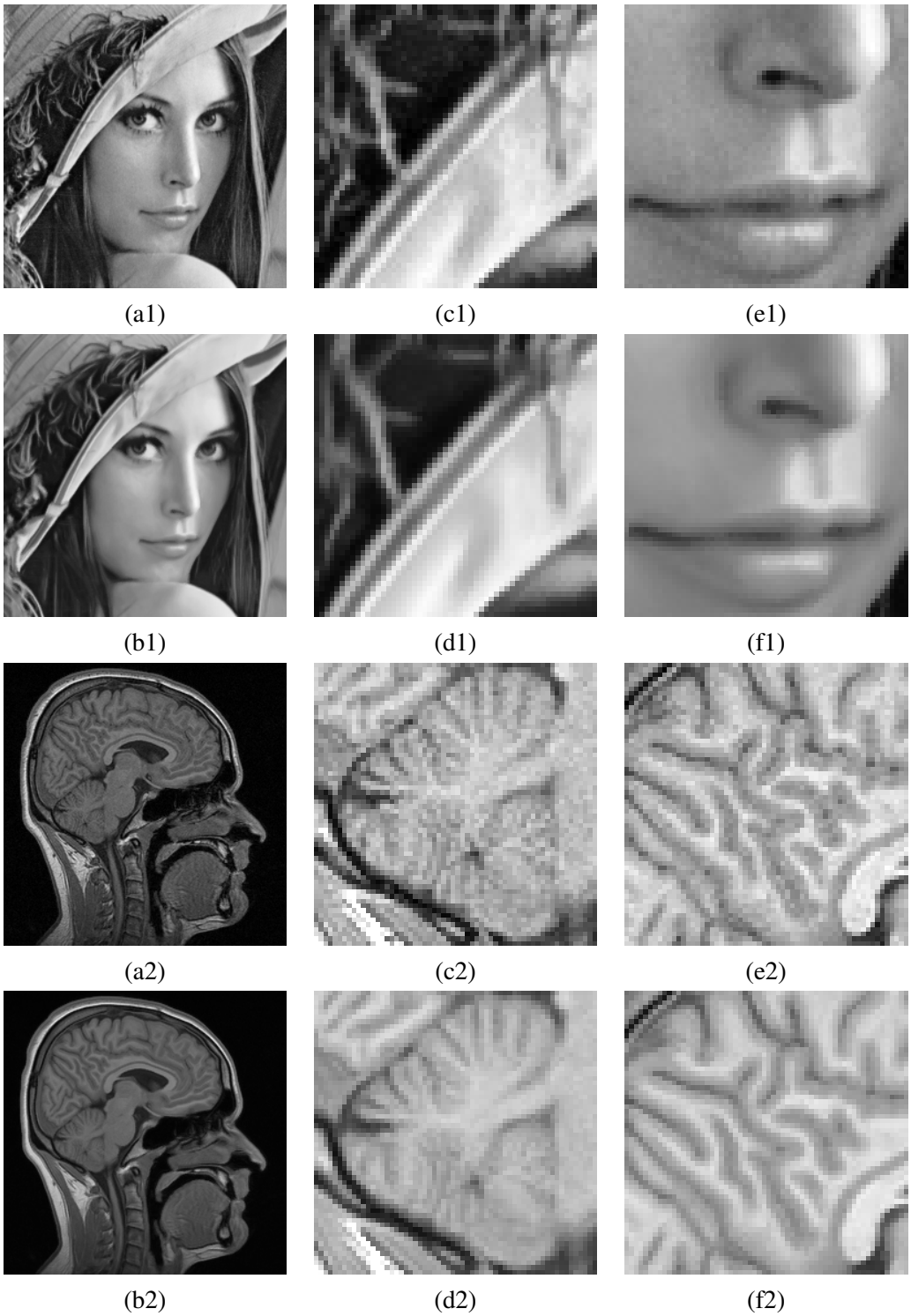
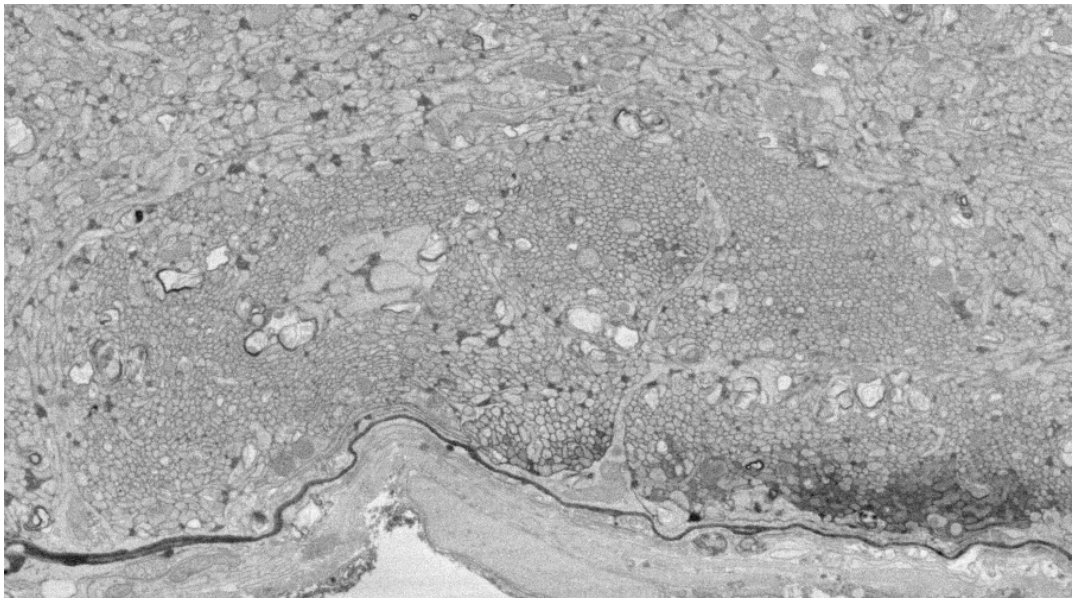
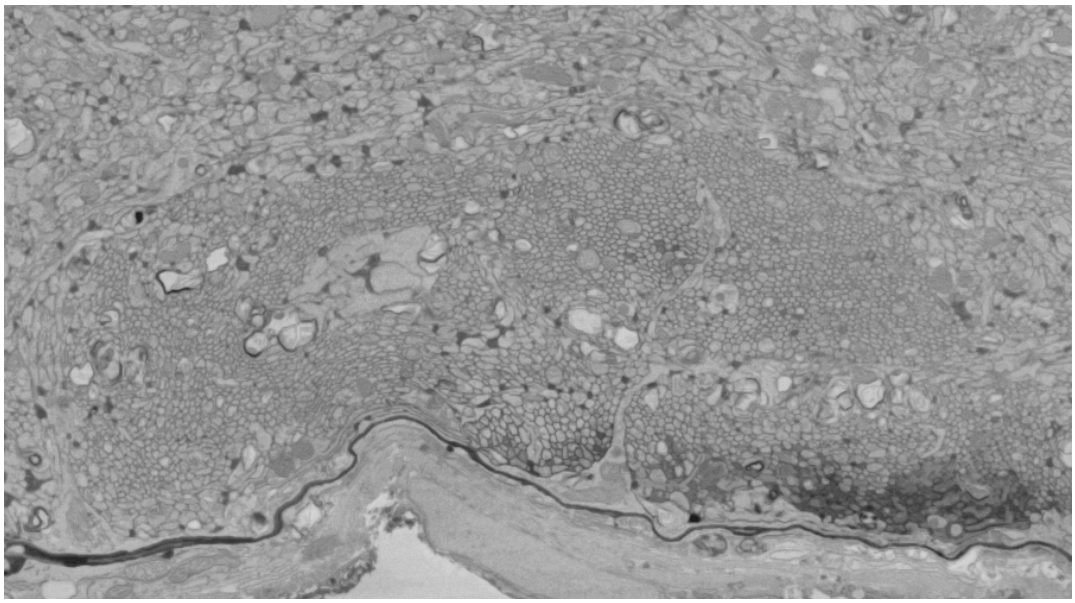



**Figure 4.5**. UINTA results. (a1),(a2) Noisy images: *Lena* and MR image of the human head. (b1),(b2) UINTA-restored images after about 5 iterations. (c1),(c2),(e1),(e2) and (d1),(d2),(f1),(f2) show magnified portions of the degraded and filtered images, respectively.

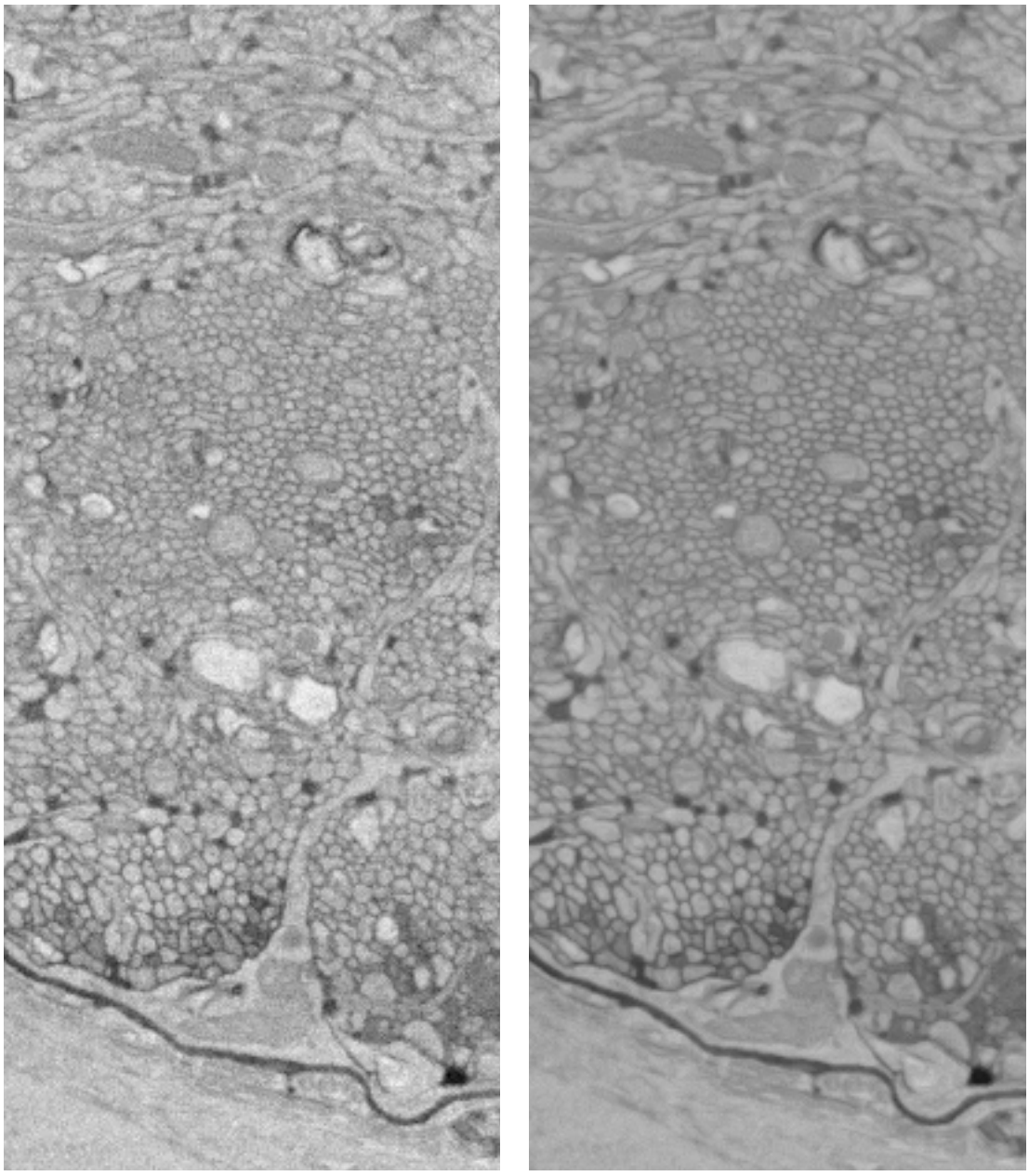


(a)



(b)

**Figure 4.6**. UINTA results. (a) Corrupted electron-microscopy image of rabbit retinal cells. (b) UINTA-restored image after 5 iterations.



(a) (b)

**Figure 4.7**. UINTA results. Zoomed insets of the (a) corrupted electron-microscopy image of rabbit retinal cells, and (b) UINTA-restored image after 5 iterations.

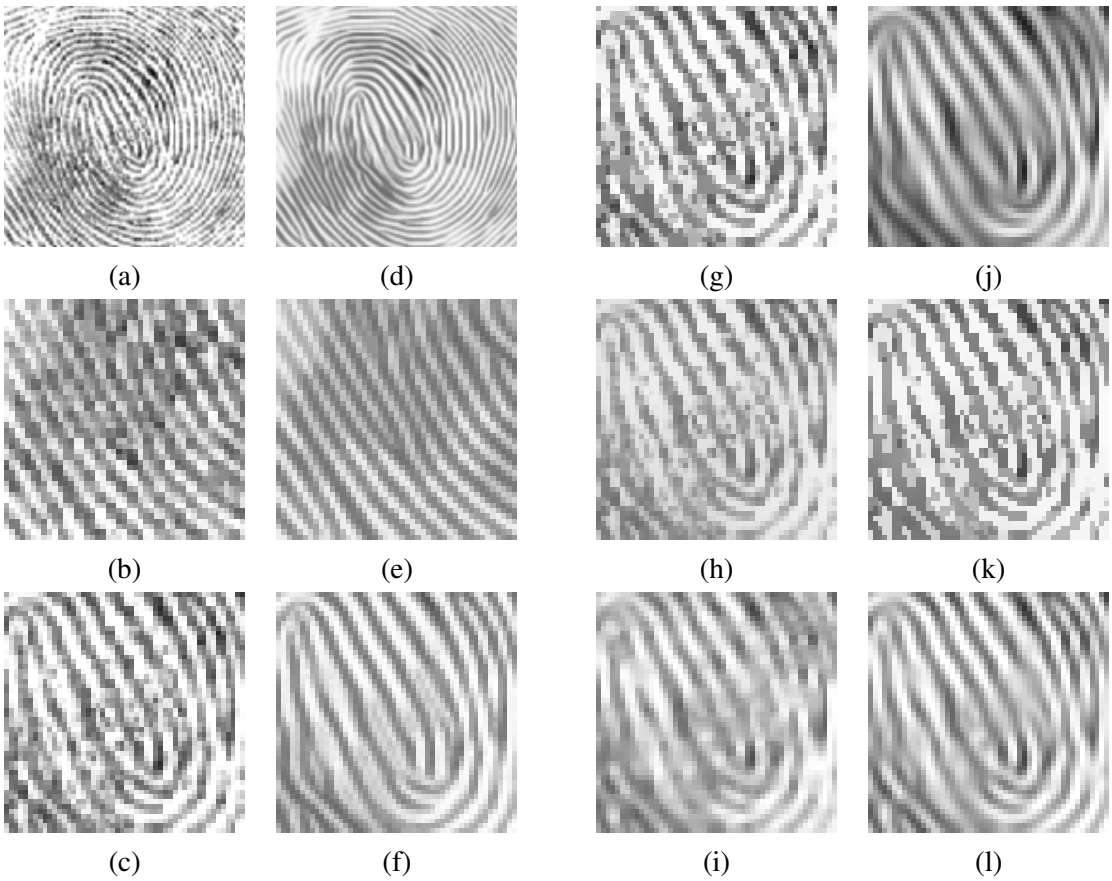



**Figure 4.8**. UINTA results compared with the state of the art. (a) Degraded fingerprint image with (b),(c) zoomed insets. (d) UINTA restored image with (e),(f) zoomed insets. Zoomed insets of the fingerprint image processed with (g) anisotropic diffusion: $K$=0.45 grayscale values, $99$ iterations, (h) bilateral filtering: $\sigma_{\mathrm{domain}}$=3 pixels, $\sigma_{\mathrm{range}}$=$15$ grayscale values, (i) curvature flow: time step=$0.2$, $5$ iterations, (j) coherence-enhancing diffusion: $\sigma$=$0.1$ pixels, $\rho$=2 pixels, $\alpha$=$0.0001$, $C$=$0.0001$, $15$ iterations, (k) unrestricted mean shift [10]: $\sigma_{\mathrm{domain}}$=2 pixels, $\sigma_{\mathrm{range}}$=$5$ grayscale values, $5$ iterations, and (l) wavelet denoising [133]: $\sigma_{\mathrm{noise}}$=$14$ grayscale values.

piecewise-smooth image models associated with anisotropic smoothing, bilateral filtering, and curvature flow (Figures 4.8(g)-(i)) are clearly inappropriate for this image. The coherence-enhancing filter (Figure 4.8(j)) does not succeed in retaining or enhancing the light-dark contrast boundaries. It also forces some elongated structures to grow or connect. An unrestricted mean-shift filtering (Figure 4.8(k)) on image intensities (with the PDF not changing with iterations) yields a thresholded image, while retaining most

of the noise. Wavelet denoising (Figure 4.8(l)) is unable to get rid of the smudges and excessively smoothes other regions of the image.

Figure 4.9 gives an example of restoring the standard *House* image [133] corrupted with i.i.d. additive Gaussian noise having variance $10^2$. The wavelet denoising technique yields a lower RMS error for this image, but introduces ringing-like artifacts in smooth regions. Table 4.1 shows the RMS errors with the standard test images of the *House*, *Lena*, *Barbara*, and *Peppers* [133].

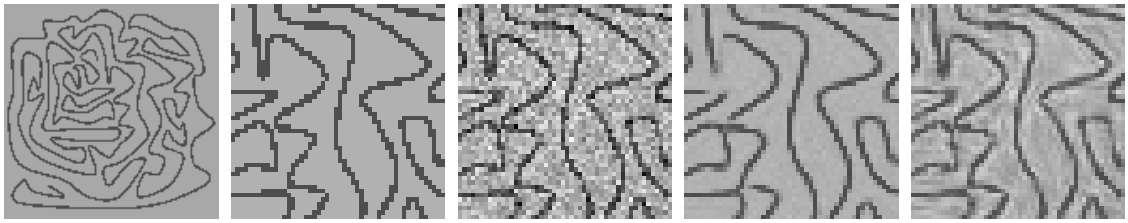
Figure 4.10 shows the application of UINTA to an image of hand-drawn curves (noise $N(0, 25^2)$). The noise level is high enough so that thresholding can not yield the original

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

**Figure 4.9**. UINTA results. (a) *House* image and its (b) zoomed inset. Zoomed insets of the (c) Noisy image. (d) UINTA filtered image. (e) Wavelet denoised [133] image.

**Table 4.1**. RMS errors comparing UINTA with the current state-of-the-art wavelet denoisers. Note: The standard test images of *Barbara* [133] and *Peppers* [133] do not appear in this dissertation. All uncorrupted images have an intensity range between $0$ and $100$.

| Example | Initial RMS error | UINTA | [133] | [152] | [128] |
|---|---|---|---|---|---|
| Standard image: *House* | 10.0 | 3.5 | 2.9 | 3.1 | 3.5 |
| Standard image: *Lena* | 10.0 | 4.6 | 3.6 | 3.8 | 4.1 |
| Standard image: *Barbara* | 10.0 | 4.8 | 3.8 | 4.2 | 4.5 |
| Standard image: *Peppers* | 10.0 | 4.5 | 3.5 | 3.7 | 3.9 |
| Hand-drawn curves | 25.0 | 15.4 | 16.0 | 18.5 | 18.0 |
| Simulated fingerprint | 10.0 | 3.4 | 4.1 | 4.7 | 4.7 |
| Simulated range data (head) | 1.0 | 0.35 | 0.34 | 0.36 | 0.5 |
| *Reptile* | 10.0 | 3.5 | 2.9 | 3.0 | 3.4 |
| Building Facade | 10.0 | 4.5 | 4.4 | 5.1 | 5.4 |
| MRI (with learning) | 10.0 | 3.1 | 3.4 | 3.7 | 3.9 |
| MRI (multimodal) | 10.0 | 3.3 | 3.4 | 3.7 | 3.9 |

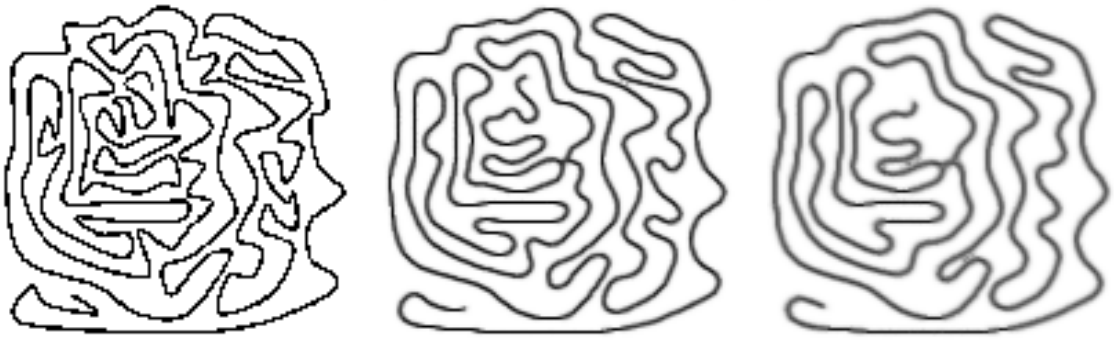


(a) (b) (c) (d) (e)

**Figure 4.10**. UINTA results. (a) Hand-drawn curves with a (b) zoomed inset. Zoomed insets of the (c) noisy image, (d) UINTA-filtered image, and (e) wavelet-denoised [133] image.

image. UINTA learns the pattern of black-on-white curves and forces the image to adhere to this pattern. However, UINTA does make mistakes when curves become too close, exhibit very sharp bends, or when the noise introduces ambiguous gaps. The wavelet denoised image depicts significant artifacts around the edges, giving a higher RMS error (Table 4.1).

The entropy reduction associated with UINTA *does* impose a kind of statistical simplification on the image, and that statistical simplicity corresponds, in many cases, to geometric simplicity. Figure 4.11 shows the results of many UINTA iterations on the hand-drawn image of Figure 4.10(a). UINTA has no explicit geometrical model and yet it gradually smooths out the bends in these curves producing progressively simpler geometric structures. The entropy of straighter curves is lower, because of reduced



(a) (b) (c)

**Figure 4.11**. UINTA results. (a) Hand-drawn curves. (b) and (c) show UINTA filtered images after 100 and 200 iterations, respectively.

variability in the associated neighborhoods. The result is qualitatively similar to that of curvature-reducing geometric flows [118, 153, 117], suggesting a strong link between variational and statistical characterizations of images [188].

In order to better analyze the behavior of UINTA and compare its performance with state-of-the-art wavelet denoisers, we present results with a diverse collection of synthetic images. We provide examples on the simulated fingerprint image (Figure 4.12(a1)), the simulated range data of the human head (Figure 4.12(a2)), and the synthetic *Reptile* image [55] (Figure 4.12(a3)). Table 4.1 shows the RMS errors. UINTA performs better on the fingerprint, almost equally well on the range data and poorer on the *Rep-*

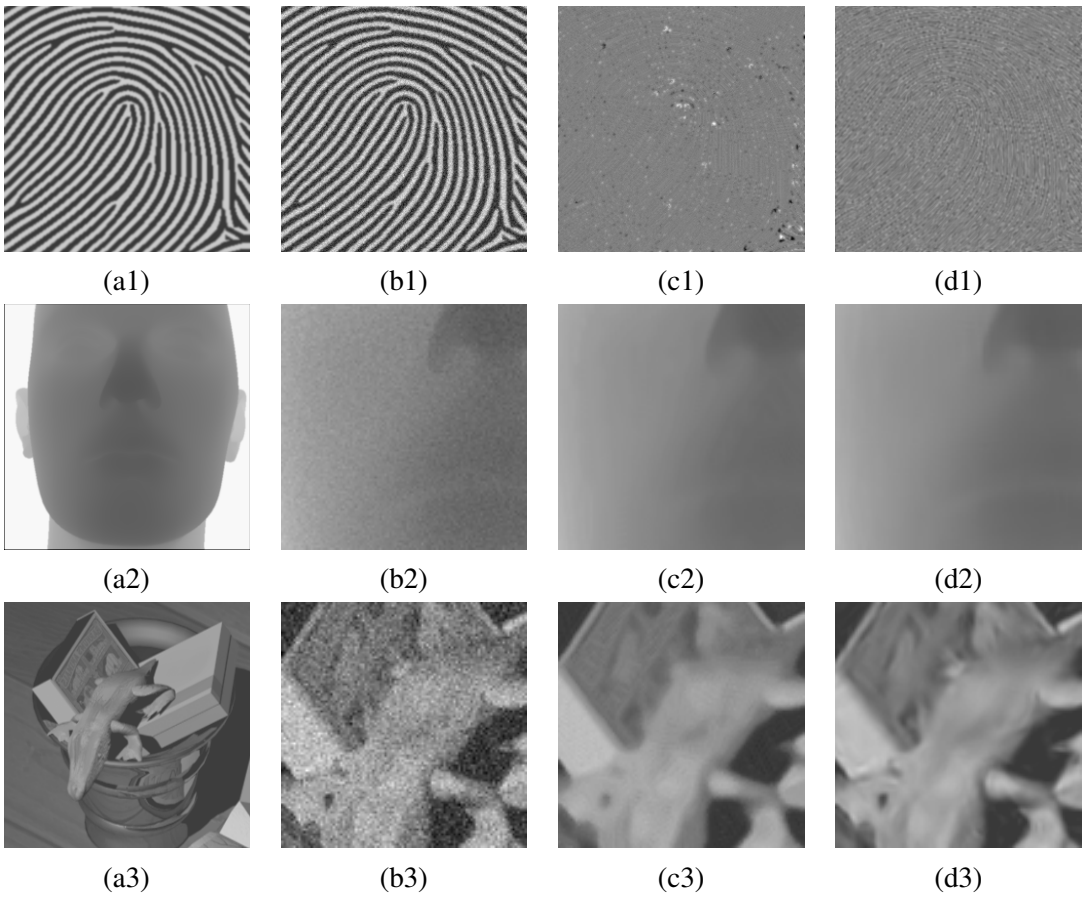
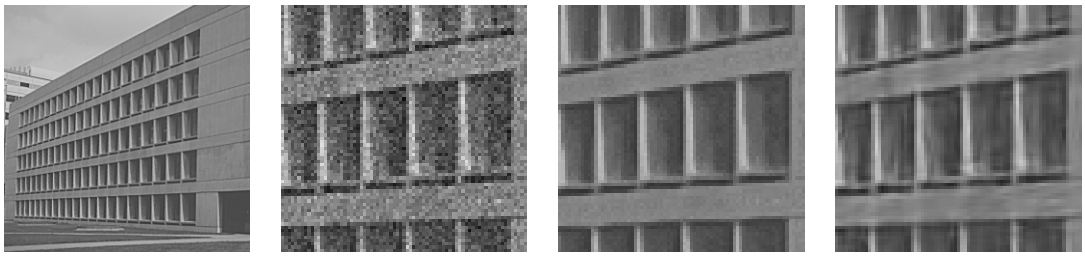



**Figure 4.12**. UINTA results. (a1) Simulated fingerprint image. (b1) Noisy image. Difference between the filtered and the noiseless images for (c1) UINTA and (d1) the wavelet denoiser [133]. (a2) Head range data. (a3) *Reptile* image [55]. Zoomed insets of the (b2)-(b3) noisy images, (c2)-(c3) UINTA filtered images, and (d2)-(d3) wavelet denoised images [133].

