Information Technology: Transmission, Processing, and Storage

Stochastic Image Processing

Chee Sun Won

and

Robert M. Gray

STOCHASTIC IMAGE PROCESSING

Chee Sun Won
Dongguk University
Seoul, South Korea

and

Robert M. Gray
Stanford University
Stanford, California

Kluwer Academic / Plenum Publishers
NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

Library of Congress Cataloging-in-Publication Data

Won, Chee Sun.
Stochastic image processing/Chee Sun Won, Robert M. Gray.
p. cm. — (Information technology: transmission, processing, and storage)
Includes bibliographical references and index.
ISBN 0-306-48192-8
1. Image processing—Digital techniques. 2. Stochastic processes. I. Gray, Robert, M., 1943- II. Title. III. Series.

TA1637.W66 2004
621.36'7—dc22

2003064139

ISBN: 0-306-48192-8

233 Spring Street, New York, New York 10013

http://www.kluweronline.com

10 9 8 7 6 5 4 3 2 1

A C.I.P. record for this book is available from the Library of Congress

Printed in the United States of America

STOCHASTIC IMAGE PROCESSING

Information Technology: Transmission, Processing, and Storage

---

Coded Modulation Systems
John B. Anderson and Arne Svensson

Communication System Design Using DSP Algorithms: With Laboratory Experiments for the TMS320C6701 and TMS320C6711
Steven A. Tretter

A First Course in Information Theory
Raymond W. Yeung

Interference Avoidance Methods for Wireless Systems
Dimitrie C. Popescu and Christopher Rose

Nonuniform Sampling: Theory and Practice
Edited by Farokh Marvasti

Simulation of Communication Systems, Second Edition: Methodology, Modeling, and Techniques
Michel C. Jeruchim, Philip Balaban, and K. Sam Shanmugan

Stochastic Image Processing
Chee Sun Won and Robert M. Gray

To the memory of
Haluk Derin (1945–2002)
and
Roland L. Dobrushin (1929–1995)

Preface

The field of statistical signal processing has contributed a variety of ideas and techniques to the field of image processing during its roughly forty years of existence. Like other areas of signal processing, image processing has incorporated probability, random processes, systems, and transforms into its basic toolbox of approaches and techniques. During recent years an increasingly large role has been played by a particular collection of techniques that were little used during the formative years of image processing: Markov random fields, a generalization of the idea of Markov chains to the two or three dimensional context of images. Although promoted as early as the 1970s by the great Russian information theorist Roland Dobrushin as an excellent model for theoretical studies of images, the ideas were slow to spread to the engineering literature and applications. Even more recently the ideas of hidden Markov models developed with such success in speech processing applications for coding and recognition have been extended to the two and three dimensional context of images, allowing a rich new class of models for image processing applications such as compression, coding, classification, recognition, segmentation, and a variety of forms of image analysis. Some of these methods have extended ideas from one dimensional signals such as speech to two and more dimensions by artificially ordering pixel indexes in a manner analogous to the natural ordering of time for a one dimensional signal, forcing two or three dimensional signals to be spatially "causal" signals. Models and methods not imposing this restraint have required more effort, but the profusion and success of such methods demonstrates that they are coming into their own in terms of both theory and implementation. This book is an effort to provide a survey and comparative development of an exciting and rapidly evolving field of multidimensional Markov and hidden Markov random fields.

CHEE SUN WON AND ROBERT M. GRAY

Acknowledgments

This book would not have been written without the motivation and inspiration provided by our graduate students and the research support for those students and ourselves.

The first author expresses his gratitude to Department of Electrical Engineering at Stanford University for providing facilities while he was writing this book and the support of Dongguk university.

The second author specifically acknowledges the support of the National Science Foundation, the Xerox Foundation, and Hewlett Packard Laboratories for research relating to the topics developed in this book.