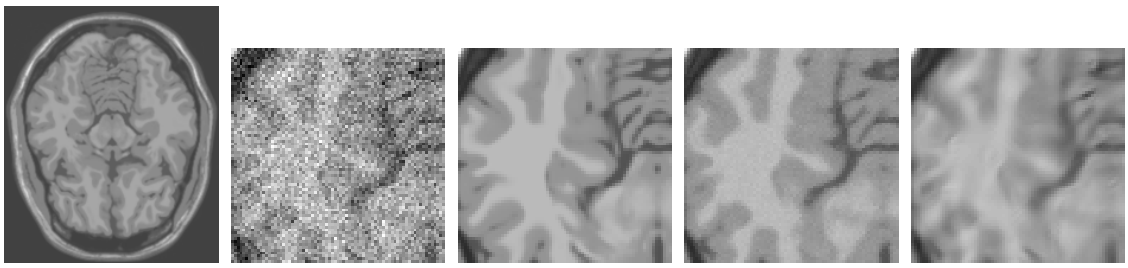
*tile* image. Thus, UINTA performs better as a denoiser when it can find sufficiently many patterns in the degraded image to be able to distinguish the degradation from the underlying signal. Indeed, this stems from the stationarity assumption on the MRF model underpinning UINTA. Moreover, the statistical models underlying the wavelet denoisers are empirically derived from photographs, similar to the *Reptile* image. Figure 4.13 shows a photograph of a building facade that exhibits a certain degree of redundancy. UINTA is able to exploit that to perform almost as well as the best wavelet denoiser in terms of RMS error (see Table 4.1) and with fewer visual artifacts.

When operating within a specific application domain, UINTA can perform much better by learning from ideal or noiseless-image examples. Figure 4.14 shows a demonstration of this concept on simulated MRI data from the BrainWeb [31] project. We corrupt a head MRI T1 image with i.i.d. additive Gaussian noise and use two other similar, but not identical, images for learning the neighborhood statistics of typical brain MR images.



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

**Figure 4.13**. UINTA results. (a) Building facade image. Zoomed insets of the (b) noisy image, (c) UINTA-filtered image, and (d) wavelet-denoised image [133].



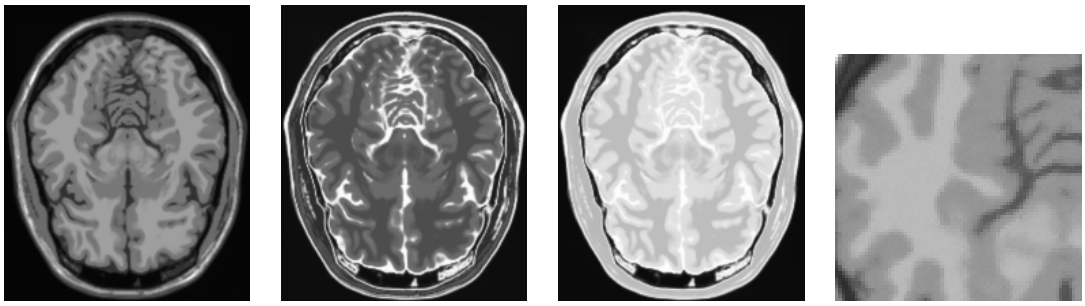
(a) (b) (c) (d) (e)

**Figure 4.14**. UINTA results in a supervised scenario. (a) Image used for learning neighborhood statistics. Zoomed insets of the (b) noisy image, (c) original image, (d) UINTA-filtered image, and (e) wavelet-denoised image [133].

Figure 4.14(a) shows one of the two images representing the nonparametric prior model. This example shows the power of such learning—the UINTA restored image exhibits structures that are barely visible in the degraded version and fares considerably better than the wavelet denoiser, both qualitatively and quantitatively.

The UINTA formulation also generalizes easily to simultaneous restoration of a sequence of images, e.g., multimodal MRI, exploiting the relationships between images to further enhance performance. Figure 4.15 shows an example with multimodal restoration. This entails a simultaneous restoration of T1, T2, and PD images in a coupled manner, treating the combination of three images as an image of vectors, and analyzing PDFs in the *combined* probability space. Although in this chapter we show results with multimodal images that are well aligned, our experiments suggest that the restoration is fairly robust to minor registration errors. Here again, UINTA fares better than the wavelet denoiser.

We now provide qualitative comparison between UINTA and the NL-means algorithm [23]. The updates in both methods have similar mathematical form. However, there are several important differences. While UINTA is iterative and formulated in an information-theoretic context, NL-means is not iterative and relies on optimal nonparametric regression estimation [156]. The derivation of the NL-means update is closely related to the ICE update. UINTA relies on a stopping criterion based on an information-theoretic or statistical optimality metric. Concerning the engineering aspects, while UINTA chooses the Parzen sample $\mathcal{A}_t$ stochastically from a Gaussian PDF over the im-



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

**Figure 4.15**. UINTA results. (a)-(c) Multimodal MR images comprising T1, T2, and PD scans. (d) Zoomed inset of the UINTA-restored image.

age coordinates, NL-means chooses $\mathcal{A}_t$ from a small square neighborhood. More importantly, while UINTA dynamically tunes the Parzen-kernel parameter $\sigma$ via a data-driven manner optimality metric, NL-means exposes this $\sigma$ as a free parameter. NL-means relies on a heuristic to tune $\sigma$ to approximately $10$-$15$ times the estimated standard-deviation of the (assumed) i.i.d. Gaussian noise in the image. UINTA, on the other hand, automatically chooses $\sigma$ to be close to the noise level.

Figure 4.16(a1)-(a4) gives some images denoised by the NL-means algorithm. Each of the original images was corrupted with $10\%$ i.i.d. additive Gaussian noise. For an accurate comparison with UINTA, we used the same $9 \times 9$ pixels neighborhood mask in NL-means as we do for UINTA (see Figure 3.3). We found that choosing $\sigma$ as $10$ times the noise level leads to extreme smoothing/averaging that destroys all significant image details. We choose $\sigma$ to be $6$ times the noise level. The RMS errors are: $3.2$ for the *House* , $4.75$ for the building facade, $3.8$ for the simulated fingerprint, and $15.9$ for the hand-drawn curves. Comparing these values with those in Table 4.1, we observe that UINTA produces better results on the three images other than the *House* image. Moreover, the edges in the NL-means-restored images appear noisy. Figure 4.16(b1)-(b4) and Figure 4.16(c1)-(c4) show the difference between degraded images and restored images (termed *method-noise* [23]) for NL-means and UINTA, respectively. While the method noise in UINTA has a wider intensity range showing poor performance for unique image structures, e.g., corners, that in NL-means appears more correlated along long edges. Figure 4.16(d1)-(d4) and Figure 4.16(e1)-(e4) show the difference between restored images and original images for NL-means and UINTA, respectively—for a perfect restoration, these images would comprise all zero values. We can observe the higher correlation along long structures in the NL-means-restored images a bit more clearly as compared to the method-noise images. Note: the method-noise images in Figure 4.16(b1)-(b4) and Figure 4.16(c1)-(c4) can be obtained by negating the images in Figure 4.16(d1)-(d4) and Figure 4.16(e1)-(e4) followed by addition of the noise.

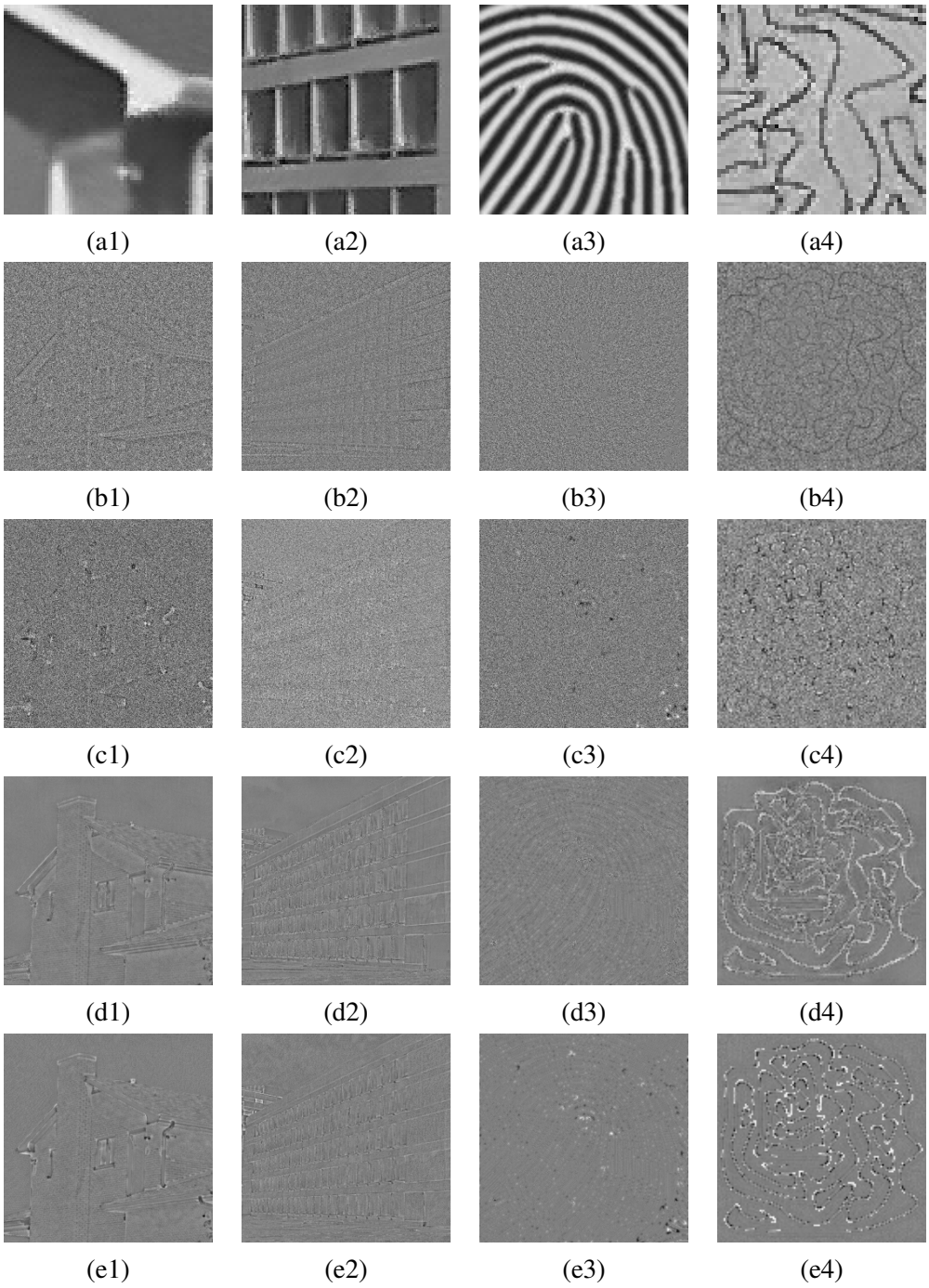



**Figure 4.16**. Comparison of UINTA with NL-Means. (a1)-(a4) Images denoised via NL-means [23]: cropped and zoomed for comparison with UINTA-restored images shown previously. (b1)-(b4) Difference between the degraded images and the restored images (*method-noise* [23]) for NL-means [23]. (c1)-(c4) Difference between the degraded images and the restored images (*method-noise* [23]) for UINTA. (d1)-(d4) Difference between the restored images and the original images for NL-means. (e1)-(e4) Difference between the restored images and the original images for UINTA.

# CHAPTER 5

# DENOISING MR IMAGES USING EMPIRICAL-BAYES METHODS

Over the last several decades, magnetic resonance imaging (MRI) technology has benefited from a variety of technological developments resulting in increased resolution, signal-to-noise ratio (SNR), and acquisition speed. However, fundamental trade-offs between resolution, speed, and SNR combined with scientific, clinical, and financial pressures to obtain more data more quickly, result in images that still exhibit significant levels of noise. In particular, the need for shorter acquisition times, such as in dynamic imaging, often undermines the ability to obtain images having both high resolution and high SNR. Furthermore, the efficacy of higher-level, post processing of MR images, including tissue classification and organ segmentation, that assume specific models of tissue intensity (e.g., homogeneous), are sometimes impaired by even moderate noise levels. Hence, denoising MR images remains an important problem. From a multitude of statistical and variational denoising formulations proposed, no particular one appears as a clear winner in all relevant aspects, including the reduction of randomness and intensity bias, structure and edge preservation, generality, reliability, automation, and computational cost.

This paper presents a novel framework for denoising MR images that relies on the adaptive Markov-random-field (MRF) image model described in [9, 5]. The work in this paper is a significant modification of our previous approach in [8]. The key idea in the modeling approach is to *adapt* or infer the model from the corrupted input data itself and subsequently process the data based on the infer model. The proposed denoising method produces an optimal reconstruction based on principles in empirical-Bayesian estimation [141, 140] and information theory. The method bootstraps itself by estimat-

ing the uncorrupted-signal Markov statistics, using an information-theoretic optimality metric, from the corrupted input data and the knowledge of the Rician noise model. It then employs the *inferred* uncorrupted-signal Markov statistics as an *adaptive prior* in a Bayesian denoising process at each pixel. In this way, it avoids the need of imposing ad hoc prior models. Furthermore, it proposes a novel iterative Bayesian-inference algorithm on MRFs that incorporates entropy reduction on posterior PDFs. We call this new approach as *iterated conditional entropy reduction* (ICER). The results demonstrate that the method denoises conservatively while ensuring the preservation of most of the important features in the brain MR images. Qualitative and quantitative comparisons with the state of the art clearly depict the advantages of the proposed method.

## 5.1 Overview of MRI Denoising

A multitude of variational methods based on partial differential equations have been developed for a wide variety of images and applications [127, 117], with some of these having applications to MRI [64, 100, 56]. However, such methods impose certain kinds of models on local image structure that are often too simple to capture the complexity of anatomical MR images. These methods, typically, do not take into account the bias introduced by Rician noise. Furthermore, such methods usually involve manual tuning of *critical* free parameters that control the conditions under which the models prefer one sort of structure over another; this has been an impediment to the widespread adoption of these techniques.

Another class of methods relies on statistical inference on multiscale representations of images. A prominent example includes methods based on wavelet transforms. Healy *et al.* [75] were among the first to apply soft-thresholding based wavelet techniques for denoising MR images. Hilton *et al.* [77] apply a threshold-based scheme for functional-MRI data. Nowak [115], operating on the square magnitude MR image, includes a Rician noise model in the threshold-based wavelet denoising scheme and thereby corrects for the bias introduced by the noise. Pizurica *et al.* [129] rely on the prior knowledge of the correlation of wavelet coefficients that represent significant features across scales. They first detect the wavelet coefficients that correspond to these signifi-

cant features and then empirically estimate the PDFs of wavelet coefficients conditioned on the significant features. They employ these probabilities in a Bayesian denoising scheme.

In our previous work [9, 6], we described UINTA which restores images by generalizing the mean-shift to incorporate neighborhood information. UINTA, however, relies neither on the knowledge of a noise model nor a prior model. Some MR-inhomogeneity correction methods are based on the quantification of information content in MR images [157, 103]. They follow from the observation that inhomogeneities increase the entropy of the 1D *gray scale* PDFs. However, entropy measures on first-order image statistics are insufficient for effective denoising; thus this paper extends the information-theoretic strategy to higher-order Markov PDFs.

The proposed method takes the *empirical-Bayes* approach [141, 140, 24], pioneered by Robbins [141, 140], for Bayesian denoising without making any ad hoc assumptions on the prior PDFs. The empirical-Bayes approach is applicable when we encounter multiple independent instances of a Bayesian decision problem (i.e., denoise each pixel) that all rely on exactly the same *fixed*, *but unknown*, prior PDF (i.e., uncorrupted-signal Markov PDF). In this special case, the empirical-Bayes approach allows accurate data-driven computation of the posterior PDF without the need to impose ad hoc or ill-fitting prior models. In this way, the decision procedure automatically *adapts* to the unknown prior PDFs. Robbins employed the empirical-Bayes approach to first obtain a *maximum likelihood* (ML) estimate of the prior distribution using the observations corrupted by a known noise model, and then employ the estimated prior model to compute the posterior [90]. The strategy in this paper closely follows Robbin's strategy.

Weismann *et al.* [175] address optimal image denoising using Markov statistics and empirical-Bayes approach [175]. Their discrete universal denoiser (DUDE) focuses on *discrete* signal intensities and, subsequently, relies on inverting the channel transition matrix (noise model) to give a closed-form estimate for source statistics from the observed statistics. The proposed method addresses continuous-valued signals, which is essential for medical-imaging applications, and thus entails estimating uncorrupted-signal statistics nonparametrically through the reduction of a *Kullback-Leibler* (KL) divergence.

Snyder *et al.* [158] also use kernel density estimators for density deconvolution. The proposed approach also presents a method for practically dealing with the nonstationarity of real MRI data.

Cordy and Thomas [33] employ the expectation-maximization (EM) algorithm [43, 104] for *deconvolving* PDFs corrupted with i.i.d. additive Gaussian noise. They model the uncorrupted-signal PDF as a Gaussian mixture model, but use the EM algorithm to estimate only the weights of Gaussians in the mixture—the means and variances of the Gaussians are tuned manually before EM is applied. They constrained the Gaussians to be spread uniformly over the entire domain of the PDF. Such a strategy, however, is not likely to be effective for density estimation in high-dimensional domains because of the enormous numbers of Gaussians needed to cover the space and sparsity of the data in the space—uniformly-distributed Gaussians will tend to oversmooth the PDF structure in high-curvature regions and will be inefficient in the tails of the PDF.

## 5.2 Bayesian Denoising by Entropy Reduction

The proposed strategy relies on several pieces of technology that interact to provide accurate, practical models of image statistics. For clarity, the discussion begins at a high level and successive sections discuss how each of these pieces is developed from the input data.

Given the noisy image $\tilde{\mathbf{x}}$, our goal is to find the *maximum-a-posteriori* (MAP) estimate $\mathbf{x}^*$ of the true image $\mathbf{x}$:

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmax}}\ P(\mathbf{x}|\tilde{\mathbf{x}}). \tag{5.1}$$

Writing the posterior as

$$P(\mathbf{x}|\tilde{\mathbf{x}}) = P(x_t|\{x_u\}_{u\in\mathcal{T}\setminus\{t\}}, \tilde{\mathbf{x}})P(\{x_u\}_{u\in\mathcal{T}\setminus\{t\}}|\tilde{\mathbf{x}}), \tag{5.2}$$

where $t$ is an arbitrary pixel, motivates us to employ an iterative restoration scheme where, starting from some initial image estimate, we update the estimate pixel-wise so that the posterior never decreases. Besag's ICM algorithm [16] gives one such strategy that updates $x_t$ to the mode of the PDF $P(x_t|\{x_u\}_{u\in\mathcal{T}\setminus\{t\}}, \tilde{\mathbf{x}})$. Finding modes of

PDFs, however, is not always straightforward or computationally efficient. Therefore, we propose a new algorithm that updates $x_t$ by moving it closer to the *local mode* of $P(x_t|\{x_u\}_{u\in\mathcal{T}\setminus\{t\}}, \tilde{\mathbf{x}})$. The proposed algorithm is similar in spirit to the ICM algorithm, but relies on entropy reduction on the PDF that updates pixel intensities by performing a gradient ascent on the logarithm of the PDF—hence called *iterated conditional entropy reduction* (ICER). The relationship between reducing Shannon's entropy of Parzen-window PDFs and gradient ascent on the logarithm of the posterior PDF is described in detail in [9, 6]. It follows that by updating intensities $x_t$ to reduce the entropy $h(x_t|\{x_u\}_{u\in\mathcal{T}\setminus\{t\}}, \tilde{\mathbf{x}})$ and bringing them closer to their local modes, we can guarantee nondecreasing values for $P(x_t|\{x_u\}_{u\in\mathcal{T}\setminus\{t\}}, \tilde{\mathbf{x}})$ and, thereby, convergence.

Let us assume for simplicity that, given the true image $\mathbf{x}$, the RVs in the MRF $\tilde{\mathbf{X}}$ are *conditionally independent* ( 2.86). Subsequently, Bayes rule gives [16]

$$\underset{x_t}{\text{argmax}}\ P(x_t|\{x_u\}_{u\in\mathcal{T}\setminus\{t\}}, \tilde{\mathbf{x}}) = \underset{x_t}{\text{argmax}}\ P(x_t|\mathbf{y}_t)P(\tilde{x}_t|x_t), \tag{5.3}$$

where $P(x_t|\mathbf{y}_t)$ is the *unknown prior PDF* and $P(\tilde{x}_t|x_t)$ is the *likelihood* as determined from the Rician noise model. We model the prior using nonparametric Parzen-window density estimates with Gaussian kernels. The next section describes a method for adaptively inferring the prior based on the input data and the knowledge of the noise model.

## 5.3 Estimating Uncorrupted-Signal Markov Statistics

A Bayesian denoising framework implicitly assumes the existence of a prior statistical model of the uncorrupted signal. We can, potentially, derive such priors from a suitable database of high-SNR brain MR images (e.g., different images of the same modality and anatomy). This effectively amounts to *training* the denoising system. Effective training data, however, are not easily available for many applications. Alternatively, we can *infer* the uncorrupted signal statistics from the observed data by making suitable assumptions. Let us assume a *fixed*, but *unknown*, Markov model $P(\mathbf{Z})$ for the uncorrupted signal that generates *all* uncorrupted data. These data, subsequently, get corrupted by Rician noise. What we observe is only the corrupted data—the prior remains unknown. However, the following analysis provides a way of inferring the prior.

Given sufficiently many corrupted observations, we can infer the Markov statistics of the corrupted signal accurately [9, 5]. With this knowledge of the corrupted-signal Markov statistics and knowing the properties of the corruption process, we can accurately estimate the uncorrupted-signal Markov statistics. In this way, we can *empirically* estimate the unknown prior PDF. This essentially amounts to solving an *inverse problem*, which we discuss in detail in the next section.

### 5.3.1 Forward Problem: Numerical Solution

Let us denote the Markov PDF of the corrupted signal by $P_C(\tilde{\mathbf{Z}})$. Let us model the Markov PDF of the uncorrupted signal using Parzen-windowing as:

$$P_{\mathcal{U}}(\mathbf{z}) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} G(\mathbf{z} - \mathbf{z}_u, \sigma), \tag{5.4}$$

where $\{\mathbf{z}_u\}_{u \in \mathcal{U}}$ denotes the means of the Gaussians and $\sigma$ their standard deviation along each dimension. This nonparametric model is a general model capable of representing arbitrary PDFs for large $|\mathcal{U}|$. The goal is to estimate the set $\{\mathbf{z}_u\}_{u \in \mathcal{U}}$ and $\sigma$, i.e., the parameters of the model, based on the knowledge of the observed corrupted-signal Markov statistics and the Rician corruption process. The key idea is as follows. An estimate of the uncorrupted-signal model parameters and the Rician noise level gives us an estimate of the corrupted-signal statistics. In the inverse-methods literature, this is the process of solving the so-called *forward problem*. We must match this estimate of the corrupted-signal Markov PDF with the Markov PDF obtained from the corrupted data by suitably updating the prior-model parameters. We use the KL-divergence measure to quantify the goodness of the match. We now analyze the noise model in detail and present a numerical scheme for solving the forward problem.

The Rician noise model corresponds to a linear *shift-variant* system whose *impulse response* for an impulse PDF located at $x \geq 0$ is

$$P(\tilde{x}|x) = \frac{\tilde{x}}{\sigma_R^2} \exp\left( - \frac{\tilde{x}^2 + x^2}{2\sigma_R^2} \right) I_0\left( \frac{\tilde{x}x}{\sigma_R^2} \right), \tag{5.5}$$

where $\sigma_R$ is the noise level and $I_0(\cdot)$ is the zero-order modified Bessel function of the first kind. For $x \gg 3\sigma_R$, Rician noise corrupts in a way very similar to additive independent

Gaussian noise. For smaller $x$, though, the effect is more complex. For a Gaussian input PDF $G(x-\mu, \sigma)$, a general analytical formulation of the output PDF makes the denoising framework very cumbersome. To alleviate this problem, we compute the system response numerically and approximate it by a Gaussian. We construct two *lookup tables* $\mathcal{L}_\mu(\cdot)$ and $\mathcal{L}_\sigma(\cdot)$ that provide the means and variances of the output Gaussians $G(x' - \mu', \sigma')$, given the means $\mu$ and variances $\sigma^2$ of input Gaussians and the noise level $\sigma_R$. We discretize the input parameters at a sufficiently-high resolution and employ bilinear interpolation to read values from the table.

We must be aware of some important issues while computing the system response. The Rician PDF $P(\tilde{x}|x)$ is defined only for nonnegative $x$. However, the Parzen-window model with Gaussian kernels extends to negative values too. This model approximates the system poorly in cases where $\sigma$ values are relatively large as compared to the magnitude of their means $\| \mathbf{z}_u \|$. In such cases, the Rician corruption process that applies only to the nonnegative part of the Gaussian input (a truncated Gaussian) and produces an output that may not be fitted well by a Gaussian. However, we can view the situation more positively because of the implications of the *central limit theorem* [167, 123, 78, 12]. This classic theorem [167, 123] states that the PDF for the sum of *independent* RVs asymptotically approaches a Gaussian. In the same vein, there exists a central limit theorem for arbitrary *dependent* RVs too [78, 12] that proves their sum to approach a Gaussian RV. The theorem concerning dependent RVs applies to the Rician corruption process—the functional form of $P(\tilde{X}|x)$ depends on $x$. In our case, while one of the RVs is a Gaussian (input PDF), the other (Rician PDF) resembles a Gaussian in general and approaches a Gaussian for specific parameter values. These facts help us obtain good fits. Figure 5.1 shows that the fitted Gaussians approximate the Rician-corrupted output PDFs reasonably well. We observe that for input Gaussians that extend significantly to the negative axis, in Figure 5.1(a)-(b), the fit is not perfect while for the other cases, the fit is close to perfect. We use a Levenberg-Marquardt curve-fitting technique [137] to fit Gaussians to the output corrupted PDFs.

Given the uncorrupted PDF $P_{\mathcal{U}}(\cdot)$ and the Rician noise level $\sigma_R$, we can approximate the corrupted-signal Markov PDF as

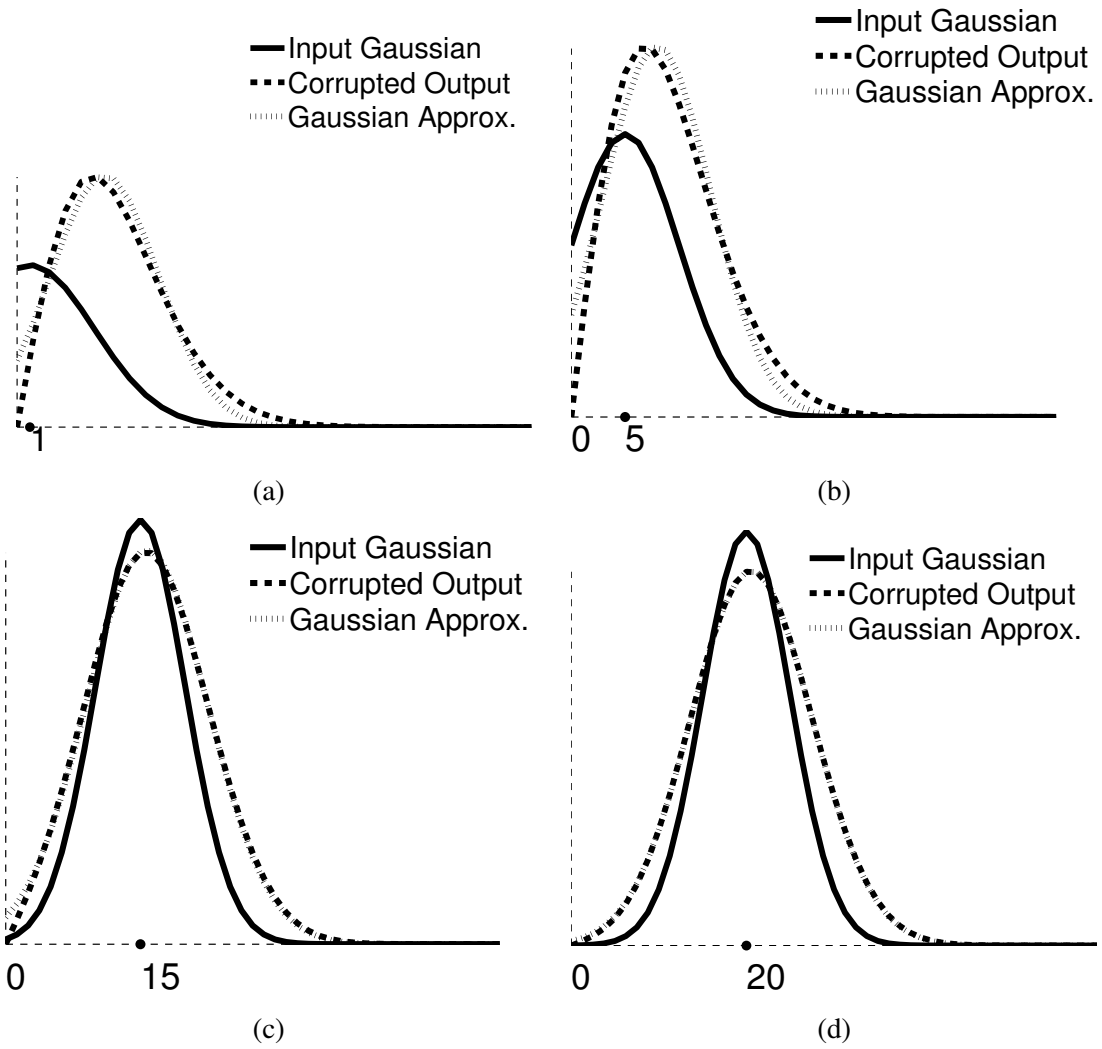



**Figure 5.1**. These graphs depict the Rician corruption process in 1D with $\sigma = 5$ and $\sigma_R$ = 5. The input Gaussian PDF is corrupted by Rician noise resulting in the output corrupted PDF. We fit a Gaussian to approximate this corrupted PDF. The graphs show this process for different means of the input Gaussian: (a) $x_u$ = 1, (b) $x_u$ = 5, (c) $x_u$ = 15, and (d) $x_u$ = 20. We have numerically found that the maximum relative error between the output and its Gaussian approximation is always less than $0.1$.

$$\hat{P}_C(\tilde{\mathbf{z}}) \approx \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} G(\tilde{\mathbf{z}} - \mathbf{z'}_u, \Psi'_u), \tag{5.6}$$

where we define the $i$-th component of the neighborhood-intensity vector $\mathbf{z'}_u$ as

$$\mathbf{z'}_u(i) = \mathcal{L}_\mu(\mathbf{z}_u(i), \sigma, \sigma_R) \tag{5.7}$$

and the entry on the $i$-th row of the diagonal covariance matrix $\Psi'_u$ as

$$\Psi'_u(i,i) = \mathcal{L}_\sigma(\mathbf{z}_u(i), \sigma, \sigma_R). \tag{5.8}$$

### 5.3.2 Inverse Problem: KL-Divergence Optimality

We want the corrupted-signal PDF $\hat{P}_C(\tilde{\mathbf{z}})$, derived from the uncorrupted-signal model $P_{\mathcal{U}}(\mathbf{Z})$, to match the Markov PDF $P_C(\tilde{\mathbf{z}})$ estimated from the observed corrupted data. We propose the Kullback-Leibler (KL) divergence as a measure of the discrepancy between the two PDFs. If we define $\Theta = \{\mathbf{z}_u\}_{u\in\mathcal{U}}$, then we want to find

$$\begin{aligned}
\{\Theta^*, \sigma^*\} &= \underset{\Theta,\sigma}{\text{argmin}}\ \text{KL}\left(P_C \parallel \hat{P}_C\right) \\
&= \underset{\Theta,\sigma}{\text{argmin}}\ E_{P_C}\left[\log \frac{P_C}{\hat{P}_C}\right] \\
&= \underset{\Theta,\sigma}{\text{argmin}}\ E_{P_C}\left[\log P_C - \log \hat{P}_C\right] \\
&= \underset{\Theta,\sigma}{\text{argmax}}\ E_{P_C}\left[\log \hat{P}_C\right] \\
&\approx \underset{\Theta,\sigma}{\text{argmax}} \sum_{t\in\mathcal{T}} \log \hat{P}_C(\tilde{\mathbf{z}}_t) \\
&= \underset{\Theta,\sigma}{\text{argmax}} \sum_{t\in\mathcal{T}} \log\left(\sum_{u\in\mathcal{U}} G(\tilde{\mathbf{z}}_t - \mathbf{z'}_u, \Psi'_u)\right).
\end{aligned} \tag{5.9}$$

What we have here is a ML optimization problem. ML estimation procedures, however, are well known to need regularization to reduces the chances of the optimization getting stuck in local maxima and to produce effective estimates, e.g., the classic *method-of-sieves* regularization by Grenander [68]. We propose to regularize the ML estimation by fixing the value of $\sigma$ beforehand. The enforcement of this regularization is similar in spirit to that used by Geman and Hwang [63] for nonparametric density estimation.

We can produce an effective optimal estimate for $\sigma$ as follows. We first find a ML-based estimate $\tilde{\sigma}$ for the nonparametric Markov PDF of the corrupted observed sample $\{\tilde{\mathbf{z}}_t\}_{t\in\mathcal{T}}$ (details in [9, 5]). We know that a significant fraction of intensities in the image are much larger than the noise level $\sigma_R$ where the Rician noise model is close to an additive independent Gaussian noise model. Therefore, we approximate

$$\sigma^* \approx \sqrt{\tilde{\sigma}^2 - \sigma_R^2}. \tag{5.10}$$

Fixing this $\sigma$ value, we subsequently obtain an optimal ML estimate for the set $\Theta$ relying on the EM algorithm. We have found that this approximation for $\sigma$ works effectively in practice.

### 5.3.3 Optimization Using the EM Algorithm

The inverse problem we have here is that of mixture-density parameter estimation—the parameter here is the set $\Theta = \{\mathbf{z}_u\}_{u \in \mathcal{U}}$ of the means of Gaussians that defines the uncorrupted-signal Markov PDF. We propose to solve this using the EM algorithm [43, 104]. The EM algorithm computes a ML parameter estimate when the data are *incomplete*, i.e., a part of the data remains unobserved or *hidden*. We now describe the key idea behind the working of the EM algorithm.

The optimization formulation in (5.9) is a little unwieldy because it contains the logarithm of a sum. If we knew which Gaussian component generated each observation, then we could obtain the probability $\hat{P}_C(\tilde{\mathbf{z}}_t)$ by evaluating a single Gaussian: the one that generated $\tilde{\mathbf{z}}_t$. The EM approach gets rid of the summation that the logarithm applies to. The key idea behind EM is that it assumes the existence of one *hidden* RV associated with each observation $\tilde{\mathbf{z}}_t$. The PDF of this hidden RV gives the probabilities for different Gaussian components to have generated $\tilde{\mathbf{z}}_t$. Let us call this RV $L$. The values of $L$ are, however, never observed. The EM algorithm starts by assuming a joint PDF $P(\tilde{\mathbf{Z}}, L)$ of the observed and hidden RVs, i.e., the *complete data*. It defines the probability of the observation $\tilde{\mathbf{z}}_t$ assuming that it came from the $l$-th Gaussian as

$$P(\tilde{\mathbf{z}}_t|l) = G(\tilde{\mathbf{z}}_t - \mathbf{z'}_l, \Psi'_l), \tag{5.11}$$

where $\mathbf{z'}_l$ and $\Psi'_l$ are the mean and covariance values, respectively, for the $l$-th Gaussian. The goal of the EM algorithm is to iteratively find the ML estimate of the parameter $\Theta$ as

$$\begin{aligned} \Theta^* &= \underset{\Theta}{\text{argmax}} \ \log P(\tilde{\mathbf{z}}|\Theta) \\ &= \underset{\Theta}{\text{argmax}} \ \log \Big( \int_{\mathcal{S}_L} P(\tilde{\mathbf{z}}, l|\Theta) dl \Big), \end{aligned} \tag{5.12}$$

where $\mathcal{S}_L$ is the support of $P(L)$. Each iteration comprises the *E* (expectation) step and the *M* (maximization) step. The E step formulates an expectation of the complete-data

likelihood function over the PDF of the hidden RV conditioned on the observed data and current parameter estimate. The M step maximizes this expectation with respect to the parameter. After much simplification [19], the maximization performed in the $m$-th iteration reduces to

$$\underset{\Theta}{\text{argmax}} \sum_{u \in \mathcal{U}} \sum_{t \in \mathcal{T}} P(u|\tilde{\mathbf{z}}_t; \Theta^{m-1}) \log P(\tilde{\mathbf{z}}_t|u, \mathbf{z}_u), \tag{5.13}$$

where $\Theta^{m-1}$ is the $(m-1)$-th parameter estimate that is held constant and $\Theta = \{\mathbf{z}_u\}_{u \in \mathcal{U}}$ is the free variable. The parameter updates guarantee no decrease in the likelihood $P(\tilde{\mathbf{z}}|\Theta)$ of the observed data and, hence, the sequence of estimates converge to a local maximum of the likelihood function.

An important element in this entire process of inferring the uncorrupted-signal Markov statistics is the initial choice of the sample $\{\hat{\mathbf{z}}_u^0\}_{u \in \mathcal{U}}$ for the EM algorithm. We initialize $\{\hat{\mathbf{z}}_u^0\}_{u \in \mathcal{U}}$ to comprise a small random fraction of the entire set of observed neighborhood-intensities $\{\tilde{\mathbf{z}}_t\}_{t \in \mathcal{T}}$, spread uniformly over the image domain $\mathcal{T}$. This ensures the representation of all important features in the image and produces an initial estimate close to the global maximum of the likelihood function.

The EM updates, for density estimation using a sum of Gaussians, are as follows.

1. Let $\{\hat{\mathbf{z}}_u^m\}_{u \in \mathcal{U}}$ be the parameter estimate at the $m$-th iteration.
2. Use the lookup tables to compute $\hat{\mathbf{z}'}_u^m$ and $\Psi'^m_u$, $\forall u \in U$, where

$$\begin{aligned} \hat{\mathbf{z}'}_u^m(i) &= \mathcal{L}_\mu(\hat{\mathbf{z}}_u^m(i), \sigma, \sigma_R) \text{ and} \\ \hat{\Psi'}_u^m(i,i) &= \mathcal{L}_\sigma(\hat{\mathbf{z}}_u^m(i), \sigma, \sigma_R). \end{aligned} \tag{5.14}$$

3. Compute

$$\forall u \in \mathcal{U}, \forall t \in \mathcal{T}, P(\tilde{\mathbf{z}}_t|u) = G(\tilde{\mathbf{z}}_t - \hat{\mathbf{z}'}_u^m, \Psi'_u) \tag{5.15}$$

4. Use Bayes rule to evaluate $P(u|\tilde{\mathbf{z}}_t), \forall t \in \mathcal{T}, \forall u \in \mathcal{U}$. Because we derive the initial set of observations $\hat{\mathbf{z}}_u^0$ from the PDF $P(\tilde{\mathbf{Z}})$ that is close to $P(\mathbf{Z})$, we can ignore the *a priori* probabilities $P(u)$—treat them equal for all $u$. Thus, we compute

$$\forall u \in \mathcal{U}, \forall t \in \mathcal{T}, P(u|\tilde{\mathbf{z}}_t) \approx \frac{P(\tilde{\mathbf{z}}_t|u)}{\sum_{v \in \mathcal{U}} P(\tilde{\mathbf{z}}_t|v)}. \tag{5.16}$$

5. Update the current parameter estimate using a gradient-ascent scheme using first-order finite forward differences:

$$\forall u \in \mathcal{U}, \hat{\mathbf{z}}_u^{m+1} = \hat{\mathbf{z}}_u^m + \left(\frac{\partial \hat{\mathbf{z}}_u}{\partial \hat{\mathbf{z}'}_u^m}\right)\left(\frac{\sum_{t \in \mathcal{T}} P(u|\tilde{\mathbf{z}}_t)\tilde{\mathbf{z}}_t}{\sum_{t \in \mathcal{T}} P(u|\tilde{\mathbf{z}}_t)} - \hat{\mathbf{z}'}_u^m\right), \tag{5.17}$$

   where the Jacobian is a diagonal matrix—each component of the vector neighborhood $\hat{\mathbf{z}}_u$ is corrupted independently because of the conditional independence assumption on the noise model—that can be computed numerically using the lookup table $\mathcal{L}_\mu(\cdot)$. The partial derivatives in the Jacobian are the reciprocal of the rate of change of the shift in the mean of the Rician-corrupted Gaussians with respect to the change in the means of the input Gaussian (for i.i.d. additive Gaussian noise the Jacobian is exactly identity). We have numerically found that this derivative is always greater than unity, and approaches unity for large SNR (where Rician noise behaves very similar to i.i.d. additive Gaussian noise). For low SNR, however, the derivative can be much larger than unity and this may lead to numerically-large updates. In practice, we treat the Jacobian as identity. This results in a projected-gradient ascent strategy that is still guaranteed to converge.
6. If $\sum_{u \in \mathcal{U}} \parallel \hat{\mathbf{z}}_u^{m+1} - \hat{\mathbf{z}}_u^m \parallel_2^2 < \epsilon$, where $\epsilon$ is a small threshold, then stop, otherwise go to Step 3.

### 5.3.4 Engineering Enhancements for the EM Algorithm

Our initialization strategy gives $|\mathcal{U}| = \alpha|\mathcal{T}|$, where $\alpha$ is a free parameter and $0 < \alpha \leq 1$. Too small an $\alpha$ reduces the ability of the nonparametric PDF to well approximate the uncorrupted-signal Markov PDF. Too large an $\alpha$ increases the number of parameters to be estimated—equal to $|\mathcal{U}|$—thereby increasing the chance of the EM algorithm getting

stuck on local maxima. A large $\alpha$ also increases the space requirements of the algorithm: $O(|\mathcal{U}||\mathcal{T}|)$. We have found that, in practice, the algorithm is not very sensitive to the specific choice of $\alpha$ and a choice of $\alpha = 0.33$ works well in practice.

To further reduce the computational and space requirements of the algorithm, we can replace the set $\mathcal{T}$ itself by a uniformly-distributed random sample of observations $\mathcal{T}^\dagger$, with $|\mathcal{T}^\dagger| = \beta|\mathcal{T}|, 0 < \beta \leq 1$, and subsequently choose $U$ as a random sample from $\mathcal{T}^\dagger$, with $|\mathcal{U}| = \alpha|\mathcal{T}^\dagger|$. This makes the computational and space complexity of the EM algorithm both to be $O(\alpha\beta^2|\mathcal{T}|^2)$. The results in this paper use $\alpha = 0.33$ and $\beta = 0.66$.

## 5.4 Iterated Conditional Entropy Reduction (ICER)

At each pixel $t$, the prior PDF is

$$P(x_t|\mathbf{y}_t) = \frac{\sum_{u\in\mathcal{U}} G(\mathbf{y}_t - \mathbf{y}_u, \sigma)G(x_t - x_u, \sigma)}{\sum_{u\in\mathcal{U}} G(\mathbf{y}_t - \mathbf{y}_u, \sigma)} \tag{5.18}$$

and the likelihood PDF is

$$P(\tilde{x}_t|x_t) = \frac{1}{\eta(\tilde{x}_t, \sigma_R)} \frac{\tilde{x}_t}{\sigma_R^2} \exp\left( - \frac{\tilde{x}_t^2 + x_t^2}{2\sigma_R^2}\right) I_0\left(\frac{\tilde{x}_t x_t}{\sigma_R^2}\right), \tag{5.19}$$

where $\eta(\tilde{x}_t, \sigma_R)$ is the normalization factor that depends on the observed value $\tilde{x}_t$ and the noise level $\sigma_R$. We propose updating pixel intensities $x_t$, to increase the posterior probability $P(x_t|\{x_u\}_{u\in\mathcal{T}\setminus\{t\}}, \tilde{\mathbf{x}})$ in ( 5.2), by performing a gradient ascent on the logarithm of the posterior. In [9, 6], we showed the equivalence between a gradient ascent on the logarithm of a PDF and entropy reduction using the Shannon's entropy measure. Entropy reduction on this posterior PDF results in the following update rule for all pixel intensities $x_t$

$$\begin{aligned} x_t &\leftarrow x_t - \frac{\partial h(x_t|\mathbf{y}_t, \tilde{x}_t)}{\partial x_t} \\ &= x_t + \left[\frac{\partial \log P(x_t|\mathbf{y}_t)}{\partial x_t} + \frac{\partial \log P(\tilde{x}_t|x_t)}{\partial x_t}\right] \\ &= x_t \\ &\quad + \frac{\sum_{u\in\mathcal{U}} G(\mathbf{y}_t - \mathbf{y}_u, \sigma)G(x_t - x_u, \sigma)(x_u - x_t)}{\sum_{u\in\mathcal{U}} G(\mathbf{y}_t - \mathbf{y}_u, \sigma)G(x_t - x_u, \sigma)} \end{aligned}$$

$$-\frac{\hat{x}_t^m}{\sigma^2} + \frac{\tilde{x}_t}{\sigma^2}\frac{I_1(\tilde{x}_t\hat{x}_t^m/\sigma^2)}{I_0(\tilde{x}_t\hat{x}_t^m/\sigma^2)}, \tag{5.20}$$

where $I_1(\cdot)$ is the first-order modified Bessel function of the first kind. ( The expression for the gradient of the logarithm of the Rician likelihood PDF appears in [11]. ) These sequence of updates leads to image estimates with nondecreasing posterior probabilities and, hence, guarantee convergence to a local maximum of the posterior PDF. We call this novel proposed algorithm for performing Bayesian estimation on MRFs as the *iterated conditional entropy reduction* (ICER).

## 5.5 MRI-Denoising Algorithm

The proposed iterative denoising algorithm requires an initial estimate. We obtain an initial estimate entirely based on the knowledge of the noise model, without any use of Markov prior. Thus, the initialization is a ML estimate of the image. The MRI-denoising algorithm finally produces the MAP image estimate as follows:

1. Infer the prior PDF $P(\mathbf{Z})$ (as described in Section 5.3) by minimizing the KL divergence, using the EM algorithm, between the observed corrupted-signal Markov PDF and its estimate derived from the prior-PDF model. The prior PDF is represented by a Parzen-window sum of isotropic Gaussian kernels with means $\{\mathbf{z}_u\}_{u\in\mathcal{U}}$ and standard deviation $\sigma$.
2. Obtain an initial denoised ML image $\hat{\mathbf{x}}^0 = \{\hat{x}_t^0\}_{t\in\mathcal{T}}$:

$$\forall t \in \mathcal{T}, \hat{x}_t^0 = \underset{x_t}{\text{argmax}}\ P(\tilde{x}_t|x_t). \tag{5.21}$$

   We compute the mode of each likelihood PDF numerically using the iterative mode-seeking *mean-shift procedure* [60, 57].
3. Given the denoised-image estimate $\hat{\mathbf{x}}^m$ at iteration $m$, obtain the next estimate $\hat{\mathbf{x}}^{m+1}$ as

$$\begin{aligned}\forall t \in \mathcal{T}, \hat{x}_t^{m+1} \;=\; & \hat{x}_t^m \\ & + \frac{\sum_{u\in\mathcal{U}} G(\hat{\mathbf{y}}_t^m - \mathbf{y}_u, \sigma)G(\hat{x}_t^m - x_u, \sigma)(x_u - \hat{x}_t^m)}{\sum_{u\in\mathcal{U}} G(\hat{\mathbf{y}}_t^m - \mathbf{y}_u, \sigma)G(\hat{x}_t^m - x_u, \sigma)} \\ & - \frac{\hat{x}_t^m}{\sigma^2} + \frac{\tilde{x}_t}{\sigma^2}\frac{I_1(\tilde{x}_t\hat{x}_t^m/\sigma^2)}{I_0(\tilde{x}_t\hat{x}_t^m/\sigma^2)},\end{aligned} \tag{5.22}$$

   where all the symbols have the same meaning as in Section 5.4.

4. If $\| \hat{\mathbf{x}}^{m+1} - \hat{\mathbf{x}}^{m} \|_2 < \epsilon$, where $\epsilon$ is small threshold, then stop, otherwise go to Step 3.

## 5.6 Results and Validation

This section gives validation results on synthetic brain MR images with a wide range of noise and bias values as well as real MR data. The computation for each iteration is $O(|\mathcal{A}_t||\mathcal{T}||\mathcal{N}_t|)$. We have found empirically that, with simulated MR images from the BrainWeb [31] database, ICER produces the largest reduction in RMS errors after a single iteration itself. Subsequent iterations converge at a nearby RMS-error value. All results in this paper employ a single iteration of ICER.

### 5.6.1 Validation on Simulated and Real MR Images

Figure 5.2 presents the results of denoising a particular slice from volumetric T1-weighted simulated BrainWeb data. The proposed MRI-denoising algorithm acts conservatively, reducing the RMS error by about $40\%$. Figure 5.2(d) shows the difference between the corrupted and the uncorrupted images. The shift in the intensity PDF introduced by Rician noise is evident in the lighter background region (higher intensity on the average) corresponding to low signal intensities. The intensities in this difference image also possess a very low degree of spatial correlation. Figure 5.2(e) shows the difference between the denoised and the uncorrupted images. We see that algorithm reduces the Rician-noise-introduced shift in intensities in the low-intensity background region—fewer bright spots. Empirical analysis shows that denoised image effectively corrects the for the shift in the corrupted-intensity PDF caused by Rician noise—as measured by the average value of the background intensities in the uncorrupted, corrupted, and denoised images. For the case of T1-weighted BrainWeb data with $5\%$ noise and $40\%$ bias in Figure 5.2 the average background values are: (a) $0.1$ for the uncorrupted image, (b) $3.1$ for the corrupted image, and (c) $0.03$ for the denoised image. The difference images in Figure 5.2(e) show low magnitudes for errors in the background region. The difference image also possesses low correlation indicating that the proposed algorithm retained the significant image features more-or-less intact. The power spectrum of the difference image in Figure 5.2(f) shows the *whiteness* [81] of the residual.

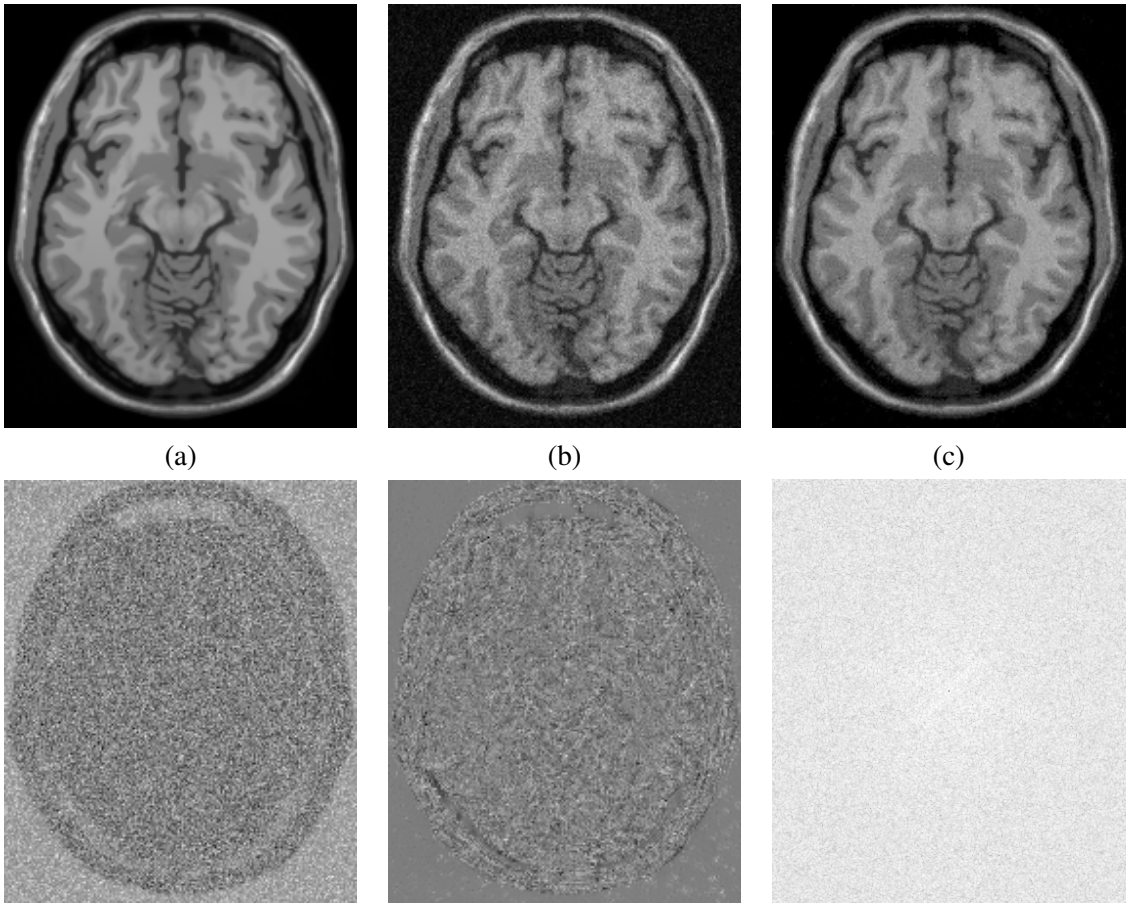
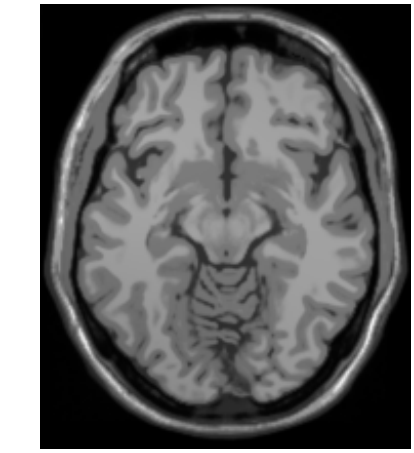
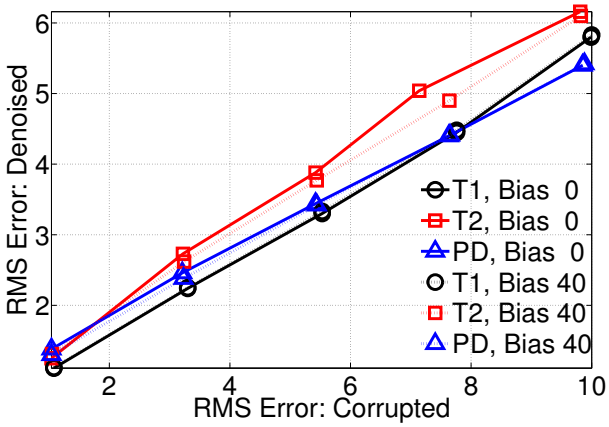


**Figure 5.2**. Results with T1-weighted simulated BrainWeb data (intensity range $0:100$) with the Rician noise level $\sigma_R = 5$ and a $40\%$ bias field. (a) Uncorrupted image. (b) Rician-noise corrupted image: RMSE = $5.53$. (c) Denoised image: RMSE = $3.3$. (d) Difference between the corrupted and uncorrupted images. (e) Difference between the denoised and uncorrupted images. (f) Power spectrum of the image in (e): close to *white*.