Contents

Chapter 1

INTRODUCTION

1.1 Notation

An image will be modeled as a *random field* : $Y = \{Y_s : s \in \Omega\}$, where

- $\Omega = \{(i,j)|0 \le i \le N_1 - 1, 0 \le j \le N_2 - 1\}$ is an index set, a set of site indices on a 2-D discrete $N_1 \times N_2$ rectangular integer lattice as depicted in Figure 1.1.
- For each lattice point or *pixel* $s = (i,j) \in \Omega$, Y_s is a real-valued random variable. For convenience we will use any of the notations Y_s, Y_{ij}, or $Y_{i,j}$ to denote the random variable at a location $s = (i,j) \in \Omega$.
- The random field Y is characterized by a joint probability distribution P_Y, which in turn may be characterized by an associated parameter set θ_Y. When the alphabet is discrete, the joint distribution will be completely described by a joint probability mass function $p_Y(y) = P_Y(\{y\}) = P(Y = y)$ for which

  $$P_Y(F) = P(Y \in F) = \sum_F p_Y(y).$$

  When the alphabet is continuous, we assume the existence of a joint probability density function (pdf) p_Y for which

  $$P_Y(F) = P(Y \in F) = \int_F dy p_Y(y)$$

  for all Borel sets F. Whether lower case p corresponds to a pmf or pdf should be clear from context.

A random field Y is simply a random object with a two dimensional index set. We will use the symbol Y to correspond to an observable image in the

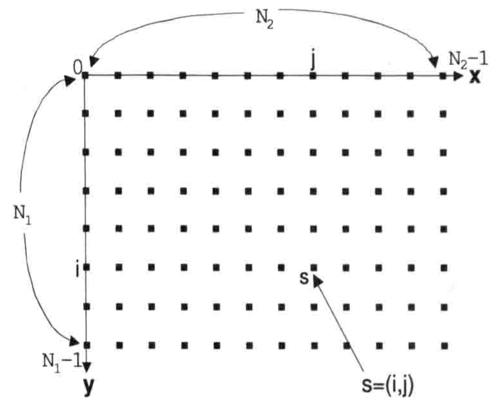

Figure 1.1. Site convention for 2-D rectangular image lattice, where $s = (i, j) \in \Omega$.

sense that its sample values will be directly seen or measured. It will often be the case that the random variables Y_s will take on sample values or realizations y_s from a common finite set of integers $\{0, 1, 2, \cdots, L_Y - 1\}$.[1] For example, for a graylevel image with 8-bit quantization, we have $L_Y = 256$ and $y_s \in \{0, 1, 2, \cdots, L_Y - 1\}$. Depending on the application at hand, however, y_s may represent some feature values extracted from the graylevels such as the discrete cosine transform (DCT) or discrete wavelet transform (DWT) coefficients.

Let the range space or alphabet of the random variables Y_s be $\mathcal{A}_{Y_s}$ and the set of all possible realizations of Y be the Cartesian product $\mathcal{A}_Y = \prod_{s\in\Omega} \mathcal{A}_{Y_s}$. We will assume that all of the $\mathcal{A}_{Y_s}$ are identical.

We also consider a second random field $X = \{X_s : s \in \Omega\}$, which is not observable but which may represent a hidden state (i.e., which is not directly observable and which must be estimated based on the observed Y). The random field X is used to represent unknown and unobservable labels corresponding Y. For example, X_s might identify a type or class to which Y_s belongs or the observation Y might be a noisy or distorted version of X. If X is discrete, it indicates a *segmentation* of the image Y into regions with common values of X and the realization x_s of X_s takes one of L_X class labels, i.e. $x_s \in \{0, 1, \cdots, L_X - 1\}$. When X is continuous, it may be interpreted as a denoised or enhanced version of Y.

[1] We always use upper case letters for random variables and lower case letters for their values (or realizations).

As with the random field Y, the random field X is similarly described by a range space or alphabet $\mathcal{A}_X$, a distribution P_X, and a parameter set θ_X. When the alphabet is discrete, the joint distribution will be completely described by a joint probability mass function $p_X(x);\ x \in \mathcal{A}_X$. When it is continuous, the joint distribution will be completely described by a probability density function $p_X(x);\ x \in \mathcal{A}_X$.

In order to relate random fields X and Y we also require the existence of an underlying joint probability distribution $P_{X,Y}$ which implies the induced conditional distributions $P_{X|Y}$ and $P_{Y|X}$ and the "marginal" distributions P_X and P_Y for the individual random fields. A general goal will be to find an "optimal" value x^* for X based on an observation $Y = y$.