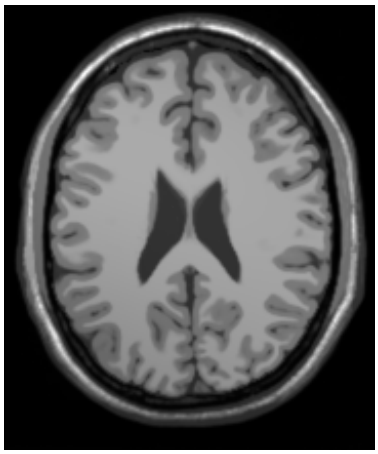
Figure 5.3 gives the performance of the proposed algorithm on three different slices of the BrainWeb MR data for varying noise and bias levels. We observe that the performance on biased and unbiased data is equivalent. This stems from the ability of adaptive-MRF model to effectively infer the appropriate Markov statistics for each case and denoise based on the inferred model. We also observe that for very low Rician noise, i.e., $\sigma_R \approx 1$, the algorithm does not effectively reduce the RMS error. This may be because of a similar level of variability inherent in the data, and in the estimated uncorrupted-signal Markov PDFs, which makes the algorithm not clearly identify the



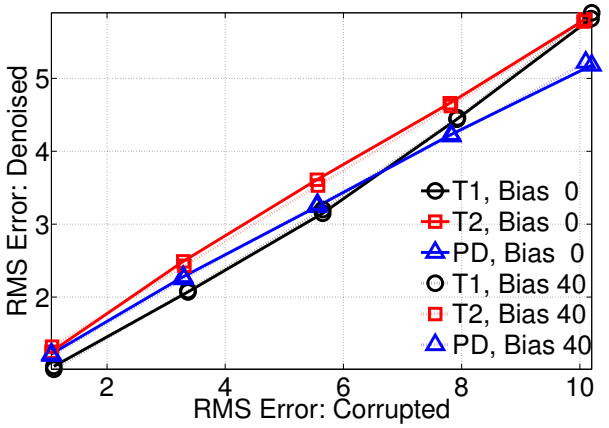
(a)



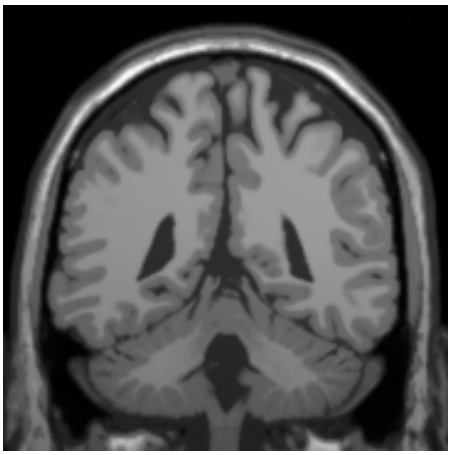

(d)



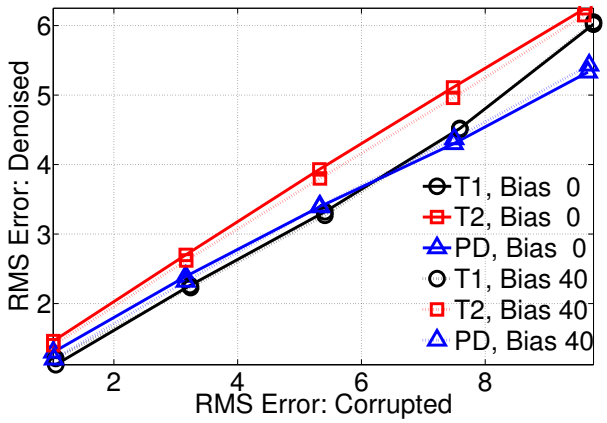
(b)




(e)



(c)



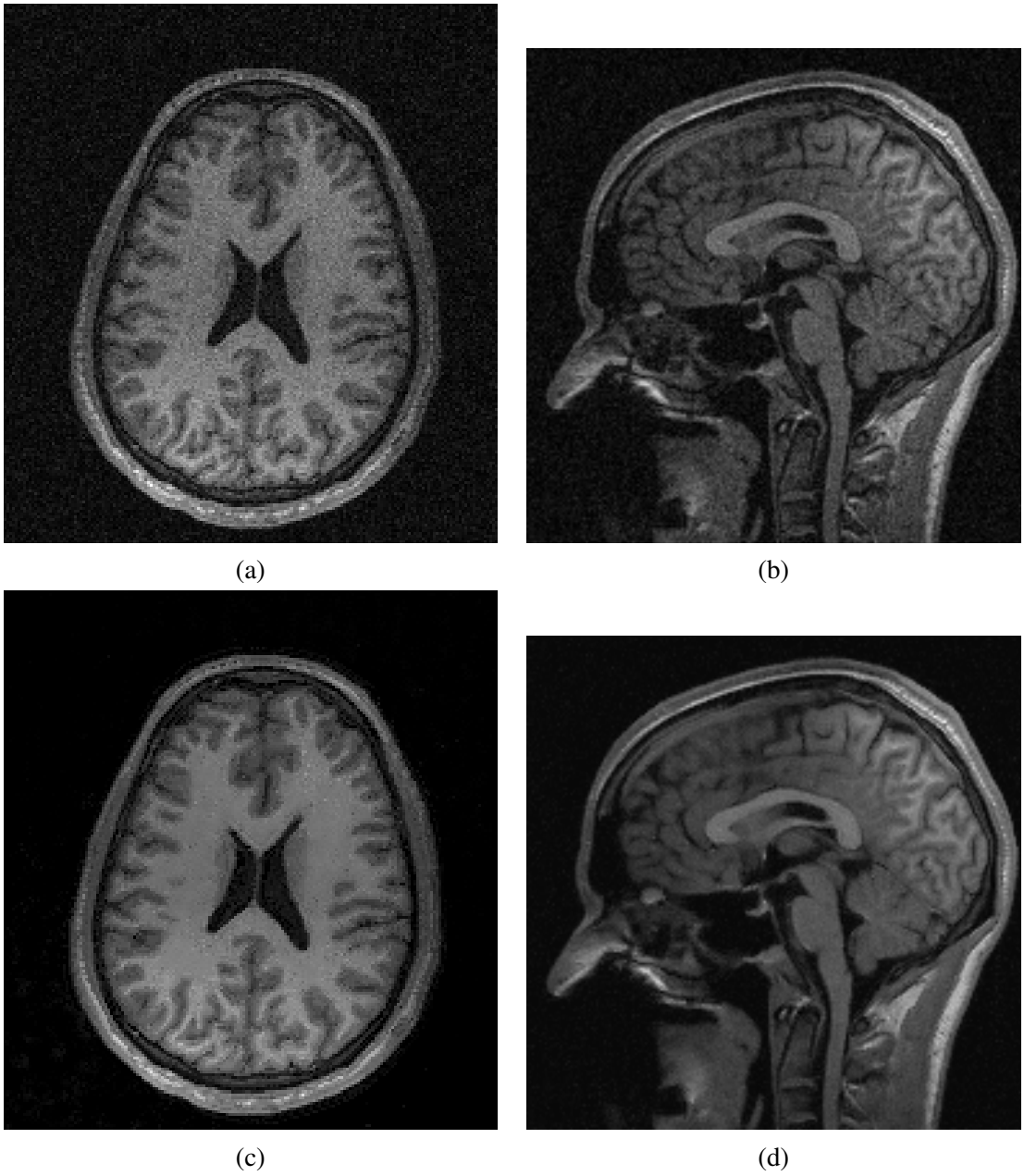

(f)

**Figure 5.3**. MRI-denoising results. (a)-(c) Three different brain slices from the BrainWeb dataset (only T1 modality shown; intensity range $0 : 100$). (d)-(f) Graphs indicating RMS errors for denoised and noisy images, with $0\%$ and $40\%$ bias fields, for T1, T2, and PD modalities on the three slices above.

noise. As the amount of noise increases, the proposed method can clearly differentiate the structure underlying the data from the noise. Figure 5.4 shows the performance of proposed algorithm on real data that depicts a significant inhomogeneity/bias.

Figure 5.5 compares, qualitatively and quantitatively, the performance of the proposed algorithm with several other recent and popular filtering algorithms. We have manually tuned all the free parameters in these other algorithms in order to give the best possible results. The proposed algorithm does better qualitatively, with an RMS error of $3.3$ (RMS error for noisy image is $5.53$) as compared to the RMS errors produced by other algorithms of around $4.0$ or more. Qualitatively too, the proposed algorithm gives a residual (difference between denoised and uncorrupted image) that is significantly less correlated. The state-of-the-art wavelet-based denoising algorithm [129] also seems to introduce artifacts in the denoised image.

Figure 5.6 show the qualitative and quantitative comparison of the proposed method with a state-of-the-art wavelet-based MRI-denoising algorithm [129]. We see that the proposed method produces lower RMS errors at all noise levels except with one image at the $9\%$ noise level. Although the RMS error for the proposed method is a little more for this high-noise case, Figure 5.6(c) and Figure 5.6(d) show that the residual for the wavelet-based method is significantly more correlated. This residual also indicates the presence of artifacts in the wavelet-denoised image.



(a) (b)

(c) (d)

**Figure 5.4**. MRI-denoising results. (a),(b) Noisy slices from a real MR volume. (c),(d) Denoised images.

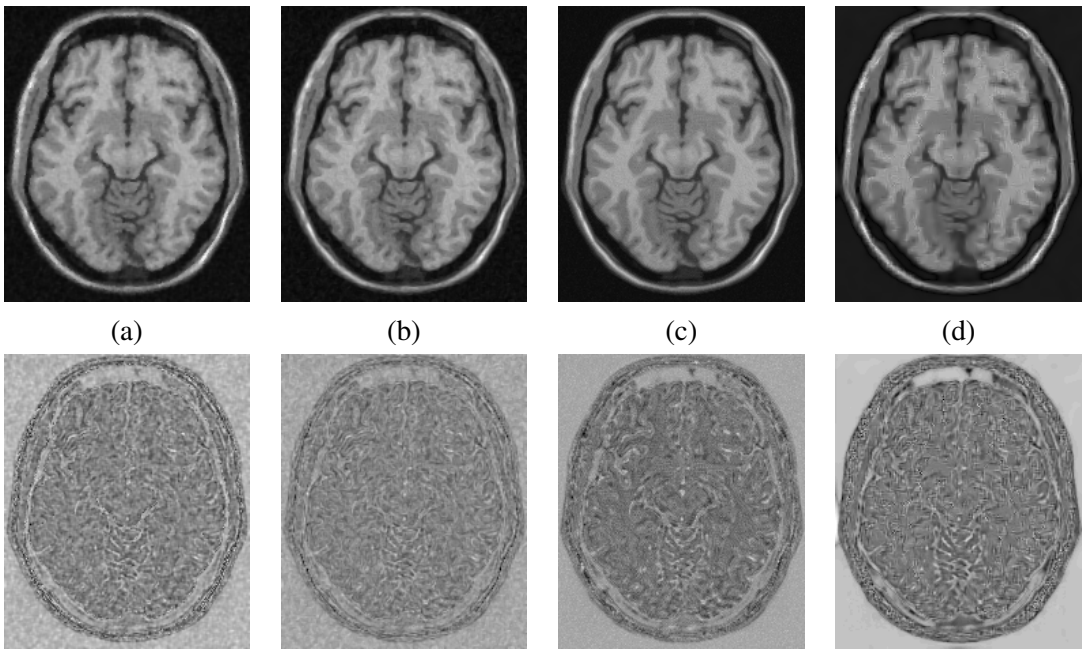



**Figure 5.5**. Results with T1-weighted simulated BrainWeb data (intensity range $0:100$) with the Rician noise level $\sigma_R = 5$ and a $40\%$ bias field. The noisy image in Figure 5.2(b) (RMSE = $5.53$) denoised using (a) anisotropic diffusion [127]: RMS error $4.03$, (b) curvature flow [153]: RMS error $3.93$, (c) UINTA [9]: RMS error $4.0$, and (d) the state-of-the-art wavelet-based MRI denoiser [129]: RMS error $5.64$, (e)-(h) show the differences between the denoised images in (a)-(d) and the uncorrupted image in Figure 5.2(a).

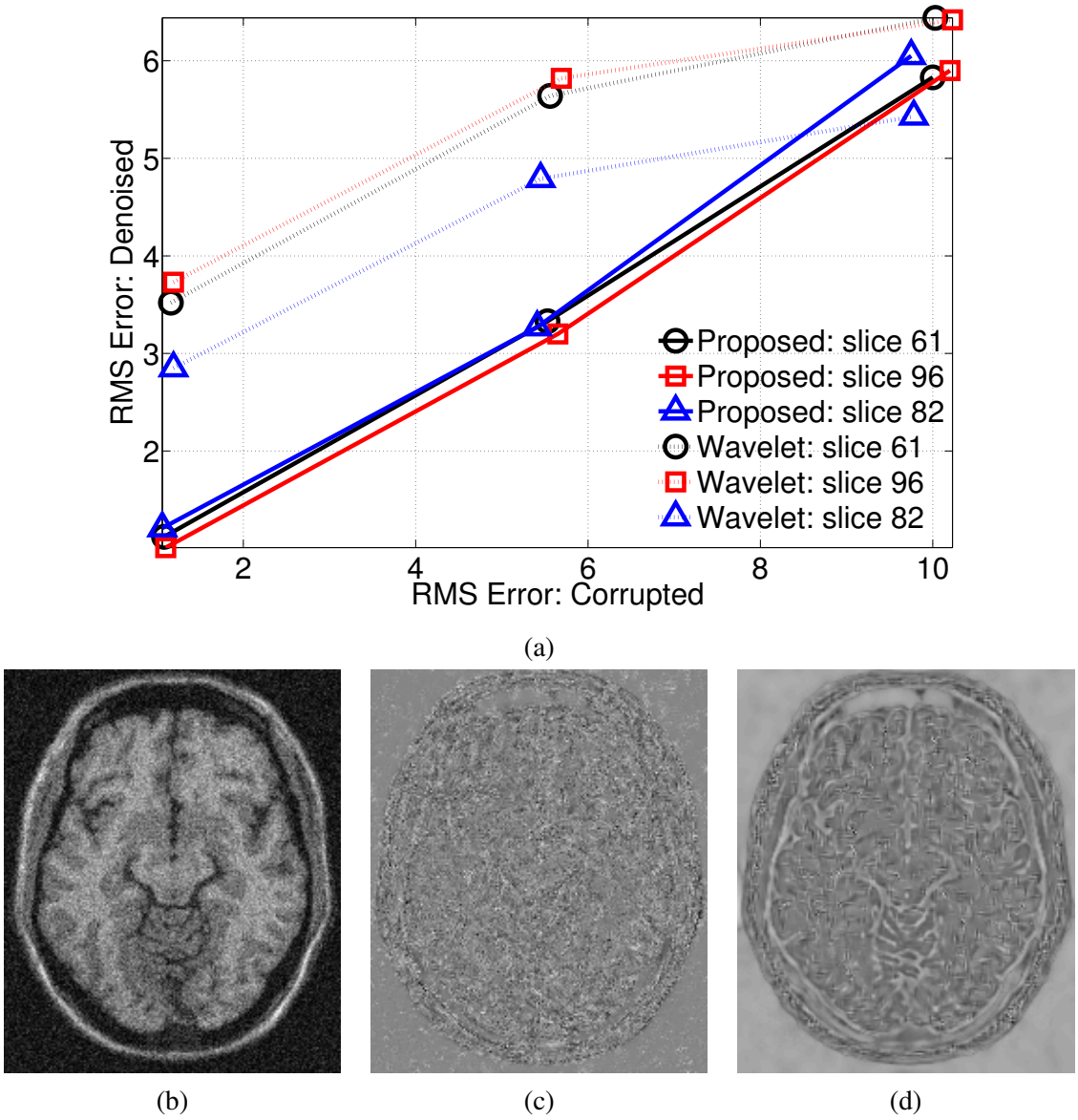



**Figure 5.6**. Comparison of proposed MRI-denoising method with the state of the art. (a) Quantitative comparison of the proposed method with a state-of-the-art wavelet-based MRI-denoiser [129] for the three different slices of T1 BrainWeb data, shown in Figure 5.3 (intensity range $0 : 100$), with varying noise levels and a $40\%$ bias field. (b) Corrupted T1 data with $9\%$ noise and $40\%$ bias field. (c) and (d) show the difference between the denoised and uncorrupted images for the proposed and wavelet-based [129] methods, respectively, when these methods are applied to the corrupted data in (b).

# CHAPTER 6

# MRI BRAIN TISSUE CLASSIFICATION BY MAXIMIZING MUTUAL INFORMATION

Tissue classification in MR images of human brains is an important problem in medical image analysis. The fundamental task in tissue classification is to classify the voxels in the volumetric (3D) MR data into gray matter, white matter, and cerebrospinal fluid tissue types. This has numerous applications related to diagnosis, surgical planning, image-guided interventions, monitoring therapy, and clinical drug trials. Such applications include the study of neuro-degenerative disorders such as Alzheimer's disease, generation of patient-specific conductivity maps for EEG source localization, determination of cortical thickness and substructure volumes in Schizophrenia, and partial-volume correction for low-resolution image modalities such as positron emission tomography.

Manual segmentation or classification of high-resolution 3D images is a tedious task, which is impractical for large amounts of data. Because of the complexity of this task, such classifications can be very error prone and exhibit nontrivial inter-expert and intra-expert variability [30]. *Fully automatic* or *unsupervised* methods, on the other and, virtually eliminate the need for manual interaction, and thus such methods for brain tissue classification have received significant attention in the literature.

Current state-of-the-art methods for automatic brain tissue classification typically incorporate the following strategies: (a) parametric statistical modeling, e.g., Gaussian, of voxel grayscale intensity for each tissue class, (b) Markov-random-field (MRF) modeling to enforce spatial smoothness on the classification, (c) methods to explicitly correct for the inhomogeneities inherent in MR images, and (d) probabilistic-brain-atlas information in the classification method. Several factors, however, continue to pose significant challenges to the state of the art:

- The intensities and contrast in MR images varies significantly with the pulse sequence, and several other scanner parameters. The quality of MR data also shows a certain amount of variation when produced at multiple sites with different MR scanners.
- MRI-acquisition artifacts, which include the Rician nature of the noise in magnitude-MR data [115] and partial voluming effects [94], can cause the data to significantly deviate from the Gaussian tissue-intensity models, thereby compromising the quality of the classification.
- Many methods treat the inhomogeneity as multiplicative noise (bias field) and explicitly correct the MR intensities to reduce its effect. For certain kinds of coil configurations or applications, such as neonatal brain MRI, however, inhomogeneities do not adhere to standard multiplicative models [134].

To address these issues in an effective way, we propose an unsupervised classification approach that *adapts* to the data. One adaptation strategy is to automatically learn the underlying image statistics from the data and construct a classification strategy based on that model. This chapter presents a novel method [163, 5] for MRI brain tissue classification that incorporates an adaptive nonparametric model of neighborhood/Markov statistics. The method incorporates the information content in the neighborhoods in the classification process. Together with a weak smoothness constraint on the estimated Markov statistics, it virtually eliminates the need for explicit smoothness constraints on the class-label image. The method produces an optimal classification by iteratively maximizing a mutual-information metric that relies on Markov PDFs. The algorithm adjusts all its important internal parameters automatically using a data-driven approach and information-theoretic metrics. Combined with an atlas-based initialization, it is fully unsupervised. It incorporates *a priori* information in probabilistic-brain-atlases via a Bayesian formulation. Experiments on real, simulated, and multimodal data demonstrate the significant advantages of the method over the current state-of-the-art. The method also performs reasonably well without any explicit inhomogeneity correction.

## 6.1 Overview of MRI Brain Tissue Classification

This section discusses works in MRI brain tissue classification and nonparametric Markov modeling along with their relationships to the proposed method. It compares and contrasts the proposed strategy, in brief, with the key ideas around which various classification strategies have evolved, including (a) partitioning based on grayscale voxel-intensity data, (b) regularization schemes based on local interactions among class labels, and (c) spatial priors based on probabilistic and anatomical atlases.

Wells *et al.* [176] present a method that couples tissue classification with inhomogeneity correction based on ML parameter estimation. They use the EM algorithm of Dempster *et al.* [43] to simultaneously estimate the unknown bias field and the classification. Leemput *et al.* [93, 94] extend this approach by posing the problem in the context of mixture density estimation to estimate the grayscale intensity PDFs for each tissue type. They apply the EM algorithm to estimate these PDFs as well as the bias and, in turn, the classification. Their approach assumes that each tissue-intensity distribution conforms to a parametric Gaussian PDF whose parameters are obtained via the EM algorithm. The proposed method, in contrast to typical EM-based strategies, does not impose any parametric model on the tissue intensities. Instead, it automatically adapts to the data using neighborhood sampling and nonparametric density estimation.

The EM-classification algorithm [176] does not impose any smoothness constraint on the classification and it is therefore susceptible to outliers in the tissue intensities. Some approaches for tissue classification do not explicitly account for noise, but employ image-denoising methods as a preprocessing step [64, 100]. Many subsequent works incorporate noise models into the classification without such preprocessing. Several authors [86, 76, 93, 94, 121, 187] have extended the EM-classification algorithm to incorporate spatial smoothness via Gibbs/Markov priors on the label image. For instance, Kapur *et al.* [86] use spatially-stationary Gibbs priors to model local interactions between neighboring labels. Typically, these methods modify single-voxel tissue-probabilities based on energies defined on local configurations of classification labels. They assign lower energies to spatially-smooth segmentations, making them more likely. Such strong Markov models, however, can over regularize the fine-structured interfaces, e.g., the one

between gray matter and white matter. Hence, it is often necessary to impose additional heuristic constraints [76, 93, 94]. Ruf *et al.* [146] extend the EM approach to perform spatial regularization by incorporating the spatial coordinates of the voxels, in addition to their grayscale intensities, in the feature vector.

This tissue-classification work dovetails with the mainstream image-processing literature, which presents a variety of algorithms that rely on MRF models of images [61, 16, 120, 99, 161]. Such methods typically involve iterative stochastic-relaxation schemes that compute local image updates based on random sampling from local conditional PDFs. These conditional PDFs on neighborhood configurations define an energy that is progressively reduced. Typically, the methods specify the conditional PDFs in parametric forms, e.g., Gaussian [99]. In this way, they encode a set of probabilistic assumptions (priors) about the geometric/statistical properties of the image data, and thus they are effective only when the data conform sufficiently well to the prior. Furthermore, the previous work on MRI classification models each tissue class with Gaussian-mixture models, which is homogeneous across the image. The proposed method, rather than enforcing a particular Markov prior on the data, *learns* the relevant Markov statistics nonparametrically from the input data and bases the classification on this adaptive model.

Researchers have also used active contour models [38, 36] to impose smoothness constraints for segmentation. These methods typically attempt to minimize the area of the segmentation boundary (smoothness) simultaneously with proper fidelity to the data. These models produce results that can be quite sensitive to the contour parameters that control the influence of the data and the smoothness. Hence, these methods typically require careful manual parameter-tuning. The proposed method, on the other hand, sets its important internal free parameters via data-driven techniques using information-theoretic optimality criteria. As a result, it easily applies to a wide spectrum of data with little parameter tuning.

An important component in MRI brain tissue classification is the correction of intensity inhomogeneities or bias fields. Several approaches propose an approach that couples iterative updates of the class labels with the bias-field correction based on polynomial least-squares fitting [176, 69, 93]. Although the focus of this chapter is not on

inhomogeneity correction, it is compatible with all such schemes. The literature also presents many methods that aim at *implicitly* dealing with the inhomogeneities in MR data in the classification method itself [184, 183, 92, 136, 113]. For instance, Yan and Karp [184] employ an adaptive K-means clustering strategy that, over many iterations, gradually takes the feature-space points from increasingly-local neighborhoods. The initial segmentation uses all points in the image but the final segmentation implicitly accounts for local intensity variations such as those cause by the inhomogeneity field.

More recently, researchers have realized the importance of the *nonstationarity* of head images in tissue classification—different anatomical structures in the brain represent different image patterns, each possessing unique higher-order/Markov statistics—and several authors introduce global information in the form of anatomical atlases [165, 37, 142]. Typically, they use atlases in one of two ways. First is to convert the classification problem into a deformable-registration problem between the MR-image and the anatomical brain atlas. Once the registration is done, the method uses the resulting transformation to map the anatomical structure from the atlas onto the data to produce a segmentation based on the labels in the atlas. Several authors use *probabilistic atlases*, which are generated from ensembles of head images. These atlases encode tissue probabilities (rather than discrete label values) at each voxel, and are used as a prior in the EM estimation described previously [35]. The proposed method uses probabilistic atlases for the initialization, which is important to the success of the algorithm, and can include probabilities from atlases in the posterior estimation.

## 6.2 Learning Per-Class Markov Statistics Nonparametrically

The proposed method constructs a segmentation strategy based on a Markov statistical image model [99] that it *learns* automatically from the input data. It formulates the segmentation problem as an optimization problem to maximize the dependency or *mutual information* [34] between the segmentation labels and the Markov image statistics.

The proposed approach models brain MRI images as derived from piecewise stationary-ergodic MRFs. For brain MR images, the Markov PDFs at voxels in individual parts of the brain, such as white matter or gray matter, are similar and, hence, the piecewise-

stationary model holds to some degree. Indeed, the successful high-quality classifications produced by the proposed method corroborate this claim.

Consider a discrete RV $L : T \rightarrow \mathbb{Z}$, where $\mathbb{Z}$ is the set of integers, that maps each voxel $t \in \mathcal{T}$ to the class it belongs to, i.e., $L(t) = k$ if voxel $t$ is in class $k$. Let $\{\mathcal{T}_k\}_{k=1}^{K}$ denote a mutually-exclusive and collectively-exhaustive decomposition of the image domain $\mathcal{T}$ into $K$ regions—assumed stationary—such that $\mathcal{T}_k = \{t \in \mathcal{T} : L(t) = k\}$. The stationarity assumption implies that for each class $k$ the Markov PDFs are exactly the same, i.e.,

$$\forall k = 1, 2, \ldots, K, \forall t \in \mathcal{T}, P(\mathbf{Z}_t | L(t) = k) = P_k(\mathbf{Z}). \tag{6.1}$$

Based on the piecewise stationary-ergodic assumption, the Parzen-window density estimate gives the PDF for class $k$ as

$$P_k(\mathbf{z}) \approx \frac{1}{|\mathcal{A}|} \sum_{s \in \mathcal{A}} G_d(\mathbf{z} - \mathbf{z}_s, \Psi_d), \tag{6.2}$$

where the set $\mathcal{A}$ is a small subset of $\mathcal{T}_k$ chosen at random for each voxel $t_k$.

## 6.3 Classification via Mutual-Information Maximization

This section formulates the classification problem as an optimal-segmentation problem using with an information-theoretic goodness measure associated with the Markov PDFs. It begins by forming a connection between information-theoretic measures, such as mutual information, entropy [34], and classification.

Loosely speaking, the mutual information between two random variables quantifies the degree of *functional dependence* between them. For functionally-dependent random variables, each variable uniquely determines the other, and the mutual information is maximized. On the other hand, independent random variables convey no information about each other, and their mutual information is zero (minimal). For image segmentation [87], we can say that a good segmentation is one in which the voxel-neighborhood intensity values provide the most information about the class labels. Likewise, knowing the voxel class should provide the most reliable estimate of the voxel neighborhood. Clearly, there is no strict functional dependence and images are inherently stochastic,

but mutual information provides a well-founded mechanism for quantifying the degree to which these properties hold.

Using the set of conditional PDFs $\{P_k(\mathbf{Z})\}_{k=1}^{K}$ for the $K$ classes, we can define a joint PDF $P(L, \mathbf{Z})$ between the RVs $L$ and $\mathbf{Z}$. At each voxel $t$, an instance $(l_t, \mathbf{z}_t)$ is drawn from the joint PDF. What we observe, however, are only the intensity-neighborhood vectors $\mathbf{z}_t$. The label values $l_t$ define the classification and must be estimated. We define the optimal segmentation as the one that maximizes the mutual information between $L$ and $Z$, i.e.,

$$\begin{aligned} I(L, \mathbf{Z}) &= h(\mathbf{Z}) - h(\mathbf{Z}|L) \\ &= h(\mathbf{Z}) - \sum_{k=1}^{K} P(L = k) h(\mathbf{Z}|L = k), \end{aligned} \tag{6.3}$$

where $I(\cdot)$ is the mutual information function and $h(\cdot)$ is the entropy. Entropy is a measure of randomness or uncertainty associated with a PDF [34], and regions $\mathcal{T}_k$ having low entropies $h(\mathbf{Z}|L = k)$ for Markov PDFs exhibit a high degree of predictability in their neighborhoods.

The entropy of class $k$ is

$$h(\mathbf{Z}|L = k) = -\int_{\Re^{|d|}} P_k(\mathbf{z}) \log P_k(\mathbf{z}) d\mathbf{z}, \tag{6.4}$$

where $d = |\mathcal{N}_t|$ is the neighborhood size.

The entropy of the Markov PDF associated with the entire image, $h(\mathbf{Z})$, is independent of the label assignment $L$ and we can ignore it during the optimization. Thus, (6.3) implies that the optimal segmentation is the one that minimizes a weighted average of entropies $h(\mathbf{Z}|L = k)$ of the $K$ Markov PDFs associated with the $K$ stationary-ergodic regions. The present mutual-information-based energy gives more importance, or weight, to reducing entropies of larger regions in the image in direct proportion to their size—the weights are the probability of occurrence of the classes $P(L = k)$ in the image. Rewriting $I(L, \mathbf{Z}) = h(L) - h(L|\mathbf{Z})$ provides more insight into this optimality metric. We see that the metric encourages segmentations with equal voxel counts for the classes—uniform PDF for $L$ implying maximal $h(L)$—while demanding high predictability of the label at each voxel $t$ given its neighborhood intensities $\mathbf{z}_t$—low $h(L|\mathbf{z}_t)$ leading to low $h(L|\mathbf{Z})$.

Equations (6.3) and (6.4) give the optimal segmentation as

$$\begin{aligned}\{\mathcal{T}_k^*\}_{k=1}^K &= \underset{\{\mathcal{T}_k\}_{k=1}^K}{\text{argmin}} \sum_{k=1}^K P(L=k)h(\mathbf{Z}|L=k) \\ &= \underset{\{\mathcal{T}_k\}_{k=1}^K}{\text{argmax}} \left( \sum_{k=1}^K P(L=k) \int_{\Re^d} P_k(\mathbf{z}) \log P_k(\mathbf{z}) d\mathbf{z} \right). \end{aligned} \tag{6.5}$$

Treating entropy as the expectation of negative log-probability and approximating the expectation, in turn, by the sample mean [34], gives

$$\begin{aligned}\{\mathcal{T}_k^*\}_{k=1}^K &= \underset{\{\mathcal{T}_k\}_{k=1}^K}{\text{argmax}} \left( \sum_{k=1}^K P(L=k) E_{P_k(\mathbf{Z})}\Big[ \log P_k(\mathbf{Z}) \Big] \right) \\ &\approx \underset{\{\mathcal{T}_k\}_{k=1}^K}{\text{argmax}} \left( \sum_{k=1}^K P(L=k) \frac{1}{|\mathcal{S}_k'|} \sum_{\mathbf{z} \in \mathcal{S}_k'} \log P_k(\mathbf{z}) \right), \end{aligned} \tag{6.6}$$

where $\mathcal{S}_k'$ is a *random sample* [47, 161] derived from the PDF $P_k(\mathbf{Z})$. Assuming ergodicity [47], in addition to stationarity, enables us to approximate ensemble averages using $\mathcal{S}_k$ with spatial averages using $\mathcal{T}_k$. Hence we have

$$\{\mathcal{T}_k^*\}_{k=1}^K \approx \underset{\{\mathcal{T}_k\}_{k=1}^K}{\text{argmax}} \left( \sum_{k=1}^K P(L=k) \frac{1}{|\mathcal{T}_k|} \sum_{t \in \mathcal{T}_k} \log P_k(\mathbf{z}_t) \right). \tag{6.7}$$

To estimate $P(L=k)$ from the data, we observe that the discrete random variable $L$ can take only $K$ possible values. Furthermore, $|\mathcal{T}_k|$ voxels, out of a total of $|\mathcal{T}|$ voxels, have $L=k$. Thus,

$$P(L=k) = \frac{|\mathcal{T}_k|}{|\mathcal{T}|}. \tag{6.8}$$

Substituting (6.8) in (6.7) gives

$$\{\mathcal{T}_k^*\}_{k=1}^K \approx \underset{\{\mathcal{T}_k\}_{k=1}^K}{\text{argmax}} \left( \frac{1}{|\mathcal{T}|} \sum_{k=1}^K \sum_{t \in \mathcal{T}_k} \log P_k(\mathbf{z}_t) \right). \tag{6.9}$$

The probabilities $P_k(\mathbf{z}_t)$ are given by the Parzen-window density estimate in (6.2), i.e.,

$$P_k(\mathbf{z}_t) \approx \frac{1}{|\mathcal{A}_t|} \sum_{u \in \mathcal{A}_t} G_d(\mathbf{z}_t - \mathbf{z}_u; \Psi_d), \tag{6.10}$$

where the set $\mathcal{A}_t$ is a small subset of $\mathcal{T}_k$. Section 3.5.1 describes how to construct $\mathcal{A}_t$, unique for each voxel $t$, to effectively estimate the probability.

So far, we have not taken into account any *a priori* information in the segmentation process and we have derived all probabilities solely from the data. The formulation, however, extends in a straightforward manner to include *a priori* information using standard Bayesian strategies followed by optimization involving the resulting posterior probabilities. Section 6.4.3 discusses how to integrate *a priori* information in the form of brain tissue probabilistic atlases into the proposed method. For the minimization in (6.9), we manipulate the regions $\mathcal{T}_k$ using a gradient-descent optimization strategy. Section 6.4.2 gives the details.

## 6.4 Brain Tissue Classification

For brain MR images, our goal is to segment the image into $K = 4$ regions corresponding to the (a) white matter, (b) gray matter, (c) cerebrospinal fluid, and (d) all other tissue types. This section starts by giving a high-level version of the proposed iterative classification algorithm along with an initialization strategy. It gives a few ways of incorporating *a priori* information in the probabilistic atlases into the proposed method. It describes the details of an efficient strategy for choosing the Parzen-window sample $\mathcal{A}_t$, explains why the method performs reasonably well without explicit inhomogeneity correction, and describes a optimal data-driven choice of important internal parameters.

### 6.4.1 Initial Classification Using Probabilistic Atlases

The proposed classification algorithm seeks local optima of mutual information from an initial assignment of class labels, $\{\mathcal{T}_k^0\}_{k=1}^K$. These labels must be sufficiently close to the solution to provide distinct density estimates for the different classes. For this, we use co-registered probabilistic atlases for the white matter, gray matter, and cerebrospinal fluid. We obtain these atlases from the ICBM repository [139], which also provides an average-T1 image registered with these atlases. These atlases give the *a priori* probability for a voxel belonging to one of these tissue types. The probabilities are obtained using an empirical procedure whose goal is to obtain an average-anatomy of the human brain. The procedure for constructing these atlases involved averaging 452 brain tissue-class images, after aligning all of them to a common coordinate system [139].

We define the initialization as the maximum-a-priori estimate. We first register the average-T1 image to the data using an affine transformation and then use the transformation to resample the three probability images. The initialization is therefore:

1. Perform affine registration between the average-T1 image, associated with the atlas, and the data.
2. Resample the white matter, gray matter, and cerebrospinal fluid atlases based on the transformation obtained in the previous step.

   Let $P_k^a(t), k = 1, 2, 3$ be the *a priori* probability, given by the atlas, for the $t$-th voxel belonging to the $k$-th tissue type.
3. Compute the probabilities for the class (say class $k = 4$) comprising all the non-brain tissue types:

$$\forall t \in \mathcal{T} : P_4^a(t) = 1 - \sum_{k=1}^{3} P_k^a(t). \tag{6.11}$$

4. Assign the initial class labels:

$$\forall t \in \mathcal{T} : L^0(t) = \underset{k}{\text{argmax}}\ P_k^a(t). \tag{6.12}$$

### 6.4.2 Classification Algorithm

From the Markov PDFs, which are estimated from the initial classification, we reassign voxels based on optimizing the information content of the labels. We observe that the energy in (6.9) can be reduced, based on a steepest-descent strategy, if each voxel $t$ is assigned to the class $k$ that maximizes the probability $P_k(\mathbf{z}_t)$. This is an iterative process where the Markov PDFs define a classification that, in turn, redefines the PDFs. Because the PDFs get implicitly redefined after every iteration, via the updated classification, the PDF estimates *lag*, so to speak, the classification. We have found this to be an acceptable approximation, although some recent work [17] introduces an additional term in the update rule to avoid this lag.

Given a classification $\{\mathcal{T}_k^m = \{t \in \mathcal{T} : L_t^m = k\}\}_{k=1}^K$ at iteration $m$, the algorithm iterates as follows:

1. For $k = 1, 2, 3, 4$ and $\forall t \in \mathcal{T}$, estimate $P_k^m(\mathbf{z}_t)$ nonparametrically, as described in Section 6.2.

2. Update the classification labels:

$$\forall t \in \mathcal{T} : L_t^{m+1} = \underset{k}{\text{argmax}} \; P_k^m(\mathbf{z}_t). \tag{6.13}$$

3. Stop upon convergence, i.e., when $\sum_{t \in \mathcal{T}} \delta(L^{m+1}(t) - L^m(t)) < \epsilon$, where $\delta(\cdot)$ is the Kronecker-delta (unit impulse) function and $\epsilon$ is a small threshold.

### 6.4.3 Bayesian Classification with Probabilistic-Atlas Priors

The registered, probabilistic atlas plays another role in the proposed classification algorithm. Instead of using data-driven probabilities alone for the classification updates, we can employ a Bayesian estimation strategy to compute the probabilities. The likelihood terms are the data-driven probabilities $P_k(\mathbf{z}_t)$ that we have computed via Parzen-window density estimation. The posterior is therefore the likelihood multiplied by the prior $P_k^a(t)$, which we derive from the probabilistic atlas. The Bayesian label updates are based on the MAP estimate:

$$\begin{aligned} \underset{k}{\text{argmax}} \; P(L = k|\mathbf{z}_t, t) &= \underset{k}{\text{argmax}} \; \Big( P(\mathbf{z}_t|L = k, t) P(L = k|t) \Big) \\ &= \underset{k}{\text{argmax}} \; \Big( P_k(\mathbf{z}_t) P_k^a(t) \Big). \end{aligned} \tag{6.14}$$

For the proposed method, our empirical evidence suggests that using the atlas directly as a prior can strongly dominate the likelihood and introduce systematic biases in the classification. Pohl *et al.* [130] report similar findings with a direct use of an atlas prior. For instance, for regions where the prior probability is zero, or near zero, the likelihood can have little effect. In such a case, the final segmentation may be very much like the initialization. Such behavior is likely an artifact from either (a) the limited variability in the atlas resulting from a limited-size population, or (b) the degree of misfit that remains after the registration process during atlas construction. In practice, the prior strictly interpreted from the atlas is too strong, and we have investigated two ways of weakening its affect on the final solution. Section 6.5.2 discusses empirical results and the effect of different priors on the proposed method in more detail. Section 6.5.2 shows the performance with both these priors.

One way of weakening the atlas prior is to use the atlas for discriminating only between two tissue types, namely the brain and nonbrain tissue. In this way, the prior does not interfere with the more subtle distinctions between the different brain tissues. For this, we sum the atlas probabilities for the white matter, gray matter, and cerebrospinal fluid to create one composite atlas that only gives the spatial probability for any kind of brain tissue. This is equivalent to redefining $P_k^a(t), \forall t \in \mathcal{T}$ as

$$\text{For } k = 1, 2, 3, \forall t \in \mathcal{T} : P_k^a(t) = 1 - P_4^a(t) \tag{6.15}$$

We call this the *2-class* prior.

Another way of reducing the strength of the prior is to voxel-wise rescale the atlas probabilities in such a way that the probabilities continue to add up to one but are less discriminating between the tissue types. We have used the following function for the desired effect.

$$\text{For } k = 1, 2, 3, 4, \forall t \in \mathcal{T} : P_k^a(t) = \frac{1 - v}{4} + v P_k^a(t), \tag{6.16}$$

where $v \in [0, 1]$ is a free parameter. The redefined prior probabilities continue to add up to unity: $\forall t \in \mathcal{T} : \sum_{k=1}^{4} P_k^a(t) = 1$. A value of $v = 1$ makes no change to the atlas probabilities, whereas $v = 0$ makes every class equiprobable. In this chapter we provide experimental results with a moderate value of $v = 0.5$. We call this the *scaled-atlas* prior.

### 6.4.4 Parzen-Window Kernel Parameter

The Parzen-window parameter $\sigma$, effectively controls the smoothing of the data in the feature space $< \mathbf{z} >$ of neighborhood-intensity vectors. However, $\sigma$ must be commensurate with the number and density of observations in that space, and thus it should adapt to different sampling strategies and applications. We have found that the optimal (cross-validated ML) $\sigma$, estimated from limited data, does not properly "connect" all of the configurations of gray matter neighborhoods in a single class, thereby breaking the manifold into many distinct pieces prone to misclassification. Indeed, this method of regularization is known to under-smooth the PDF and be sensitive to outliers. In practice,

to obtain desirable results with finite data, we impose additional smoothness on the Markov PDFs of each class, by multiplying the optimal $\sigma$ by a factor $\alpha$ larger than unity. This strategy is somewhat ad hoc and a different strategy based on plug-in bandwidth estimators [156, 171] that produces over-smooth, but more robust, PDF estimates might work better. We have found that the choice of the precise value of this *multiplicative factor* $\alpha$ is not critical and Table 6.1 in the next section confirms that the algorithm is quite robust to small changes in $\alpha$, i.e., $\alpha$ varying between $5$ and $10$. All of the results in this chapter employ $\alpha = 10$.

## 6.5 Results and Validation

This section gives validation results on real and synthetic brain MR images along with the analysis of the method's behavior. It also provides quantitative comparisons with a current state-of-the-art classification method [93, 94]. For all the results in this paper, we use a first-order neighborhood system for the MRF model. Thus, each pixel has $6$ neighbors—$2$ neighbors along each of the $3$ coordinate axes for the volumetric MR data. For all of the results in this chapter, we use $\sigma_{\text{spatial}} = 15$ voxels along

**Table 6.1**. The proposed method is fairly robust to changes in the values of the local-sampling Gaussian variance parameter and the Parzen-window $\sigma$ multiplicative factor. This table gives the Dice metrics for the BrainWeb T1 data with $5\%$ noise and a $40\%$ bias field.

| *Local-sampling* Gaussian variance | Gray matter | White matter |
|---|---|---|
| 100 | 0.9033 | 0.9386 |
| 225 | 0.9079 | 0.9427 |
| 400 | 0.9082 | 0.9422 |
| 625 | 0.9043 | 0.9368 |
| Parzen-window $\sigma$ *multiplicative factor* | Gray matter | White matter |
| 1.0 | 0.7634 | 0.9105 |
| 2.5 | 0.8988 | 0.9502 |
| 5.0 | 0.9106 | 0.9487 |
| 7.5 | 0.9095 | 0.9451 |
| 10.0 | 0.9079 | 0.9427 |
| 12.5 | 0.9066 | 0.9411 |
| 15.0 | 0.9058 | 0.9402 |

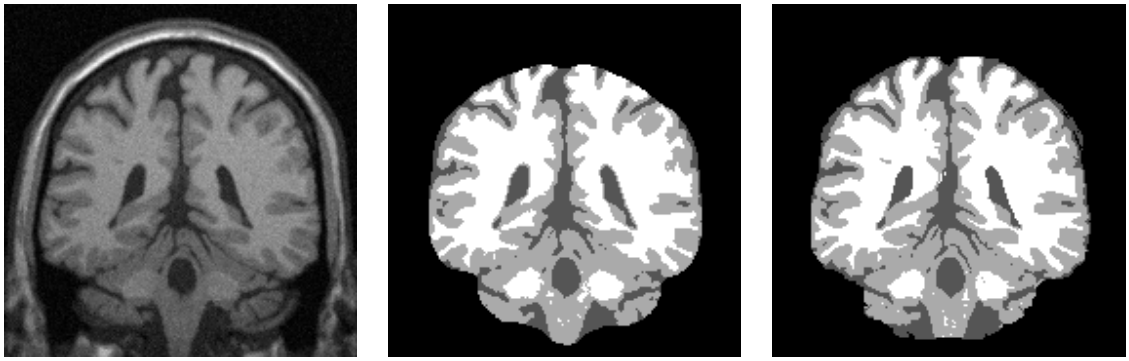
each cardinal direction. The empirical results in Table 6.1 confirm that the performance of the proposed method degrades gracefully for suboptimal values of this parameter. This local-sampling strategy also plays an important role in implicit inhomogeneity handling by enabling the method to subsume the bias field in the estimated Markov statistics that determine the segmentation. For all voxels $t$, the proposed method sets $|\mathcal{A}_t| = 500$, based on the method explained in Section 3.5.2. The computation for each iteration is $O(K|\mathcal{A}_t||\mathcal{T}||\mathcal{N}_t|)$. The algorithm typically takes about $4$ to $7$ iterations to converge depending on the noise/bias level. The implementation takes about 45 minutes to process a 181-voxels $\times$ 217-voxels $\times$ 181-voxels volume on a single Pentium-IV $2.8$GHz workstation. The implementation runs about twice as fast on a dual-processor shared-memory Pentium machine. The implementation in this chapter relies on the Insight Toolkit [2].

Leemput *et al.* [94] use the Dice metric [44] to evaluate the classification performance of their state-of-the-art approach, which is based on EM and Gibbs/Markov priors on the segmentation labels. For a direct comparison, we use the same metric. Let $\{\tilde{\mathcal{T}}_k\}_{k=1}^K$ denote the ground-truth classification and $\{\mathcal{T}_k^*\}_{k=1}^K$ denotes the classification obtained from the proposed method. Then, the Dice metric $D_k$ that quantifies the quality of the classification for class $k$ is $2|\mathcal{T}_k^* \cap \tilde{\mathcal{T}}_k|/(|\mathcal{T}_k^*| + |\tilde{\mathcal{T}}_k|)$, where the $|\cdot|$ operator gives the cardinality of sets.

### 6.5.1 Validation on Simulated MR Images

This section validates the proposed approach on simulated brain MR images with a known ground truth. We use 1 mm isotropic T1-weighted images from the BrainWeb simulator [31] with varying noise levels and bias fields. Figure 6.1 shows some data along with the classification and the ground truth.

We first show results on simulated T1-weighted data without any bias field and with noise levels varying from $0\%$ to $9\%$ . We use the *2-class* prior. The BrainWeb simulator defines the noise-level percentages with respect to the mean intensity of the brightest tissue class. Figures 6.2(a) and 6.2(b) plot the Dice metrics for gray-matter ($D_{\text{gray}}$) and white-matter ($D_{\text{white}}$) classifications for the proposed algorithm and compare them with



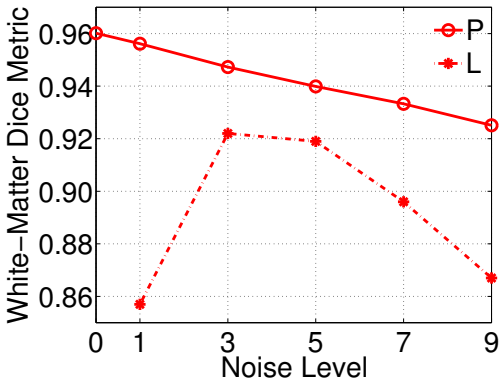
(a) (b) (c)

**Figure 6.1**. Qualitative analysis of the proposed algorithm with BrainWeb data [31] with $5\%$ noise and a $40\%$ bias field. (a) A coronal slice of the data. (b) The classification produced by the proposed method. (c) The ground truth.

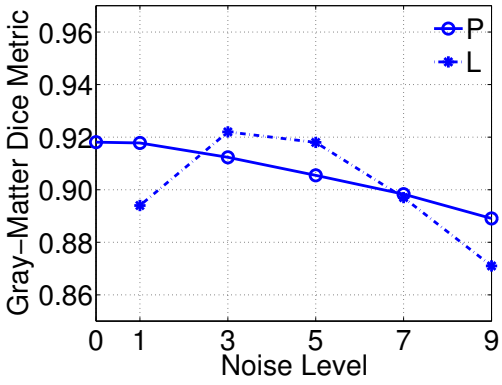
the corresponding values for the current state-of-the-art [94]. We see that the proposed method is consistently better for the white matter. For a few noise levels for the gray matter, its performance level is slightly below the state-of-the-art. We have found that this is caused by the *2-class* prior which biases the results against the gray matter, as compared to the *scaled-atlas* prior. With the *scaled-atlas* prior the results are consistently better than the state-of-the-art for all noise levels. Section 6.5.2 describes that both priors perform equally well as measured by the average of the Dice metric for the white matter and gray matter, i.e., $(D_{\text{white}} + D_{\text{gray}})/2$.

Figure 6.2(c) shows that for the average Dice metric, the proposed algorithm performs consistently better than the state-of-the-art at all noise levels for gray matter and white matter. Furthermore, it exhibits a slower performance degradation with increasing noise levels than the state-of-the-art method. For $3\%$ noise, which is typical for real MRI [94], the improvement in the average Dice metric is approximately $1.1\%$. The performance gain at $9\%$ noise is $3.8\%$. The larger gain over the state-of-the-art for large noise levels should prove useful for classifying noisier fast-acquisition clinical MRI.

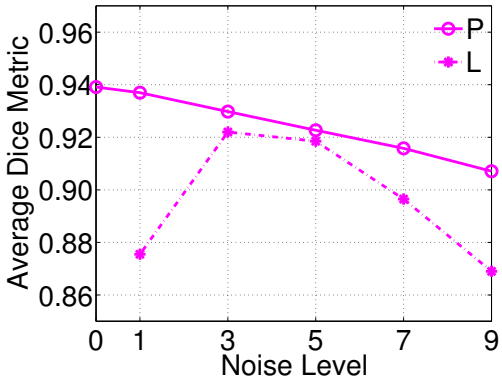
Figure 6.2 shows that for low noise levels, the performance of the parametric EM-based algorithm drops dramatically. This is because it systematically assigns voxels close to the interface between gray matter and white matter to the class which happens to have a larger intensity variability [94]. This class is, inherently, the gray matter class. It




(a)




(b)




(c)

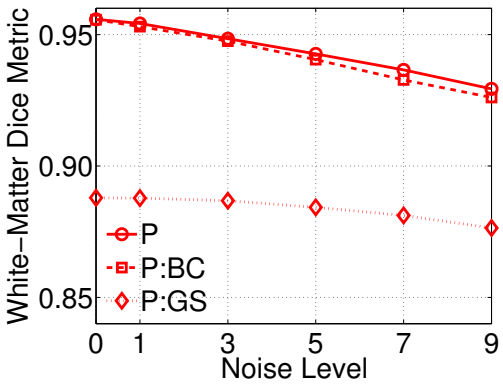
**Figure 6.2**. Validation, and comparison with the state-of-the-art [94], on simulated T1-weighted data without any bias and varying noise levels. Here, the proposed method uses the *2-class* prior. Dice metrics for (a) white matter: $D_{\text{white}}$, (b) gray matter: $D_{\text{gray}}$, and (c) their average: $(D_{\text{white}} + D_{\text{gray}})/2$. *Note*: In the graphs, P: Proposed method, L: State-of-the-art method of Leemput *et al.* [94].

turns out that, in such low-noise cases, partial voluming seems to dictate the MR-tissue intensity model which deviates significantly from the assumed Gaussian [94]. Hence, approaches enforcing Gaussian intensity PDFs on the classes, such as [94, 146], would face a serious challenge in this case. In contrast, the proposed adaptive modeling strategy, which is based on nonparametric density estimation, does not suffer from this drawback. Figure 6.2 clearly depicts this advantage of the proposed method.

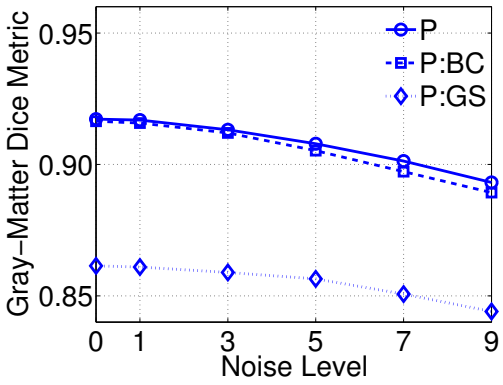
Strictly speaking, all methods trying to classify partial-volume voxels to one specific class are, in a way, fundamentally flawed. The proposed method, however, approaches this problem in a relatively more principled manner as compared to the EM-based method [94]. A partial-volume voxel $t$ comprising a larger contribution from tissue-class $k$ will produce a $\mathbf{z}_t$ lying "closer" to the feature-space distribution of class $k$. The results show that the data-driven nonparametric estimation of all tissue-class PDFs, employing the same Parzen-window $\sigma$ for each class, prevents any undesirable biases (unlike [94]) in the classification.