We will usually focus on discrete random variables and hence pmfs will be emphasized and we will write sums rather than integrals and treat the probability functions as actual probabilities, e.g., $p_{X|Y}(x|y) = P(X = x|Y = y)$. Our aim, however, is to simplify the development and not to restrict consideration to discrete-valued random fields. The formulas extend to the continuous case by the usual method of replacing sums by integrals and the densities can be interpreted as approximate probabilities by multiplying them by a differential volume. In order to simplify notation and the effort required for handling continuous, discrete, and mixed cases, we adopt the following notation: $P(x)$ will represent the pmf $P(X = x) = p_X(x)$ in the discrete case and the pdf $p_X(x)$ in the continuous case. This employs the common abuse of notation of letting the independent variable imply the random object in question. $P(y)$ will similarly represent the pmf or pdf as appropriate for Y. Conditional distributions will be similarly abbreviated with $P(x|y)$ representing a conditional pdf or pmf $p_{X|Y}(x|y)$ and similarly for $P(y|x)$. Usually we will implicitly assume the discrete case and use sums, but in continuous examples such as the Gaussian examples, the obvious replacement of sums by integrals and probability mass functions by density functions should be made. When necessary for clarity, we will occasionaly use the longer notation, e.g., $P(X = x|Y = y)$ for a conditional pmf.

1.2 A Brief History of MRF-Based Image Modeling

Early work on Markov random fields (MRFs) and Gibbs random field (GRFs) began in the 1950's [179], 1960's [1, 58], and 1970's [13, 131, 189]. These seminal works concentrated primarily on the Markovian and Gibbsian models on 2-D lattice structures. Since 1970, MRF models have been applied to image spaces for image segmentation and restoration problems (see for example [14, 50, 72]). These works exploited the analogy between very large spatial stochastic systems of digital images and the lattice-based systems of statistical mechanics. It is not surprising that most of their answers to the following important questions are based on statistical mechanics:

(i) How are the contextual relationships among pixel labels established in a noncausal way?

(ii) With the stochastic models defined in (i), how is a segmentation or restoration problem formulated as an optimization process with a well defined cost function?

(iii) How is the optimal solution defined in (ii) obtained?

For the first question, the obvious obstacle is the extension of *causal* 1-dimensional (1-D) signal processing techniques to *noncausal* 2-dimensional (2-D) spatial data. Since there is no natural definition for past and future for pixels in the spatial domain, a direct application of the 1-D Markov chain (MC) model to 2-D image data is not possible. This led researchers to introduce a 2-D noncausal Markovian property, which is defined by a neighborhood system. Although the MRF model allows us to represent the spatial dependence of the class label field in a noncausal way, its local characteristics expressed by noncausal neighbors creates another problem. Specifically, unlike the 1-D MC, the joint probability of the noncausal MRF is not factored into the local characteristics. This implies that we need a consistent way of constructing the joint probability using the local characteristics. Fortunately, the joint probability can be obtained alternatively by using the Hammersley-Clifford theorem [13]. The theorem says that any MRF is equivalent to a GRF. Note that the MRF is based on the conditional probability, whereas the GRF is defined by a joint distribution. Therefore, the MRF-GRF equivalence theorem allows us to represent the joint distribution of X in terms of local conditional probabilities and vice versa. Since we have an explicit expression for the joint probability as well as the conditional probability, the noncausal random field is complete and is ready to use.

The answer to the second and third questions adopted in the earlier works is the Bayesian classifier, specifically, the maximum a posteriori (MAP) estimation criterion. Given the observed image random field $Y = y$, the MAP criterion seeks a labeling $X = x^*$ which maximizes the conditional probability function $p_{X|Y}(x|y)$ for all possible realizations x of X. In the case where the alphabet of X is discrete, this corresponds to maximizing the probability $P(x|y)$ and minimizing the probability of an error or misclassification. Following Bayes' rule, maximizing $P(x|y)$ is equivalent to maximizing $P(y|x)P(x)$. Finding the optimal class label $x^*(y) = \text{argmax}_x P(y|x)P(x)$ by examining all possible class label configurations in $\mathcal{A}_X$ is computationally prohibitive. Instead, in the earlier work, the following assumptions were made to simplify MAP optimization.

(i) Given a realization $x = \{x_s, s \in \Omega\}$ of the class label field X, the observations $Y = \{Y_s, s \in \Omega\}$ are assumed to be of the form $Y_s = X_s + W_s$, $\forall s \in$

Ω, where W is a random field that is independent of X and is independent and identically distributed (iid). In particular, given $X = x$ the Y_s are conditionally independent of each other.

(ii) The random field X is assumed to obey a noncausal MRF property.