Figure 6.3 shows the validation results with the BrainWeb data having a $40\%$ bias field with varying noise levels. Even in the absence of an explicit bias-correction scheme, the method performs quite well on biased BrainWeb MR data (Figure 6.2). To confirm the important role that the *local-sampling* Parzen-window density estimation strategy plays in enabling the automatic learning of the bias field, we perform two more experiments. In the first experiment, we use explicit bias correction with the proposed method (degree-4 polynomial fit [93] to the white matter intensities iteratively). Figure 6.3 shows that this method performs approximately as well, but not significantly better than without the bias correction. The second experiment replaced the *local-sampling* scheme with a *global-sampling* scheme that chooses the random Parzen-window sample (with the same sample size $|\mathcal{A}_t|$) uniformly over the image. Figure 6.3 shows that this scheme performs significantly worse at all noise levels in the absence of bias correction.

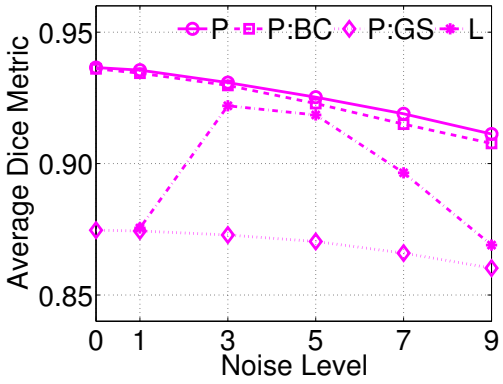
To study the sensitivity of the variance parameter $\sigma^2_{\text{spatial}}$ for the local-sampling Parzen-window Gaussian and the Parzen-window $\sigma$ multiplicative factor $\alpha$, we measure the Dice metrics for the white matter and gray matter over a range of parameter values. We use the BrainWeb T1 data with $5\%$ noise and a $40\%$ bias field. Table 6.1 gives the results




(a)




(b)




(c)

**Figure 6.3**. Validation, and comparison with the state-of-the-art [94], on simulated T1-weighted data with $40\%$ bias and varying noise levels. We compare the performance by incorporating explicit bias correction and *global sampling: same sample size* (see text). Dice metrics for (a) white matter: $D_{\text{white}}$, (b) gray matter: $D_{\text{gray}}$, and (c) their average: $(D_{\text{white}} + D_{\text{gray}})/2$. *Note*: In the graphs, P: Proposed method, BC: Bias correction, GS: *Global sampling: same sample size*, L: State-of-the-art method of Leemput *et al.* [94].

confirming that the classification performance is fairly robust to changes in the values of these two parameters, as explained before in Section 3.5.2.

We can extend the proposed method in a straightforward manner to deal with multimodal data. Multimodal segmentation entails classification using MR images of multiple modalities, e.g., T1 and PD. It treats the combination of images as an image of vectors with the associated PDFs in the *combined* probability space. Figure 6.4 shows the classification results for multimodal data using T1 and PD images, both with and without a bias field. The results demonstrate that incorporating more information in the classification framework, via images of two modalities T1 and PD, produces consistently better results than those using T1 images alone.

### 6.5.2 Validation on Real MR Images

This section shows validation results with real expert-classified MR images. We obtained this data set from the IBSR website [1]. The data set comprises T1-weighted brain MR images for 18 subjects. Figure 6.5 shows an example from the data set. We observe that the data have lower contrast and possesses certain acquisition-related artifacts that makes the classification task more challenging than that for the BrainWeb dataset. Figure 6.5 also shows an example of a classification generated by the proposed method and compares it to the ground truth.

Figure 6.6 compares the performance of the proposed method using the two different atlas-based priors. Figure 6.6(a) shows that the *2-class* prior, relative to the *scaled-atlas* prior, biases the classification more in favor of the white matter. With the *2-class* prior, which gives equal weight to all three brain tissue types, the Dice metric for the white matter is better than that for the gray matter because of lower inherent variability of the intensities in the white matter. The *scaled-atlas* prior imposes a stronger constraint which tends to shift this bias, as seen in Figure 6.6(b). Empirical evidence confirms that as the parameter $v$ varies from $0.0$ to $1.0$, the bias shifts away from white matter towards gray matter. Nevertheless, with the average Dice metric, Figure 6.6(c) shows that both priors perform equally well.

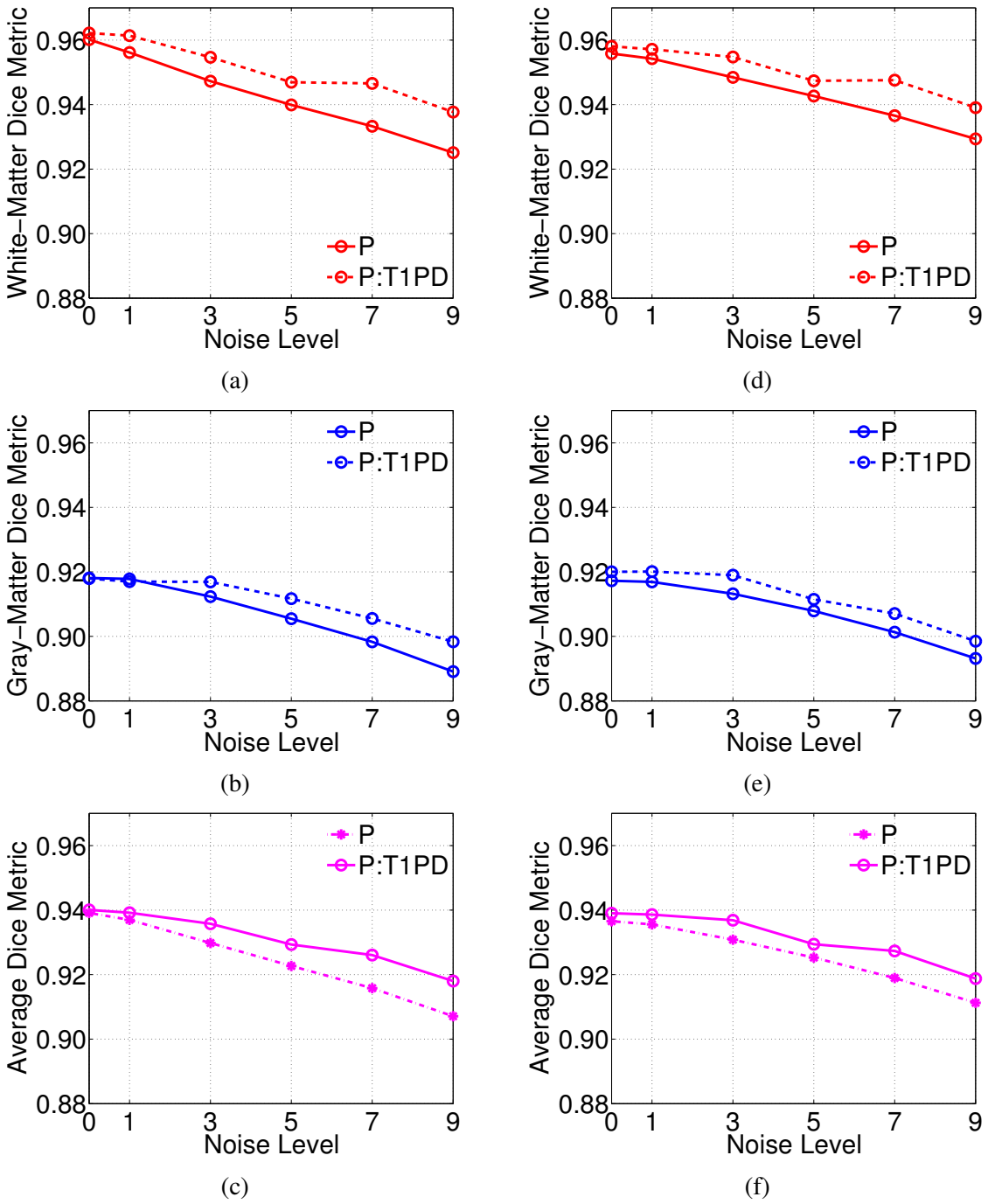



**Figure 6.4**. Validation on simulated multimodal (T1 and PD) data with varying noise levels. Dice metrics for (a) white matter: $0\%$ bias, (b) gray matter: $0\%$ bias, and (c) their average: $0\%$ bias. Dice metrics for (d) white matter: $40\%$ bias, (e) gray matter: $40\%$ bias, and (f) their average: $40\%$ bias. *Note*: In the graphs, P: Proposed method, T1PD: Using both T1 and PD images.

(a) (b) (c)

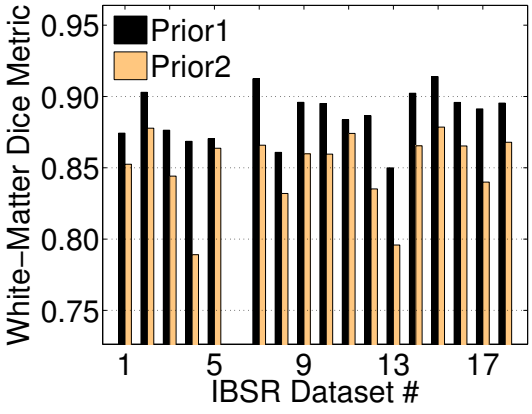
**Figure 6.5**. Qualitative analysis of the proposed algorithm with IBSR data [1]. The voxel size for this image is $0.9375 \times 0.9375 \times 1$ (coronal). (a) An axial slice of the data. (b) The classification produced by the proposed method. (c) The expert-classified ground truth.

For the proposed algorithm using the *2-class* prior, Table 6.2 gives the mean, median, and the standard deviation for the Dice metrics over the entire dataset. The proposed method yields a higher mean (by a couple of percent) and lower standard deviation for the Dice metrics over both white matter and gray matter classes, as compared to the results reported by Ruf *et al.* [146] for the state-of-the-art method of Leemput *et al.* [94] as well as their own method.

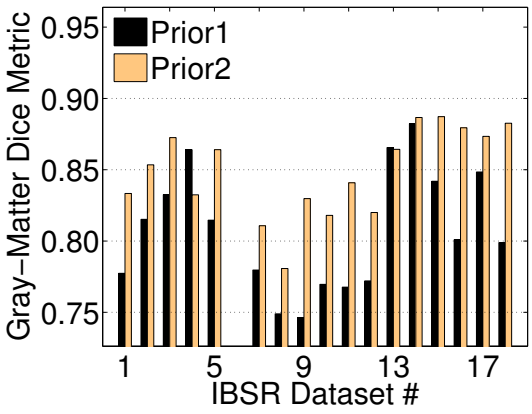
The results in the chapter empirically confirm that the piecewise stationary-ergodic Markov model conforms well to brain MR images. It shows that it is possible to learn these models via nonparametric density estimation in the high-dimensional spaces of MR-image neighborhoods. These results also suggest that the statistical structure in these spaces capture important tissue properties in brain MR images. The mathematical and engineering components in this chapter are appropriate for any kind of densely-sampled medical data, including vector-valued images (e.g., multimodal MR data) and images with higher-dimensional domains (e.g., a sequence of volumetric MR images over time).

**Table 6.2**. Mean, median, and standard deviation for the gray-matter and white-matter tissue classes in the IBSR data set using the proposed method with the *2-class* prior.

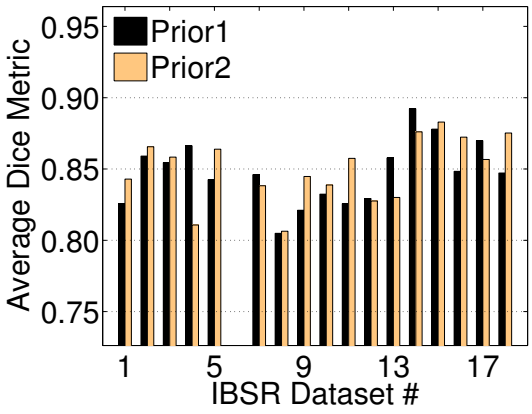
| Statistical measure | White matter | Gray matter |
|---|---|---|
| Mean | 0.8868 | 0.8074 |
| Median | 0.8913 | 0.8009 |
| Standard deviation | 0.0179 | 0.0426 |




(a)




(b)




(c)

**Figure 6.6**. Validation, of the proposed method with two different atlas-based priors, on IBSR data. Dice metrics for (a) white matter: $D_{\text{white}}$, (b) gray matter: $D_{\text{gray}}$, and (c) their average: $(D_{\text{white}} + D_{\text{gray}})/2$. *Note*: In the graphs, Prior1: *2-class* prior, Prior2: *scaled-atlas* prior.

# CHAPTER 7

# TEXTURE SEGMENTATION USING FAST LEVEL-SET PROPAGATION DRIVEN BY MUTUAL INFORMATION

This chapter addresses the problem of segmenting *textured* images. Textured regions do not typically adhere to the piecewise-smooth or piecewise-constant assumptions that characterize most intensity-based segmentation problems. Julesz [82] pioneered the statistical analysis of textures and characterized textures as possessing regularity in the higher-order intensity statistics. This establishes the description of a textured image, or a *Julesz ensemble*, as one derived from stationary MRFs [189]. This principle forms the foundation of the proposed approach.

Image segmentation is one of the most extensively studied problems in computer vision. The literature gives numerous approaches based on a variety of partitioning criteria including intensity, color, texture, depth, and motion. The state-of-the-art in texture segmentation incorporates several important pieces of technology. One important component is the mechanism used to model or quantify the regularity in image textures. Researchers have developed progressively-richer descriptions of local image geometry [18, 148, 149] and sophisticated statistically-based metrics [39, 87, 124, 83, 144]; thereby capturing more complex distinctions between textures. Another area of focus, like in general image segmentation, concerns robust mechanisms for enforcing geometric smoothness on the segmented-region boundaries [118, 153, 117].

This chapter presents a method [4] that exploits the defining characteristics of a texture coupled with the generality of nonparametric statistical modeling. The method relies on an information-theoretic metric on Markov image statistics. The nonparametric modeling of the statistics of the stationary MRF imposes very few restrictions on the

statistical structure of neighborhood intensities. This enables the method to easily adapt to a variety of textures. The method does not rely on a training stage and, hence, is unsupervised. These properties make it is easily applicable to a wide range of texture-segmentation problems. Moreover, the method incorporates relatively recent advances in *level-set* evolution strategies that use *threshold dynamics* [54, 53].

## 7.1 Overview of Texture Segmentation

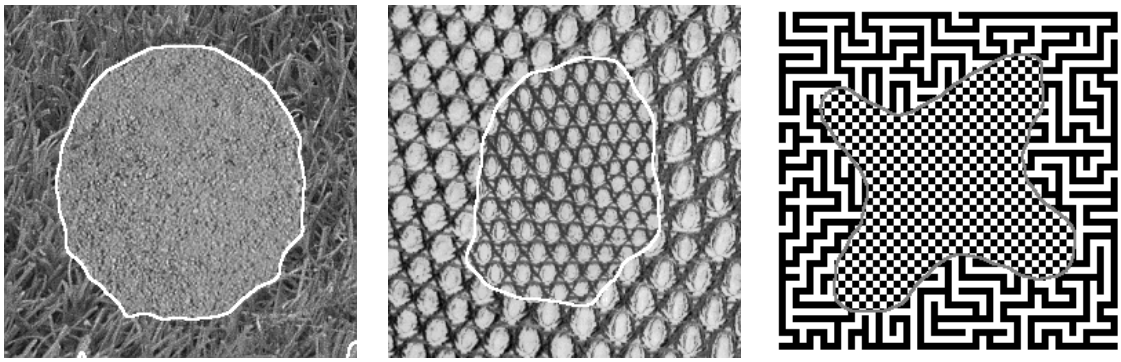
Much of the previous work in texture segmentation employs filter banks, comprising both isotropic and anisotropic filters, to capture texture statistics. For instance, researchers have used Gabor-filter responses to discriminate between different kinds of textures [124, 148, 149]. Gabor filters are a prominent example from the class of oriented multiscale filters [39, 21]. This approach emphasizes the extraction of appropriate features for discriminating between specific textures, which is typically a nontrivial task. The proposed method, on the other hand, does not rely on using specific descriptors that work for certain kinds of textures, but is based on a more generic approach that tries to adaptively capture the core properties of a wide variety of textures.

Researchers have also investigated using more compact sets of texture features. For instance, Bigun *et al.* [18] use the *structure tensor*, which includes all derivatives upto second order, to detect local orientation. Rousson *et al.* [144] refine this strategy by using vector-valued anisotropic diffusion, instead of Gaussian blurring, on the feature space formed using the components of the structure tensor. This strategy requires the structure tensors to have a sufficient degree of homogeneity within regions as well as sufficient dissimilarity between regions. However, as the coming paragraphs explain, not all texture images can be distinguished using these criteria.

Other approaches use the intensity (or grayscale) histograms to distinguish between textures [87, 83]. However, the grayscale intensity statistics (i.e., 1D histograms), may fail to capture the *geometric* structure of neighborhoods, which is critical for distinguishing textures with similar 1D histograms. The proposed method exploits higher-order image statistics, modeled nonparametrically, to adaptively capture the geometric regularity in textures.

Figure 7.1(a) shows two textures that are both *irregular* (in addition to having similar means and gradient-magnitudes) that would pose a challenge for structure-tensor-based approaches such as [18, 144]. In Figure 7.1(b) the textures differ *only in scale*. Approaches based on structure tensors at a single scale would fail to distinguish such cases, as reported in [144]. Approaches solely using intensity histograms would also fail here. In Figure 7.1(c) the textures have identical histograms, identical scale, and an almost-identical set of structure-tensor matrix components. In this case, the above-mentioned approaches [18, 144] would face a formidable challenge. The proposed method, on the other hand, incorporating a fundamentally-richer texture description, produces successful segmentations (depicted by white/gray outlines) for all the images in Figure 7.1.

Recently, researchers have investigated more direct approaches towards modeling image statistics. For instance, the dynamic-texture segmentation approach by Doretto *et al.* [46] uses a Gauss-Markov process to model the relationships among pixels within regions and over time. However, that approach assumes a Gaussian process for image intensities, a restrictive assumption that cannot easily account for complex or subtle texture differences [46, 144, 39, 189]. Rousson *et al.* [144] use nonparametric statistics for one of the channels (the image-intensity histogram) in their feature space to counter



(a) (b) (c)

**Figure 7.1**. Segmentations with the proposed approach (depicted by white/gray outlines) for (a) Brodatz textures for sand and grass— both *irregular* textures with similar gradient magnitudes, (b) Brodatz textures differing *only in scale*, and (c) synthetic textures with identical histograms, identical scales, and an almost-identical set of structure-tensor matrix components.

this restriction and the proposed method generalizes that strategy to the complete higher-order image statistics.

## 7.2 Texture Segmentation Using Mutual Information

The problem of texture segmentation is, at a high level, very similar to that of MRI classification that we considered in the previous chapter—essentially, we want to partition the image into mutually-exclusive and collectively-exhaustive sets in such a way that these partitions comprise stationarity Markov PDFs that are as compact as possible. For the current work, the number of partitions remains a free parameter of the system. For MRI classification, we modeled the intensities in each tissue class as an instance of a stationary MRF. For texture segmentation, this model continues to hold by the employed definition of a texture: regularity in Markov statistics. Therefore, we choose to employ the same optimality metric as before, i.e., the mutual information between the labels and the data.

Consider a discrete RV $L : T \rightarrow \mathbb{Z}$ that maps each voxel $t \in \mathcal{T}$ to the class it belongs to, i.e., $L(t) = k$ if voxel $t$ is in class $k$. Let $\{\mathcal{T}_k\}_{k=1}^{K}$ denote a mutually-exclusive and collectively-exhaustive decomposition of the image domain $\mathcal{T}$ into $K$ regions—assumed stationary—such that $\mathcal{T}_k = \{t \in \mathcal{T} : L(t) = k\}$. The stationarity assumption implies that for each class $k$ the Markov PDFs are exactly the same, i.e.,

$$\forall t \in \mathcal{T}, P(\mathbf{Z}_t | L(t) = k) = P_k(\mathbf{Z}). \tag{7.1}$$

We define the optimal segmentation as the one that maximizes the mutual information between $L$ and $Z$, i.e.,

$$\begin{aligned} I(L, \mathbf{Z}) &= h(\mathbf{Z}) - h(\mathbf{Z}|L) \\ &= h(\mathbf{Z}) - \sum_{k=1}^{K} P(L = k) h(\mathbf{Z}|L = k). \end{aligned} \tag{7.2}$$

Thus, the optimal segmentation is

$$\underset{\{\mathcal{T}_k\}_{k=1}^{K}}{\operatorname{argmin}} \left( -\frac{1}{|\mathcal{T}|} \sum_{k=1}^{K} \sum_{t \in \mathcal{T}_k} \log P_k(\mathbf{z}_t) \right). \tag{7.3}$$

This rather-large nonlinear optimization problem potentially has many local minima. Similar to the approach for MRI classification in the previous chapter, we impose

smoothness constraints on the Markov PDFs via a suitable choice of the kernel-parameter $\sigma$. For texture segmentation, however, we have found that we need additional smoothness constraints on the boundaries of the segmented regions because of: (a) the higher variability in textures encountered in real images that does not conform very well with the stationary-ergodic Markov model, and (b) we do not use any prior information to obtain a good initial-segmentation estimate like the one for MRI brain tissue classification. To impose such regularizations, we can use standard variational formulations, such as the *level-set* framework [118, 153, 117]. Thus, we borrow from a rather extensive body of work on variational methods for image segmentation, in particular the Mumford-Shah model [110], its extensions to motion, depth, and texture, and its implementation via level-set flows [153, 168, 117].

## 7.3 Level-Set Optimization

Level-set methods [118, 153, 117] deform implicitly-defined surfaces, i.e., boundaries of regions, using PDEs and have applications in wide-ranging fields including computer vision [88, 179, 97], image processing [181, 101, 178, 51], visualization [180, 177, 96, 95], graphics [58, 164], and computational physics [153, 117]. Level-set methods form a powerful tool in modeling surface deformations because they avoid many problems associated with deformations using parametric surfaces. For instance, the deformation of parametric surfaces often requires frequent regularization of surface elements without which the deterioration of the surface can lead to numerical inaccuracies and instabilities [117]. Moreover, handling topological changes like merging and splitting of parametrically-represented surfaces can be complicated and computationally expensive.

The level-set method represents the deforming surface using a scalar function $\Phi(t, \tau) : \Re^n \times \Re \rightarrow \Re$ of (a) $t \in \Re^n$: the pixel coordinate in an $n$D Cartesian space and (b) $\tau$: the time variable corresponding to the process of deformation. This function is called the *embedding*. In this chapter, we consider 2D images and therefore $n = 2$. The level sets in our case are curves that are the boundaries of regions and can be defined, with loss of generality, as the zero level set of the embedding $\Phi(\cdot)$, i.e., the set $\{t : \Phi(t, \tau) = 0\}$. The motion of the surface is computed by solving a corresponding PDE on the embedding

$$\frac{\partial \Phi(t,\tau)}{\partial \tau} = F(t,\tau), \tag{7.4}$$

where $F(t,\tau)$ is typically a data-driven location-dependent term that is dictated by the optimization problem—in this chapter we need to solve (7.3).

A straightforward strategy for computing the surface deformation is to solve the level-set PDE on the entire embedding, and in this way the nested family of level sets evolve simultaneously. If one is interested in just a single level set (i.e., a single curve or surface—in our case, the region boundary), this strategy is inefficient, because each level set evolves independently from the others. The narrow band approach [3] exploits this fact and solves the level-set PDE in a band of grid points around the level set of interest, generating a computational speedup of an order of magnitude. Whitaker [179] proposed the sparse-field method, which restricts the computational domain to a few layers around the designated level set. The layers are visited via a linked-list data structure, and the domain is updated as the surface moves. This approach, and the related approach of [126], have a computational complexity like that of parametric surfaces, which is proportional to the area of the surface rather than the volume of the space in which the surface is embedded.

The level-set framework is an attractive option for solving variational problems of the form (7.5), because it restricts neither the shapes nor the topologies of regions. However, classical level-set evolution schemes for front-tracking based on narrow-band strategies entail some significant computational costs—in particular, the Courant-Friedrichs-Lewy (CFL) condition for numerical stability [153, 117] limits the motion of the moving wavefront (region boundaries) to a distance of one grid pixel per iteration. The literature presents several approaches to address this computational issue including multiresolution approaches [28, 147], graphics-processor-based schemes [95, 96], and shared-memory multiprocessor schemes [7]. Recently, Esedoglu and Tsai introduced a fast level-set algorithm based on *threshold dynamics* [54, 53] for minimizing Mumford-Shah type energies. The proposed method adopts their approach for the level-set evolution but relies on a *multiphase extension* of the basic formulation to enable multiple-texture segmentation [105, 168].

## 7.4 Fast Level-Set Optimization Using Threshold Dynamics

In the method proposed by Esedoglu and Tsai, the embeddings, one for each phase, are maintained as piecewise-constant binary functions. This method, essentially, moves the level set by first updating the embeddings based on a gradient descent on the optimization metric, and then regularizing the region boundaries by Gaussian smoothing the embedding followed by thresholding. This approach does not keep track of points near interfaces or maintain distance transforms for embeddings. It allows new components of a region to crop up at remote locations—we have found that this property allows for very rapid level-set evolution when the level-set location is far from the optimum.

Let $R_k : T \rightarrow \{0, 1\}$ denote the indicator function for region $\mathcal{T}_k$, i.e., $R_k(t) = 1$ for all $t \in \mathcal{T}_k$ and $R_k(t) = 0$ otherwise. The optimal segmentation, after incorporating this penalty using a Lagrange multiplier, is

$$\underset{\{R_k\}_{k=1}^K}{\operatorname{argmin}} \left( -\frac{1}{|\mathcal{T}|} \sum_{k=1}^{K} \sum_{t \in \mathcal{T}} R_k(t) \log P_k(\mathbf{z}_t) + \alpha \sum_{k=1}^{K} \sum_{t \in \mathcal{T}} \parallel \nabla_t R_k(t) \parallel_2 \right), \tag{7.5}$$

where $\alpha \geq 0$ is the regularization parameter and $\nabla_t$ denotes a discrete spatial-gradient operator.

We now let $\{R_k\}_{k=1}^K$ be a set of level-set functions. The segment for texture $k$ is then defined as $\mathcal{T}_k = \{t \in \mathcal{T} | R_k(t) > R_j(t), \forall j \neq k\}$. Coupling (7.5) and (2.53) creates *nested region integrals* that complicate the analytical expressions for the gradient flow associated with the level-set evolution [87, 144, 17]. Besson *et al.* [17] give the level-set speed term for minimizing the energy defined in (7.5) using a gradient-descent optimization scheme as

$$\frac{\partial R_k(t)}{\partial \tau} = \log P_k(\mathbf{z}_t) + \frac{1}{|\mathcal{T}_k|} \sum_{s \in \mathcal{T}_k} \frac{G_d(\mathbf{z}_s - \mathbf{z}_t, \Psi_d)}{P_k(\mathbf{z}_s)} + \alpha \nabla_t \cdot \left( \frac{\nabla_t R_k(t)}{\parallel \nabla_t R_k(t) \parallel_2} \right), \tag{7.6}$$

where $\tau$ denotes the time-evolution variable [87, 144].

## 7.5 Segmentation Algorithm

To obtain an initial segmentation $\{R_k^0\}_{k=1}^K$ given no *a priori* information about the locations of the textures in the images, the proposed method uses randomly generated regions, as shown in Section 7.6, based on the following algorithm.

1. Generate $K$ images of i.i.d. uniform random noise, one for each $R_k^0$.
2. Convolve each $R_k^0$ with a chosen Gaussian kernel.
3. Construct the initial segmentation. $\forall k = 1, \ldots, K, \forall t \in \mathcal{T}$ do: if

$$R_k^0(t) = \max_{j=1,\ldots,K} R_j^0(t), \tag{7.7}$$

then set $R_k^0(t) = 1$, otherwise set $R_k^0(t) = 0$. In case of multiple maxima, assign the pixel to an arbitrary region among them.

The key idea behind this procedure is to try to generate an initial segmentation where each segment is spread out over the image such can we can recover the correct segments irrespective of their position in the image. The variance of the Gaussian-smoothing kernel is related to the size of the correct segments. Excessively high or low smoothing produces segments with almost-identical nonparametric Markov PDFs and, thereby, have higher chances of getting stuck in local minima during the level-set optimization. Effective smoothing produces segments with sufficiently different PDFs that drive the optimization procedure to the global minimum, i.e., the correct segmentation.

Given a segmentation $\{R_k^m\}_{k=1}^K$ at iteration $m$, the iterations in Esedoglu and Tsai's fast level-set evolution scheme [54, 53] proceed as follows.

1. Compute the level-set forces for all pixels in all classes:
   (a) Estimate

$$P_k^m(\mathbf{z}_t), \forall k = 1, \ldots, K, \forall t \in \mathcal{T} \tag{7.8}$$

   via nonparametric Parzen-window density estimation as in (6.10).
   (b) Update the level-sets:

$$R_k'(t) = R_k^m(t) + \beta\left(\log P_k^m(\mathbf{z}_t) + \frac{1}{\mathcal{T}_k}\sum_{s\in\mathcal{T}_k}\frac{G_d(\mathbf{z}_s - \mathbf{z}_t, \Psi_d)}{P_k^m(\mathbf{z}_s)}\right) \tag{7.9}$$

2. Regularize the level-sets:

$$R_k'' = R_k' \otimes G(0, \gamma^2), \tag{7.10}$$

where $\otimes$ denotes convolution and $G(0, \gamma^2)$ is a Gaussian kernel with zero mean and standard deviation $\gamma$.

3. Update the classification: $\forall k = 1, \ldots, K, \forall t \in \mathcal{T}$ do: if

$$R_k''(t) = \max_{j=1,\ldots,K} R_j''(t), \tag{7.11}$$

then set $R_k^{m+1}(t) = 1$, otherwise set $R_k^{m+1}(t) = 0$. In case of multiple maxima, assign the pixel to an arbitrary region among them.

4. Stop upon convergence, i.e., when $\| \ R_k^{m+1} - R_k^m \ \|_2^2 < \epsilon$, where $\epsilon$ is a small threshold.

For a detailed discussion on the relationship between the parameters $\{\beta, \gamma\}$ in the threshold-dynamics framework, and the parameter $\alpha$ in the traditional level-set framework, please refer to [54, 53]. In short, increasing $\beta$ corresponds to increasing the PDE-driven force on the level-set evolution and increasing $\gamma$ results in smoother region boundaries.

## 7.6 Results

This section presents results from experiments with real and synthetic data. The number of regions $K$ is a user parameter and should be chosen appropriately. The neighborhood size, in the current implementation, is also a user parameter. This can be improved by using a multiresolution scheme for the image representation. We use $9 \times 9$ pixels neighborhoods for all examples, unless we explicitly state otherwise. We choose $\beta = 2$, $\gamma = 3$, and $|\mathcal{A}_t| = 1000$. The computation for each iteration is $O(K|\mathcal{A}_t||\mathcal{T}||\mathcal{N}_t|)$. The algorithm typically takes less than $10$ iterations to converge. Each iteration of the proposed method takes about $3$ minutes for a $256 \times 256$ pixels image on a standard Pentium-IV $2.8$GHz workstation. The implementation runs about twice as fast on a dual-processor shared-memory Pentium machine. The implementation in this chapter relies on the Insight Toolkit [2].

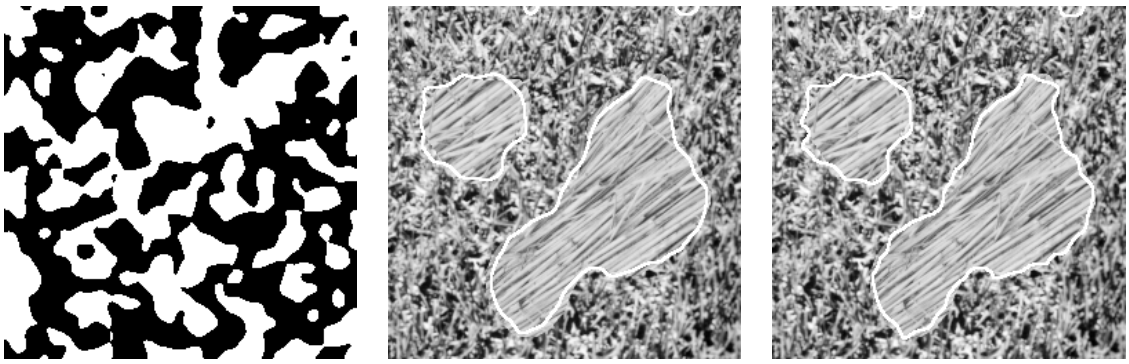
Figure 7.2(a) shows a level-set initialization $\{R_k^0\}_{k=1}^K$ as a randomly generated image with $K = 2$ regions. The level-set scheme using threshold dynamics, coupled with the *global-sampling* strategy as explained in Section 3.5.1, makes the level sets evolve very fast towards the optimal segmentation. We have found that, starting from the random initialization, just a few iterations (less than $10$) are sufficient to reach a virtually-optimal

segmentation. However, this sampling strategy sometimes falls short of giving very accurate boundaries. This is because, in practice, the texture boundaries present neighborhoods overlapping both textures and exhibiting subtleties that may not be captured by the global sampling. Figure 7.2(b) depicts this behavior.

We can handle texture boundaries better by selecting a larger portion of the samples in $\mathcal{A}_t$ from a region close to $t$ might help. Hence, we propose a second stage of level-set evolution for a few iterations that incorporates *local* sampling, in addition to global sampling, and is initialized with the segmentation resulting from the first stage. We found that such a scheme produces consistently better segmentations.

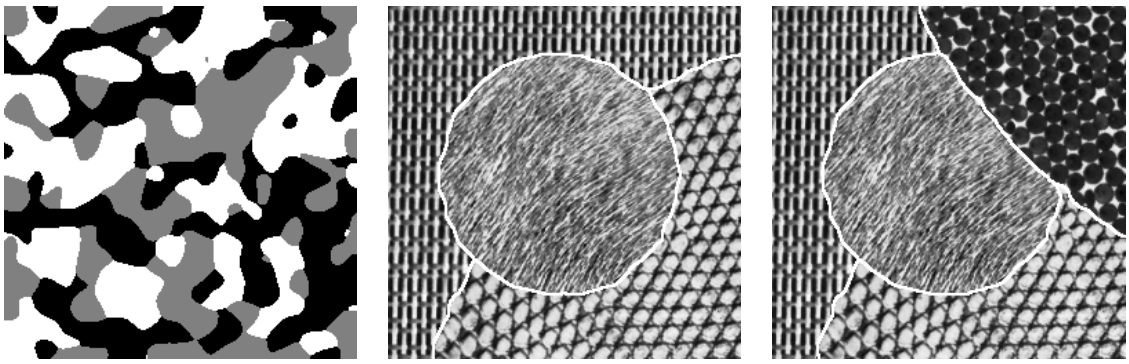
Figure 7.2(c) shows the final segmentation. For each pixel $t$, we have used a random sample of size $|\mathcal{A}_t| = 250$ taken from a Gaussian distribution, with a standard-deviation $\sigma_{\text{spatial}} = 30$ and mean at the pixel $t$. Furthermore, we have found that the method performs well for any choice of the variance such that the Gaussian distribution encompasses more than several hundred pixels. Note that given this variance, both $|\mathcal{A}_t|$ and the Parzen-window $\sigma_{\text{spatial}}$ are computed automatically in a data-driven manner, as explained before in Section 3.5.1 and Section 3.5.2.

Figure 7.3 gives examples dealing with multiple-texture segmentation. Figure 7.3(a) shows a randomly generated initialization with three regions that leads to the final segmentation in Figure 7.3(b). In this case the proposed algorithm uses a multiphase exten-



(a) (b) (c)

**Figure 7.2**. Two-texture segmentation. (a) Random initial segmentation for an image having two Brodatz textures for grass and straw. The black and white intensities denote the two regions. (b) Segmentation after stage 1; *global* samples only (see text). (c) Segmentation after stage 2; *local* and *global* samples (see text).

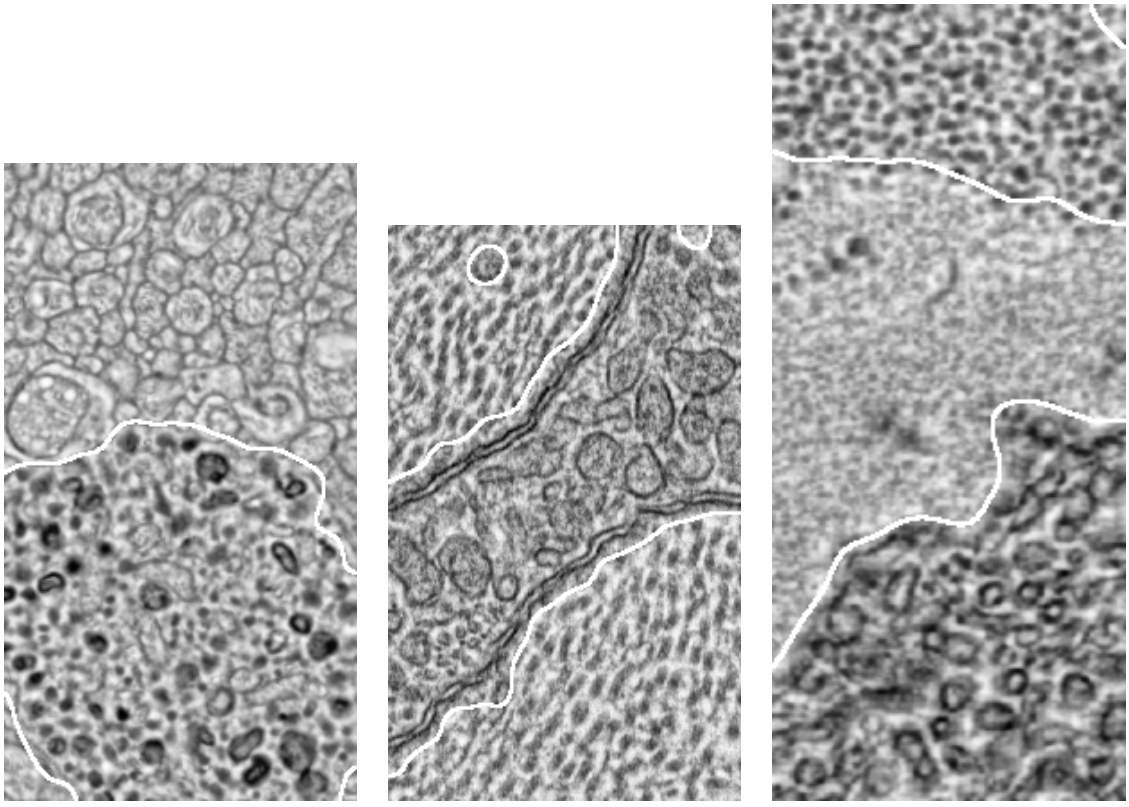


(a) (b) (c)

**Figure 7.3**. Multiple-texture segmentation. (a) Random initial segmentation containing three regions for the image in (b). (b) Final segmentation for an image with three Brodatz textures, including both irregular and regular textures. (c) Final segmentation for an image with four Brodatz textures.

sion of the fast threshold-dynamics based scheme [54, 53]. Figure 7.3(c) shows another multiple-texture segmentation with four textures.

Figure 7.4 shows electron-microscopy images of cellular structures. Because the original images severely lacked contrast, we preprocessed them using adaptive histogram equalization before applying the proposed texture-segmentation method. Figure 7.4 shows the enhanced images. These images are challenging to segment using edge or intensity information because of reduced textural homogeneity in the regions. The discriminating feature for these cell types is their subtle textures formed by the arrangements of sub-cellular structures. To capture the large-scale structures in the images we used larger neighborhood sizes of $13 \times 13$ pixels. We combine this with a higher $\gamma$ for increased boundary regularization. Figure 7.4(a) demonstrates a successful segmentation. In Figure 7.4(b) the two cell types are segmented to a good degree of accuracy; however, notice that the membranes between the cells are grouped together with the middle cell. A third texture region could be used for the membrane, but this is not a trivial extension due to the thin, elongated geometric structure of the membrane and the associated difficulties in the Parzen-window sampling. The hole in the region on the top left forms precisely because the region contains a large elliptical patch that is identical to such patches in the other cell. Figure 7.4(c) shows a successful three-texture segmentation for another image.



(a) (b) (c)

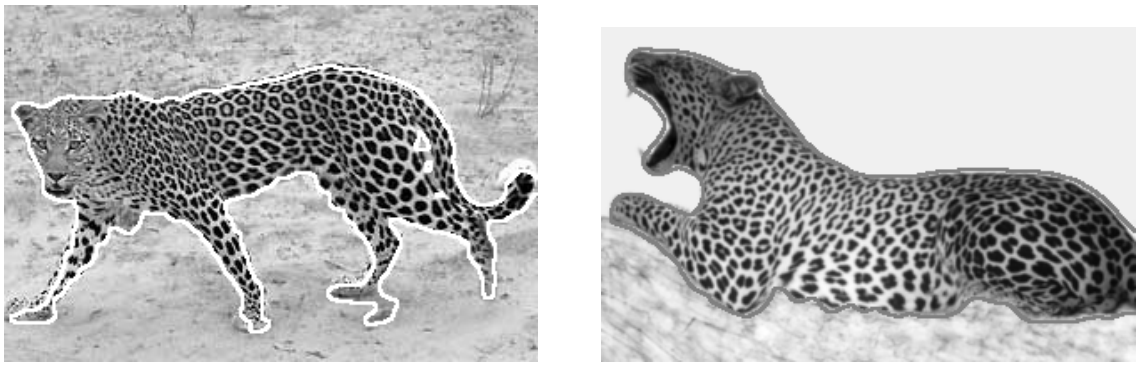
**Figure 7.4**. Final segmentations for electron-microscopy images of rabbit retinal cells for (a),(b) the two-texture case, and (c) the three-texture case.

Figure 7.5(a) shows a zebra example that occurs quite often in the texture-segmentation literature, e.g., [148, 144]. Figures 7.5(b) and 7.5(c) show other zebras. Here, the proposed method performs well to differentiate the striped patterns, with varying orientations and scales, from the irregular grass texture. The grass texture depicts homogeneous statistics. The striped patterns on the zebras' bodies, although incorporating many variations, change gradually from one part of the body to another. Hence, neighborhoods from these patterns form one continuous manifold in the associated high-dimensional feature space, which is captured by the method as a single texture class.

Figure 7.6(a) shows the successful segmentation of the Leopard with the random sand texture in the background. Figure 7.6(b) shows an image that actually contains three different kinds of textures, where the background is split into two textures. Because we

(a) (b) (c)

**Figure 7.5**. Final segmentations for real images of Zebras.



(a) (b)

**Figure 7.6**. Final segmentations for real images of Leopards. Note: The segmentation outline for image (b) is shown in gray.

constrained the number of regions to be two, the method grouped two of the background textures into the same region.

We can alleviate the sensitivity of the model to the neighborhood size by considering a multiscale adaptive-MRF model, which forms an important future engineering extension to the proposed algorithm. Such a model relies on the assumption of MRFs at each level or scale of a specific multiscale image pyramid [122]. This would significantly enhance the utility of the algorithm to images of varied resolutions comprising fractal-like textures with regularities at all scales.

# CHAPTER 8

# CONCLUSIONS

This dissertation describes an adaptive image model that relies on the assumption of image data being derived from a stationary and ergodic MRF. We *empirically* infer the model underlying the data using principles from nonparametric density estimation. The density estimation schemes based on kernel smoothing help to compensate for the sparsity of data in the high-dimensional spaces. We use this model for processing images based on optimal information-theoretic measures and Bayesian decision theory.

We applied the adaptive algorithms for many different tasks concerning image restoration and segmentation. The generic theme underlying the restoration methods was to increase the predictability of pixel intensities from their neighborhoods by reducing the entropy of the pixel-intensity PDFs conditioned on the values of their neighbors. We found that the algorithms perform well on a wide spectrum of images with little parameter tuning. For denoising MR images, we exploited the knowledge of the statistical properties of Rician noise to empirically estimate the uncorrupted-signal Markov statistics from the corrupted-signal Markov statistics. This is essentially involves deconvolving a PDF for which we used the EM algorithm. Subsequently, following the empirical-Bayes approach, we employed the inferred corrupted-signal Markov statistics as a prior in a Bayesian decision-theoretic denoising framework. For the segmentation applications, i.e., MR brain tissue classification and texture segmentation, the key idea was to formulate the problem as one to maximize the mutual information between the Markov PDFs of the data and the unobserved segmentation labels. We had to impose a higher degree of smoothness on the estimated Markov PDFs for regularizing the region boundaries.

This dissertation makes the following contributions to the field of statistical image processing. It presents novel variations on standard MRF-based deterministic algorithms for image restoration, in the form of UINTA and the MRI-denoising method. It provides high-level arguments for the convergence of UINTA, although not a rigorous mathematical proof, and a proof of convergence of the MRI-denoising algorithm. It describes the equivalence between the mode-seeking mean-shift procedure and reducing Shannon's entropy on a nonparametric Parzen-window PDF, thereby providing further insights into the behavior of these algorithms. It exploits the adaptive-MRF model for unsupervised MR brain tissue classification using unimodal and multimodal MR data. This method tries to implicitly handle the noise, inhomogeneity, and partial voluming in the data with reasonable success. It applies these concepts for the classical image-processing tasks of restoration and denoising. The resulting algorithms often perform better than the current state-of-the-art.

There exist several other works where the key ideas relate to the methods in this dissertation. The idea of nonparametrically modeling image statistics is not entirely new. Popat and Picard [131] were the pioneers in employing nonparametric MRF image modeling. Their approach models the Markov PDFs via clustering-based nonparametric density estimation. Our approach, on the other hand, relies on kernel-smoothing approaches. Some texture-synthesis algorithms rely on learning Markov statistics from a sample texture image to construct new images having the same Markov statistics as the input texture [50, 172].

The NL-means algorithm for image denoising by Buades *et al.* [22, 23] computes the denoised image intensity as a weighted average of a sample of image intensities, where the weights are derived from the neighborhoods of the pixels in the sample. The intensity updates in their method are based on the expectation of the conditional Markov PDF $P(X_t|\mathbf{y}_t)$ and closely resemble those in UINTA. While NL-means gets motivation from nonparametric regression theory, UINTA is motivated by information-theoretic concepts coupled with iterative MRF-based image processing.

The MRI-denoising strategy applies Robbins' empirical-Bayes approach for a nonparametric Parzen-window representation of the prior PDF. It also relates to the approach

by Cordy and Thomas [33] that deconvolves PDFs by corrupted with i.i.d. additive Gaussian noise employing the EM algorithm for deconvolving PDFs. Snyder *et al.* [158], similar to our approach, use kernel density estimators for density deconvolution. The DUDE approach by Weismann *et al.* [175] focuses on discrete signal intensities and, subsequently, relies on inverting the channel-transition matrix (noise model) to give a closed form estimate for source statistics from the observed statistics. DUDE then follows an empirical-Bayes strategy for denoising.

Kim *et al.* [87] propose the mutual-information metric for texture segmentation using the intensity (or grayscale) histograms to distinguish between textures. The strategy in this dissertation can be viewed as an extension of the mutual-information metric that exploits the adaptive-MRF image model.

One of the limitations of the algorithms is that they are highly computationally expensive. Therefore, many of our applications are limited to 2D images. We could alleviate this problem by exploiting parallelism in the algorithms or by developing effective fast approximations for the statistical-inference procedures. Parallelizing the algorithms to run on distributed-shared-memory multiprocessor machines or distributed clusters to obtain close-to-linear speedup is a nontrivial task, mainly because of the dynamically-changing Markov PDFs and the random memory-access patterns produced by the stochastic Parzen-window sampling schemes. Parallelization on commodity dual-processor or dual-core processors seems more straightforward, but produces limited gains. Delving into these issues would be an important part of future work.

Rapid advancements in technology, e.g., in medical-imaging and computer vision, will continue to generate new kinds of data along with the challenges of analyzing that data. Adaptive methods can play an important role in cases where accurate model formulation is difficult. Adaptive strategies using nonparametric statistics work best when sufficient data is available that allows empirical learning of the model. One may argue that this may not always be the case and a parametric model, if well designed, can effectively compensate for the scarcity of data. Nevertheless, one of the desirable properties of adaptive algorithms, which are designed to be more *general purpose*, is that their performance degrades gradually as the working conditions deviate from the optimal.

This behavior also echoes in the philosophy behind the classic *no-free-lunch* theorem in optimization theory [182, 48] that basically implies that one optimization strategy can perform better than another, on a specific problem, only if it incorporates prior information specific to that problem. Thus, specialized strategies, like those incorporating strong parametric or prior models, will show more drastic degradation in performance with sub-optimal working conditions. These advantages of adaptive strategies are corroborated by the results in this dissertation: in spite of scarce data in the high-dimensional feature spaces and the arguably-imperfect fit of the stationary-ergodic MRF model in some situations, the proposed algorithms behave robustly and perform well—many times better than the state of the art—for a wide spectrum of data and applications.

# REFERENCES


[1] *National Institute of Health supported Internet Brain Segmentation Repository (IBSR), http://www.cma.mgh.harvard.edu/ibsr*, (2006).

[2] *National Library of Medicine Insight Segmentation and Registration Toolkit (ITK), http://www.itk.org*, (2006).

[3] D. ADALSTEINSSON AND J. SETHIAN, *A fast level set method for propagating interfaces*, J. Comput. Phys., 118 (1995), pp. 269–277.

[4] S. P. AWATE, T. TASDIZEN, AND R. T. WHITAKER, *Unsupervised Texture Segmentation with Nonparametric Neighborhood Statistics*, in Proc. European Conference on Computer Vision (ECCV), 2006.

[5] S. P. AWATE, T. TASDIZEN, R. T. WHITAKER, AND N. L. FOSTER, *Adaptive, nonparametric markov modeling for unsupervised, MRI brain-tissue classification*, Medical Image Analysis, 10 (2006), pp. 726–739.

[6] S. P. AWATE AND R. T. WHITAKER, *Higher-Order Image Statistics for Unsupervised, Information-Theoretic, Adaptive, Image Filtering*, in Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 2, 2005, pp. 44–51.

[7] ——, *An interactive parallel multiprocessor level-set solver with dynamic load balancing*, tech. report, 2005.

[8] ——, *Nonparametric Neighborhood Statistics for MRI Denoising*, in Proc. Int. Conf. Information Processing in Medical Imaging (IPMI), Springer, Lect. Notes in Comp. Sci., vol. 3565, 2005, pp. 677–688.

[9] ——, *Unsupervised, Information-Theoretic, Adaptive Image Filtering for Image Restoration*, IEEE Trans. Pattern Anal. Mach. Intell. (PAMI), 28 (2006), pp. 364–376.

[10] D. BARASH AND D. COMANICIU, *A common framework for nonlinear diffusion, adaptive smoothing, bilateral filtering and mean shift*, Image Vision Comput., 22 (2004), pp. 73–81.

[11] S. BASU, *Rician noise removal in diffusion tensor mri*, Masters Thesis, School of Computing, University of Utah, (2006).

[12] K. N. BERK, *A central limit theorem for m-dependent random variables with unbounded m*, Annals of Prob., 1 (1973), pp. 352–354.

[13] M. BERTALMIO, L. VESE, G. SAPIRO, AND S. OSHER, *Simultaneous structure and texture image inpainting*, IEEE Trans. Image Proc., 12 (2003), pp. 882–889.

[14] J. BESAG, *Spatial interaction and the statistical analysis of lattice systems*, Journal of the Royal Statistical Society, series B, 36 (1974), pp. 192–236.

[15] ——, *Statistical Analysis of Non-Lattice Data*, Journal of the Royal Statistical Society, series B, 24 (1975), pp. 179–195.

[16] ——, *On the statistical analysis of dirty pictures*, Journal of the Royal Statistical Society, series B, 48 (1986), pp. 259–302.

[17] S. J. BESSON, M. BARLAUD, AND G. AUBERT, *Dream2s: Deformable regions driven by an eulerian accurate minimization method for image and video segmentation*, in Proc. European Conf. on Computer Vision-Part III, 2002, pp. 365–380.

[18] J. BIGUN, G. H. GRANLUND, AND J. WIKLUND, *Multidimensional orientation estimation with applications to texture analysis and optical flow*, IEEE Trans. Pattern Anal. Mach. Intell., 13 (1991), pp. 775–790.

[19] J. BILMES, *A gentle tutorial on the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models*, tech. report, University of Berkeley, 1997.

[20] A. BLAKE AND A. ZISSERMAN, *Visual Reconstruction*, MIT Press, 1987.

[21] R. BOOMGAARD AND J. WEIJER, *Robust estimation of orientation for texture analysis*, in 2nd Int. Workshop on Texture Analysis and Synthesis, 2002.

[22] A. BUADES, B. COLL, AND J. M. MOREL, *A non-local algorithm for image denoising.*, in CVPR (2), 2005, pp. 60–65.

[23] ——, *A review of image denoising algorithms, with a new one*, Multiscale Model. Simul., 4 (2005), pp. 490–530.

[24] G. CASELLA, *An introduction to empirical bayes analysis*, The American Statistician, 39 (1985), pp. 83–87.

[25] K. CASTLEMAN, *Digital image processing*, Prentice Hall, 1996.

[26] T. CHAN, J. SHEN, AND L. VESE, *Variational pde models in image processing*, 2003.

[27] Y. CHENG, *Mean shift, mode seeking, and clustering*, IEEE Trans. Pattern Anal. Mach. Intell., 17 (1995), pp. 790–799.

[28] D. CHOPP AND J. SETHIAN, *Motion by intrinsic laplacian of curvature*, Interfaces and Free Boundaries, 1 (1999), pp. 107–123.

[29] Y. CHOW, S. GEMAN, AND L. WU, *Consistent cross-validated density estimation*, Annals of Statistics, 11 (1983), pp. 25–38.

[30] C. COCOSCO, A. ZIJDENBOS, AND A. EVANS, *A fully automatic and robust brain MRI tissue classification method*, Medical Image Analysis, 7 (2003), pp. 513–527.

[31] D. L. COLLINS, A. P. ZIJDENBOS, V. KOLLOKIAN, J. G. SLED, N. J. KABANI, C. J. HOLMES, AND A. C. EVANS, *Design and construction of a realistic digital brain phantom.*, IEEE Trans. Med. Imag., 17 (1998), pp. 463–468.

[32] D. COMANICIU AND P. MEER, *Mean shift: A robust approach toward feature space analysis*, IEEE Trans. Pattern Anal. Mach. Intell., 24 (2002), pp. 603–619.

[33] C. B. CORDY AND D. R. THOMAS, *Deconvolution of a distribution function*, Journal of the American Statistical Association, 92 (1997), pp. 1459–1465.

[34] T. M. COVER AND J. A. THOMAS, *Elements of Information Theory*, Wiley, 1991.