Using assumption (i), the MAP criterion can be expressed as

$$\begin{aligned} x^*(y) &= \operatorname*{argmax}_x P(x|y) \\ &= \operatorname*{argmax}_x P(y|x)P(x) \\ &= \operatorname*{argmax}_x \Big[P(x) \prod_{s \in \Omega} P(Y_s = y_s | X_s = x_s)\Big]. \end{aligned}$$

Note that the noncausality of the random field X makes it difficult to arrange the unobservable class label process X in an ordered sequence. As a result, P_X can not be decomposed into (first-order) Markov transition probabilities. Instead, Geman and Geman [72] introduced an alternative stochastic relaxation method called *simulated annealing (SA)* to obtain the optimal class label x^* of the MAP criterion $P(x|y)$. The SA method is based on a *Gibbs sampler*, which generates a sequence of samples of the class labels from the Gibbs distribution of $P(x|y)$. More specifically, starting from an arbitrary realization $x(0)$, its class label of each pixel is repeatedly updated by the local conditional probability of $P(x|y)$. Here, since $P(x|y)$ has a Gibbs distribution, according to the MRF-GRF equivalence theorem, the local conditional probability of $P(x|y)$ satisfies the Markov property and can be expressed as a function of clique potentials of the neighbors. Thus, the local conditional probability is computationally feasible. It has been proven in [72] that in the limit of an infinite number of pixel visits and updates, the updated class label converges to a sample from the Gibbs distribution regardless of the choice of the starting class label field $x(0)$. Moreover, by introducing a proper cooling schedule to the energy function during the update, the Gibbs sampler converges to a maximum state of the realization, which corresponds to a MAP solution. In other words, with a proper cooling schedule, the iterative local updates overcome local maxima and eventually converge to a globally optimal realization x^* of the label field. Note that the nature of the SA algorithm is iterative, which is usually computationally expensive.

When all parameter values associated with the random field models are known *a priori*, the SA and its deterministic relaxation algorithms converge rapidly and yield an acceptable result in a finite number of iterations. However, when it comes to a realistic situation, the algorithm gets extremely complex. That is, when the parameter values are not known, it needs an additional parameter estimation step. Here, we need the known class labels to estimate the model parameter values. At the same time, the estimated model parameter values should be available for the determination of class labels. This "chicken-and-egg" problem led people to seek simultaneous or alternate methods for parameter estimations and class label updates [14, 110, 185, 197], which are computationally quite expensive.

The pixel-wise conditional independence assumption (i.e., $Y_s = X_s + W_s$, $\forall t \in \Omega$, with an iid assumption for W_s) also limits real applications. For most real images, the observed image data are correlated with their neighboring image data given the underlying class label. In this case, the random variable Y_s for observed image data may be modeled by a linear combination of the neighboring image data as well as X_s and W_s. Then, more parameters are needed to represent the contextual relationship among the extended neighbors for observed image data. This in turn increases the computational complexity because of the increased number of model parameters.

In summary, conventional MAP estimation techniques with a simulated annealing paradigm basically rely on iterative and pixel-wise local updates of the class labels. This structure may not be suitable for some realistic problems. First, the iterative nature of the techniques often requires high computational cost. It becomes even worse when the model parameter values need to be estimated during the iterative class label updates. Moreover, if there exist correlations among the observed image data, then the number of model parameters to be estimated increases. Natural images often contain texture regions in which the neighboring image data are highly correlated.

A great deal of effort has been expended to overcome these problems during the past decade. The following three approaches have shown particular promise:

(i) Multiresolution Markov random fields [20, 21, 31, 38, 109, 111, 117, 129]

(ii) Block-based approaches [114, 186, 187]

(iii) Causal Markov chain modeling [93, 103, 108, 114, 117, 135, 150].

Multiresolution extensions of the MRF model can be accomplished by decomposing the image data into different frequency components and scales, enabling the exploitation of image features in various scales. However, new issues arise from the multiresolution approaches such as the estimation of inter-model and intra-model parameters and the Markovianity between consecutive resolution levels and among the spatial data at a specific level. An alternative approach to the multiresolution methods is to deal with a group of pixels (i.e., an image block) together as a super-pixel. Long range correlation in the image data can be effectively treated by extracting the feature within the image block and considering its continuity among the neighboring image blocks, Finally, adopting causal Markov chain modeling, it is expected that well developed estimation