[35] M. D. CRAENE, A. DU BOIS D'AISCHE, B. MACQ, AND S. K. WARFIELD, *Multi-subject registration for unbiased statistical atlas construction.*, in MICCAI, 2004, pp. 655–662.

[36] R. V. CRISTERNA, V. M. BANUELOS, AND O. Y. SUAREZ, *Coupling of radial-basis network and active contour model for multispectral brain MRI segmentation*, IEEE Trans. Biomedical Engineering, 51 (2004), pp. 459–470.

[37] M. B. CUADRA, C. POLLO, A. BARDERA, O. CUISENAIRE, J. VILLEMURE, AND J. THIRAN, *Atlas-based segmentation of pathological MR brain images using a model of lesion growth*, IEEE Transactions on Medical Imaging, 23 (2004), pp. 1301– 1314.

[38] C. DAVATZIKOS AND J. PRINCE, *An active contour model for mapping the cortex*, IEEE Trans. Medical Imaging, 14 (1995), pp. 65–80.

[39] J. S. DE BONET AND P. VIOLA, *Texture recognition using a non-parametric multi-scale statistical model*, in Proc. IEEE Conf. on Comp. Vision and Pattern Recog., 1998, pp. 641–647.

[40] V. DE SILVA AND G. CARLSSON, *Topological estimation using witness complexes*, Symp. on Point-Based Graphics, (2004).

[41] J. DEBONET AND P. VIOLA, *A non-parametric multi-scale statistical model for natural images*, in Proc. Conf. on Adv. in Neural Info. Proc. Systems, 1998, pp. 773–779.

[42] F. DELLAERT, *The expectation maximization algorithm*, tech. report, 2002.

[43] A. P. DEMPSTER, N. M. LAIRD, AND D. B. RUBIN, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society, B (1977), pp. 1–38.

[44] L. R. DICE, *Measures of the amount of ecologic association between species*, Ecology, 26 (1945), pp. 297–302.

[45] D. DONOHO, *Denoising via soft threhsholding*, IEEE Trans. Info. Theory, 41 (1995), pp. 613–627.

[46] G. DORETTO, D. CREMERS, P. FAVARO, AND S. SOATTO, *Dynamic texture segmentation*, in Proc. Int. Conf. Computer Vision, 2003, pp. 1236–1242.

[47] E. DOUGHERTY, *Random Processes for Image and Signal Processing*, Wiley, 1998.

[48] R. DUDA, P. HART, AND D. STORK, *Pattern Classification*, Wiley, 2001.

[49] R. P. W. DUIN, *On the choice of smoothing parameters for parzen estimators of probability density functions*, IEEE Trans. on Computers, C-25 (1976), pp. 1175–1179.

[50] A. A. EFROS AND T. K. LEUNG, *Texture synthesis by non-parametric sampling*, in Int. Conf. Computer Vision, 1999, p. 1033.

[51] V. ELANGOVAN AND R. T. WHITAKER, *From sinograms to surfaces: A direct approach to the segmentation of tomographic data*, in Proceedings of the 4th Int. Conf. Med. Imag. Comp. and Comp.-Assisted Intervention, Springer-Verlag, 2001, pp. 213–223.

[52] I. M. ELFADEL AND R. W. PICARD, *Gibbs random fields, cooccurrences, and texture modeling*, IEEE Trans. Pattern Anal. Mach. Intell., 16 (1995), pp. 24–37.

[53] S. ESEDOGLU, S. RUUTH, AND R. TSAI, *Threshold dynamics for shape reconstruction and disocclusion*, in Proc. Int. Conf. Image Processing, 2005, pp. 502–505.

[54] S. ESEDOGLU AND Y.-H. R. TSAI, *Threshold dynamics for the piecewise constant mumford-shah functional*, no. CAM-04-63, 2004.

[55] M. EVERETT AND J. MANCUSO, *Reptiles, Stanford University*, 2001.

[56] A. FAN, W. WELLS, J. FISHER, M. ÇETIN, S. HAKER, R. MULKERN, C. TEMPANY, AND A. WILLSKY, *A unified variational approach to denoising and bias correction in mr.*, in Info. Proc. Med. Imag., 2003, pp. 148–159.

[57] M. FASHING AND C. TOMASI, *Mean shift is a bound optimization*, IEEE Trans. Pattern Anal. Mach. Intell., 27 (2005), pp. 471–474.

[58] N. FOSTER AND R. FEDKIW, *Practical animations of liquids*, in SIGGRAPH 2001, Computer Graphics Proceedings, E. Fiume, ed., ACM Press / ACM SIGGRAPH, 2001, pp. 23–30.

[59] W. T. FREEMAN AND E. H. ADELSON, *The design and use of steerable filters*, IEEE Trans. Pattern Anal. Mach. Intell., 13 (1991), pp. 891–906.

[60] K. FUKUNAGA AND L. HOSTETLER, *The estimation of the gradient of a density function, with applications in pattern recognition*, IEEE Trans. Info. Theory, 21 (1975), pp. 32–40.

[61] S. GEMAN AND D. GEMAN, *Stochastic relaxation, gibbs distributions and the bayesian restoration of images*, IEEE Trans. Pattern Anal. Mach. Intell., 6 (1984), pp. 721–741.

[62] S. GEMAN AND C. GRAFFIGNE, *Markov random field image models and their applications to computer vision*, in Proc. Int. Congress of Mathematicians, 1986, pp. 1496–1517.

[63] S. GEMAN AND C. R. HWANG, *Nonparametric maximum likelihood estimation by method of sieves*, Annals of Statistics, 10 (1982), pp. 401–414.

[64] G. GERIG, O. KUBLER, R. KIKINIS, AND F. A. JOLESZ, *Nonlinear anisotropic filtering of MRI data*, IEEE Tr. Med. Imaging, 11 (1992), pp. 221–232.

[65] R. C. GONZALEZ AND R. E. WOODS, *Digital Image Processing*, Prentice Hall, 2001.

[66] I. J. GOOD, *A nonparametric roughness penalty for probability densities*, Nature (London), 229 (1971), pp. 29–30.

[67] I. J. GOOD AND R. A. GASKINS, *Nonparametric roughness penalties for probability densities*, Biometrika, 58 (1971), pp. 255–277.

[68] U. GRENANDER, *Abstract Inference*, Wiley, 1975.

[69] R. GUILLEMAUD AND M. BRADY, *Estimating the bias field of mr images*, IEEE Tr. Med. Imaging, 16 (1997), pp. 238–251.

[70] P. HALL, *Cross-validation in density estimation*, Biometrika,, 69 (1982), pp. 382–390.

[71] H. O. HARTLEY, *Maximum likelihood estimation from incomplete data*, Biometrics, 14 (1958), pp. 174–194.

[72] R. HARTLEY, *Transmission of information*, Bell System Technical Journal, (1928), p. 535.

[73] W. HASTINGS, *Monte carlo sampling methods using markov chains and their applications*, Biometrika, 57 (1970), pp. 97–109.

[74] S. HAYKIN, ed., *Unsupervised Adaptive Filtering*, Wiley, 2000.

[75] D. HEALY AND J. WEAVER, *Two applications of wavelet transforms in magnetic resonance imaging*, IEEE Trans. Info. Theory, 38 (1992), pp. 840–860.

[76] K. HELD, E. R. KOPS, B. J. KRAUSE, W. M. WELLS, R. KIKINIS, AND H.-W. MULLER-GARTNER, *Markov random field segmentation of brain mr images*, IEEE Tr. Med. Imaging, 16 (1997), pp. 878–886.

[77] M. HILTON, T. OGDEN, D. HATTERY, G. JAWERTH, AND B. EDEN, *Wavelet denoising of functional MRI data*, 1996, pp. 93–114.

[78] W. HOEFFDING AND H. ROBBINS, *The central limit theorem for dependent random variables*, Duke Math J., 15 (1948), pp. 773–780.

[79] J. HUANG AND D. MUMFORD, *Statistics of natural images and models.*, in Proc. IEEE Comp. Vis. Pattern Recog., 1999, pp. 1541–1547.

[80] A. HYVRINEN, P. HOYER, AND E. OJA, *Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation*, Neural Computation, 11 (1999), pp. 1739–1768.

[81] A. K. JAIN, *Fundamentals of digital image processing*, Prentice-Hall, Inc., 1989.

[82] B. JULESZ, *Visual pattern discrimination*, IRE Trans. Info. Theory, IT (1962), pp. 84–92.

[83] T. KADIR AND M. BRADY, *Unsupervised non-parametric region segmentation using level sets*, in Proc. of IEEE Int. Conference on Computer Vision, 2003, pp. 1267–1274.

[84] J. N. KAPUR, *Measures of Information and their Applications*, John Wiley, 1994.

[85] J. N. KAPUR AND H. K. KESAVAN, *Entropy Optimization Principles with Applications*, Academic Press, Inc., 1992.

[86] T. KAPUR, W. E. L. GRIMSON, W. M. WELLS, AND R. KIKINIS, *Segmentation of brain tissue from magentic resonance images*, Med. Im. An., 1 (1996), pp. 109–127.

[87] J. KIM, J. W. F. III, A. J. YEZZI, M. CETIN, AND A. S. WILLSKY, *A nonparametric statistical method for image segmentation using information theory and curve evolution.*, IEEE Trans. Image Processing, 14 (2005), pp. 1486–1502.

[88] B. KIMIA, A. TANNEBAUM, AND S. ZUCKER, *Shapes, shocks, and deformations I: the components of two-dimensional shape and the reaction-diffusion space*, Int. J. Comput. Vision, 15 (1995), pp. 189–224.

[89] S. KIRKPATRICK, C. D. GELATT, AND M. P. VECCHI, *Optimization by simulated annealing*, Science, Number 4598, 13 May 1983, 220,4598 (1983), pp. 671–680.

[90] N. LAIRD, *Nonparametric maximum likelihood estimation of a mixing distribution*, Journal of the American Statistical Association, 73 (1978), pp. 805–811.

[91] A. LEE, K. PEDERSEN, AND D. MUMFORD, *The nonlinear statistics of high-contrast patches in natural images*, Int. J. Comput. Vision, 54 (2003), pp. 83–103.

[92] S. LEE AND M. VANNIER, *Post-acquisition correction of mr inhomogeneities*, Magnetic Resonance in Medicine, 36 (1996), pp. 275–286.

[93] K. V. LEEMPUT, F. MAES, D. VANDERMEULEN, AND P. SEUTENS, *Automated model-based bias field correction of mr images of the brain*, IEEE Tr. Med. Imag., 18 (1999), pp. 885–896.

[94] ——, *Automated model-based tissue classification of mr images of the brain*, IEEE Tr. Med. Imaging, 18 (1999), pp. 897–908.

[95] A. E. LEFOHN, J. E. CATES, AND R. T. WHITAKER, *Interactive, gpu-based level sets for 3d segmentation*, Medical Image Computing andComputer Assisted Intervention, (2003), pp. 564–572.

[96] A. E. LEFOHN, J. M. KNISS, C. D. HANSEN, AND R. WHITAKER, *Interactive deformation and visualization of level set surfaces using graphics hardware*, IEEE Visualization 2003, (2003), pp. 75–82.

[97] M. LEVENTON, E. GRIMSON, AND O. FAUGERAS, *Statistical shape influence in geodesic active contours*, in CVP'00, 2000.

[98] E. LEVINA, *Statistical issues in texture analysis*, Ph.D. Dissertation, Department of Statistics, University of California, Berkeley, (1997).

[99] S. Z. LI, *Markov Ramdom Field Modeling in Computer Vision*, Springer, 1995.

[100] M. LYSAKER, A. LUNDERVOLD, AND X. TAI, *Noise removal using fourth-order partial differential equation with applications to medical magnetic resonance images in space and time*, IEEE Trans. Imag. Proc., (2003).

[101] R. MALLADI AND J. SETHIAN, *A unified approach to noise removal, image enhancement, and shape recovery*, IEEE Trans. Image Processing, (1996).

[102] S. MALLAT, *A Wavelet Tour of Signal Processing*, Academic Press Inc., 1999.

[103] J. MANGIN, *Entropy minimization for automatic correction of intensity nonuniformity*, in IEEE MMBIA, 2000, pp. 162–169.

[104] G. J. MCLACHLAN, *The EM Algorithm and Extensions*, John Wiley, 1997.

[105] B. MERRIMAN, J. K. BENCE, AND S. OSHER, *Motion of multiple junctions: A level set approach,*, Tech. Report CAM-93-19, Dept. Mathematics, UCLA, 1993.

[106] N. METROPOLIS, A. ROSENBLUTH, M. ROSENBLUTH, A. TELLER, AND E. TELLER, *Equations of state calculations by fast computing machines*, Journal of Chemical Physics, 21 (1953), pp. 1087–1092.

[107] J. MILLER AND C. STEWART, *Muse: Robust surface fitting using unbiased scale estimates*, in Proc. of the Conf. Computer Vision and Pattern Recog., June 1996, pp. 300–306.

[108] T. MINKA, *Expectation maximization as lower bound maximization*, (1998).

[109] G. MOTTA, E. ORDENTLICH, I. RAMIREZ, G. SEROUSSI, AND M. J. WEINBERGER, *The dude framework for continuous tone image denoising*, in Int. Conf. Imag. Proc, 2005, pp. III: 345–348.

[110] D. MUMFORD AND J. SHAH, *Optimal approximations by piecewise smooth functions and associated variational problems*, Com. of Pure and Applied Math., 42 (1989), pp. 577–685.

[111] R. NEAL AND G. HINTON, *A new view of the EM algorithm that justifies incremental and other variants*, tech. report, Department of Computer Science, University of Toronto, 1993.

[112] ——, *A view of the EM algorithm that justifies incremental, sparse, and other variants*, in Learning in Graphical Models, M. I. Jordan, ed., Kluwer Academic Publishers, 1998, pp. 355–368.

[113] L. NOCERA AND J. GEE, *Robust partial volume tissue classification of cerebral mri scans*, in Proc. SPIE Medical Imaging: Image Processing, 1997, pp. 312–322.

[114] K. NORDSTROM, *Biased anisotropic diffusion: a unified regularization and diffusion approach to edge detection*, Image Vision Comput., 8 (1990), pp. 318–327.

[115] R. NOWAK, *Wavelet-based rician noise removal for magnetic resonance imaging*, IEEE Trans. Imag. Proc., 8 (1999), pp. 1408–1419.

[116] H. NYQUIST, *Certain factors affecting telegraph speed*, Bell System Technical Journal, (1924), p. 324.

[117] S. OSHER AND R. FEDKIW, *Level Set Methods and Dynamic Implicit Surfaces*, Springer, 2003.

[118] S. OSHER AND J. SETHIAN, *Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations*, Journal of Computational Physics, 79 (1988), pp. 12–49.

[119] A. OWEN, *Discussion of Ripley*, Canadian J. Statistics, 14 (1986), pp. 106–110.

[120] ——, *Image segmentation via iterated conditional expectations*, tech. report, 1989.

[121] C. PACHAI, Y. M. ZHU, C. GUTTMANN, R. KIKINIS, F. A. JOLESZ, G. GIMENEZ, J.-C. FROMENT, C. CONFAVREUX, AND S. K. WARFIELD, *Unsupervised and adaptive segmentation of multispectral 3d magnetic resonance images of human brain: a generic approach*, in Proc. Int. Conf. Medical Image Computing and Computer Assisted Intervention, 2001, pp. 1067–1074.

[122] R. D. PAGET AND D. LONGSTAFF, *A nonparametric multiscale markov random field model for synthesising natural textures*, Fourth International Symposium on Signal Processing and its Applications, 2 (1996), pp. 744–747.

[123] A. PAPOULIS AND S. U. PILLAI, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, 4th ed., 2001.

[124] N. PARAGIOS AND R. DERICHE, *Geodesic active regions and level set methods for supervised texture segmentation*, Int. J. Comput. Vision, 46 (2002), pp. 223–247.

[125] E. PARZEN, *On the estimation of a probability density function and the mode*, Annals of Math. Stats., 33 (1962), pp. 1065–1076.

[126] D. PENG, B. MERRIMAN, H. ZHAO, S. OSHER, AND M. KANG, *A pde based fast local level set method*, 1999.

[127] P. PERONA AND J. MALIK, *Scale-space and edge detection using anisotropic diffusion*, IEEE Trans. Pattern Anal. Mach. Intell., 12 (1990), pp. 629–639.

[128] A. PIZURICA, W. PHILIPS, I. LEMAHIEU, AND M. ACHEROY, *A joint inter and intrascale statistical model for bayesian wavelet based image denoising*, IEEE Trans. Imag. Proc., 11 (2002), pp. 545–557.

[129] A. PIZURICA, W. PHILIPS, I. LEMAHIEU, AND M. ACHEROY, *A versatile wavelet domain noise filtration technique for medical imaging.*, IEEE Trans. Med. Imaging, 22 (2003), pp. 323–331.

[130] K. POHL, W. E. L. GRIMSON, S. BOUIX, AND R. KIKINIS, *Anatomical guided segmentation with non-stationary tissue class distributions in an expectation-maximization framework*, in International Symposium on Biomedical Imaging, 2004, pp. 81–84.

[131] K. POPAT AND R. PICARD, *Cluster based probability model and its application to image and texture processing.*, IEEE Trans. Image Processing, 6 (1997), pp. 268–284.

[132] J. PORTILLA, *Full blind denoising through noise covariance estimation using gaussian scale mixtures in the wavelet domain*, in Proc. IEEE Int. Conf. Image Proc., October 2004, pp. 1217–1220.

[133] J. PORTILLA, V. STRELA, M. WAINWRIGHT, AND E. SIMONCELLI, *Image denoising using scale mixtures of gaussians in the wavelet domain*, IEEE Trans. Imag. Proc., 12 (2003), pp. 1338–1351.

[134] M. PRASTAWA, J. H. GILMORE, W. LIN, AND G. GERIG, *Automatic segmentation of neonatal brain MRI*, in Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention, 2004, pp. 10–17.

[135] J. PRINCIPE, J. F. III, AND D. XU, *Information theoretic learning*, in Unsupervised Adaptive Filtering, S. Haykin, ed., Wiley, 2000, pp. 265–319.

[136] J. RAJAPAKSE, J. GIEDD, AND J. RAPOPORT, *Statistical approach to segmentation of single-channel cerebral MR images*, IEEE Transactions on Medical Imaging, 16 (1997), pp. 176–186.

[137] S. S. RAO, *Engineering Optimization, Theory and Practice*, Wiley, 1996.

[138] A. RENYI, *On measures of information and entropy*, in Proc. 4th Berkeley Symposium on Mathematics, Statistics and Probability, 1960, pp. 547–561.

[139] D. E. REX, J. Q. MA, AND A. W. TOGA, *The LONI pipeline processing environment*, NeuroImage, 19 (2003), pp. 1033–1048.

[140] H. ROBBINS, *The empirical bayes approach to statistical decision problems*, Annals of Mathematical Statistics, 35 (1964), pp. 1–20.

[141] ——, *An empirical bayes approach to statistics*, in Proc. Third Berkeley Symp. Math. Stat. Prob., 1964, pp. 157–164.

[142] T. ROHLFING AND C. R. M. JR., *Multi-classifier framework for atlas-based image segmentation.*, in IEEE Int. Conf. Comp. Vis. Pattern Recog., 2004, pp. 255–260.

[143] B. ROMENY, ed., *Geometry-Driven Diffusion in Computer Vision*, Kluwer Academic Publishers, 1994.

[144] M. ROUSSON, T. BROX, AND R. DERICHE, *Active unsupervised texture segmentation on a diffusion based feature space*, in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, IEEE Computer Society, 2003, pp. 699–706.

[145] L. RUDIN, S. OSHER, AND E. FATEMI, *Nonlinear total variation based noise removal algorithms*, Physica D, 60 (1992), pp. 259–268.

[146] A. RUF, H. GREENSPAN, AND J. GOLDBERGER, *Tissue classification of noisy mr brain images using constrained gmm.*, in Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention, 2005, pp. 790–797.

[147] M. RUMPF AND R. STRZODKA, *Level set segmentation in graphics hardware*, in IEEE Int. Conf. Imag. Proc., 2001, pp. 1103–1106.

[148] C. SAGIV, N. A. SOCHEN, AND Y. Y. ZEEVI, *Texture segmentation via a diffusion segmentation scheme in the gabor feature space*, in 2nd Int. Workshop on Texture Analysis and Synthesis, 2002.

[149] B. SANDBERG, T. CHAN, AND L. VESE, *A level-set and gabor-based active contour algorithm for segmenting textured images*, Tech. Report CAM-02-39, Dept. Mathematics, UCLA, 2002.

[150] D. W. SCOTT, *Multivariate Density Estimation*, Wiley, 1992.

[151] D. W. SCOTT AND L. E. FACTOR, *Monte carlo study of three data-based nonparametric probability density estimators*, Journal of the American Statistical Association, 76 (1981), pp. 9–15.

[152] L. SENDUR AND I. SELESNICK, *Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency*, IEEE Trans. Signal Proc., 50 (2002), pp. 2744–2756.

[153] J. SETHIAN, *Level Set Methods and Fast Marching Methods*, Cambridge Univ. Press, 1999.

[154] C. SHANNON, *A mathematical theory of communication*, Bell System Tech. Journal, 27 (1948), pp. 379–423.

[155] B. SILVERMAN, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, 1986.

[156] J. S. SIMONOFF, *Smoothing Methods in Statistics*, Springer, 1996.

[157] J. G. SLED, A. P. ZIJDENBOS, AND A. C. EVANS, *A nonparametric method for automatic correction of intensity nonuniformity in mri data*, IEEE Trans. Medical Imaging, 17 (1998), pp. 87–97.

[158] D. L. SNYDER, M. MILLER, AND T. SCHULTZ, *Constrained probability-density estimation from noisy data*, in Proc. 22nd Annual Conference on Information Sciences and System, 1988, pp. 170–172.

[159] W. SNYDER, Y. HAN, G. BILBRO, R. WHITAKER, AND S. PIZER, *Image relaxation: Restoration and feature extraction*, IEEE Trans. Pattern Anal. Mach. Intell., 17 (1995), pp. 620–624.

[160] J. STARCK, E. CANDES, AND D. DONOHO, *The curvelet transform for image denoising*, IEEE Trans. Imag. Proc., 11 (2000).

[161] H. STARK AND J. W. WOODS, *Probability and random processes with applications to signal processing*, Prentice Hall, 2001.

[162] M. STUDENY AND J. VEJNAROVA, *The multiinformation function as a tool for measuring stochastic dependence*, in Learning in Graphical Models, M. I. Jordan, ed., Kluwer Academic Publishers, 1998, pp. 261–297.

[163] T. TASDIZEN, S. P. AWATE, R. T. WHITAKER, AND N. L. FOSTER, *Mri tissue classification with neighborhood statistics: A nonparametric, entropy-minimizing approach*, in Proc. Int. Conf. Medical Image Computing and Computer Assisted Intervention (MICCAI), vol. 3750, 2005, pp. 517–525.

[164] T. TASDIZEN, R. WHITAKER, P. BURCHARD, AND S. OSHER, *Geometric surface processing via normal maps*, ACM Trans. on Graphics, (2003).

[165] A. TOGA, *Brain Warping*, Academic Press, 1999.

[166] C. TOMASI AND R. MANDUCHI, *Bilateral filtering for gray and color images*, in Proc. Int. Conf. Comp. Vision, IEEE Computer Society, 1998, p. 839.

[167] K. S. TRIVEDI, *Probability and Statistics With Reliability, Queuing, and Computer Science Applications*, Wiley, 2001.

[168] L. VESE AND T. CHAN, *A multiphase level set framework for image segmentation using the mumford and shah model*, Tech. Report CAM-01-25, Dept. Mathematics, UCLA, 2001.

[169] L. A. VESE AND S. J. OSHER, *Modeling textures with total variation minimization and oscillating patterns in image processing*, J. Sci. Comput., 19 (2003), pp. 553–572.

[170] P. VIOLA AND W. WELLS, *Alignment by maximization of mutual information*, in Int. Conf. Comp. Vision, 1995, pp. 16–23.

[171] M. WAND AND M. JONES, *Kernel Smoothing*, Chapman and Hall, 1995.

[172] L. WEI AND M. LEVOY, *Order-independent texture synthesis*, Stanford University Computer Science Department Tech. Report TR-2002-01, (2002).

[173] J. WEICKERT, *Anisotropic Diffusion in Image Processing*, Teubner-Verlag, 1998.

[174] J. WEICKERT, *Coherence-enhancing diffusion filtering*, Int. J. Comp. Vis., 31 (1999), pp. 111–127.

[175] T. WEISSMAN, E. ORDENTLICH, G. SEROUSSI, S. VERDU, AND M. WEINBERGER, *Universal discrete denoising: Known channel*, IEEE Trans. Info. Theory, 51 (2005), pp. 5–28.

[176] W. M. WELLS, W. E. L. GRIMSON, R. KIKINIS, AND F. A. JOLESZ, *Adaptive segmentation of MRI data*, IEEE Tr. Med. Imaging, 15 (1996), pp. 429–443.

[177] R. WESTERMANN, C. R. JOHNSON, AND T. ERTL, *Topology preserving smoothing of vector fields*, IEEE Trans. Visualization and Computer Graphics, (2001).

[178] R. WHITAKER, *Reconstructing terrain maps from dense range data*, in IEEE Int. Conf. Image Processing, October 2001, pp. 165–168.

[179] R. T. WHITAKER, *A level-set approach to 3d reconstruction from range data*, Int. J. Comput. Vision, 29 (1998), pp. 203–231.

[180] ——, *Reducing aliasing artifacts in iso-surfaces of binary volumes*, in Proceedings of the 2000 IEEE symposium on Volume visualization, ACM Press, 2000, pp. 23–32.

[181] R. T. WHITAKER AND D. T. CHEN, *Embedded active surfaces for volume visualization*, in SPIE Medical Imaging 94, 1994.

[182] D. WOLPERT AND W. MACREADY, *No free lunch theorems for optimization*, IEEE Trans. Evolutionary Computation, 1 (1997), pp. 67–82.

[183] M. YAN AND J. KARP, *An adaptive bayesian approach to three-dimensional MR brain segmentation*, in Proc. of Information processing in medical imaging, 1995, pp. 201–213.

[184] ——, *Segmentation of 3D brain MR using an adaptive k-means clustering algorithm*, in Proc. of the 1994 Nuclear Science Symposium and Medical Imaging Conference, 1995, pp. 1529–1533.

[185] YANG, DURAISWAMI, GUMEROV, AND DAVIS, *Improved fast gauss transform and efficient kernel density estimation*, in Int. Conf. Comp. Vision, 2003, pp. 464–471.

[186] H. ZHANG, *Image restoration: Flexible neighborhood systems and iterated conditional expectations*, Statistica Sinica, 3 (1993), pp. 117–139.

[187] Y. ZHANG, M. BRADY, AND S. SMITH, *Segmentation of brain mr images through a hidden markov random field model and the expectation maximization algorithm*, IEEE Tr. Med. Imaging, 20 (2001), pp. 45–57.

[188] S. C. ZHU AND D. MUMFORD, *Prior learning and gibbs reaction-diffusion*, IEEE Trans. Pattern Analysis Machine Intell., 19 (1997), pp. 1236–1250.

[189] S. C. ZHU, Y. WU, AND D. MUMFORD, *Frame: Filters, random field and maximum entropy: Towards a unified theory for texture modeling*, Int. J. Computer Vision, 27 (1998), pp. 107–126.