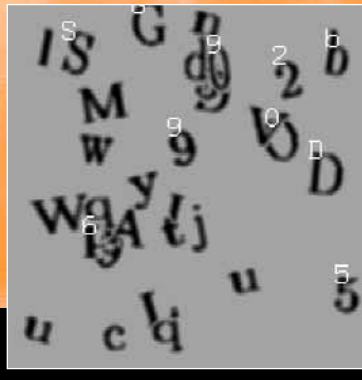
2D Object Detection and Recognition

Models, Algorithms, and Networks



Yali Amit

2D Object Detection and Recognition

Yali Amit

2D Object Detection and Recognition

Models, Algorithms, and Networks

The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Times Roman by Interactive Composition Corporation and was printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Amit, Yali.
2D object detection and recognition : models, algorithms, and networks / Yali Amit.
p. cm.
Includes bibliographical references.
ISBN 0-262-01194-8 (hc. : alk. paper)
1. Computer vision. I. Title.

TA1634 .A45 2002
006.3′7–dc21

2002016508

To Granite, Yotam, and Inbal

Contents

Preface

This book is about detecting and recognizing 2D objects in gray-level images. How are models constructed? How are they trained? What are the computational approaches to efficient implementation on a computer? And finally, how can some of these computations be implemented in the framework of parallel and biologically plausible neural network architectures?

Detection refers to anything from identifying a location to identifying and registering components of a particular object class at various levels of detail. For example, finding the faces in an image, finding the eyes and mouths of the faces. One could require a precise outline of the object in the image, or the detection of a certain number of well-defined landmarks on the object, or a deformation from a prototype of the object into the image. The deformation could be a simple 2D affine map or a more detailed nonlinear map. The object itself may have different degrees of variability. It may be a rigid 2D object, such as a fixed computer font or a 2D view of a 3D object, or it may be a highly deformable object, such as the left ventricle of the heart. All these are considered object-detection problems, where detection implies identifying some aspects of the particular way the object is present in the image—namely, some partial description of the object *instantiation.*

Recognition refers to the classification among objects or subclasses of a general class of objects present in a particular region of the image that has been isolated. For example, after detecting a face, identify the person, or classify images of handwritten digits, or recognize a symbol from a collection of hundreds of symbols. Both domains have a significant training and statistical estimation component.

Finding a predetermined object in a scene, or recognizing the object present in a particular region are only subproblems of the more-general and ambitious goal of computer vision. In broad terms, one would want to develop an artificial system that can receive an image and identify all the objects or a large part of the objects present in

a complex scene from a library of thousands of classes. This implies not only detection and recognition algorithms, but methods for sequentially learning new objects and incorporating them into the current recognition and detection schemes. But perhaps hardest of all is the question of how to start processing a complex scene with no prior information on its contents—what to look for first, and in which particular regions should a recognition algorithm be implemented. This general problem is unsolved, although our visual system seems to solve it effortlessly and very efficiently.

Deformable-template models offer some reasonable solutions to formulating a representation for a restricted family of objects, estimating the relevant parameters and subsequently detecting these objects in the image, at various levels of detail of the instantiation. Each model is defined in terms of a subset of points on a reference grid, *the template,* a set of admissible instantiations of these points, also referred to as *deformations* of the template, and a statistical model for the data—given a particular instantiation of the object is present in the image. A Bayesian framework is used, in that probabilities are assigned to the different instantiations. Bayes's rule then yields a posterior distribution on instantiations. Detections are computed by finding maxima or high values of the posterior. In chapter 2, some general and unifying elements of the Bayesian models used in all the detection algorithms are introduced, together with an overview of the models applied to a simple synthetic example. The details of the detection algorithms are provided in chapters 3–8.

Chapter 9 is devoted to recognition of isolated objects or shapes, assuming some mechanism exists for isolating the individual objects from the more-complex image. The classification schemes can be viewed as a recursive partitioning of the hierarchy of templates using classification trees. Chapter 10 is an exploration into a possible approach to complex scene analysis by merging detection and recognition, both in terms of training and in terms of implementation. Detectors are no longer geared to one particular class, but to object *clusters* containing elements from several classes. Detection can be viewed as a way to quickly choose a number of candidate regions for subsequent processing with a recognition algorithm. An overview of the models of chapters 9 and 10 are also given in chapter 2.

Chapter 11 describes schematic neural network architectures that train and implement detection and recognition algorithms based on the sparse models developed in chapters 6–9. The goal is to show that models based on binary local features, with built-in invariances, simple training procedures, and simple computational implementations, can indeed provide computational models for the visual system. Chapter 12 provides a description of the software and data sets, all of which are accessible through the web at `http://galton.uchicago.edu/~amit/book/`.

The Introduction is used to briefly describe the major trends in computer vision and how they stand in relation to the work in this book. Furthermore, in the last section of each chapter, references to related work and alternative algorithms are provided. These are not comprehensive reviews, but a choice of key papers or books that can point the reader further on.

The emphasis is on simplicity, transparency, and computational efficiency. Cost functions, statistical models, and computational schemes are kept as simple as possible—Occam's razor is too-often forgotten in the computer-vision community. Statistical modeling and estimation play an important role, including methods for training the object representations and classifiers. The models and algorithms are described at a level of detail that should enable readers to code them on their own; however, the readers also have the option of delving into the finest details of the implementations using the accompanying software. Indeed, it is sometimes the case that the key to the success of an algorithm is due to some choices made by the author, which are not necessarily viewed as crucial or central to the original motivating ideas. These will ultimately be identified by experimenting with the software. It is also useful for the readers to be able to experiment with these methods and to discover for themselves the strengths and weaknesses, leading to the development of new and promising solutions.

The images from the experiments shown in the book, and many more, are provided with the software. For each figure in the book, a parameter file has been prepared, allowing the reader to run the program on the corresponding image. This should help jump-start the experimentation stage. Even trying to change parameter settings in these files can be informative, or running them on additional images. Chapter 12 should provide the necessary documentation for understanding the parameters and their possible values.

The examples presented in this book should convince the reader that problems emerging in different computer-vision subcommunities, from the document-analysis community to the medical-imaging community, can be approached with similar tools. This comes at the expense of intensively pursuing any one particular application. Still, the book can be used as a reference for particular types of algorithms for specific applications. These include detecting contours and curves, image warping, anatomy detection in medical images, object detection, and character recognition. There are common themes that span several or all chapters, as well as discussions of connections between models and algorithms. These are, in large part, found in chapter 2 and the introductory comments and the final discussion section of each chapter. It is still possible to study individual models independently of the others.

The mathematical tools used in this book are somewhat diverse but not very sophisticated. Elementary concepts in probability and statistics are essential, including the basic ideas of Bayesian inference, and maximum-likelihood estimation. These can be found in Rice (1995). Some background in pattern recognition is useful but not essential and can be found in Duda and Hart (1973). A good understanding of multivariate calculus is needed for chapters 3 and 5, as well as some basic knowledge of numerical methods for optimization and matrix computation (which can be found in Press and colleagues 1995). The wavelet transform is used in chapters 3 and 5, where a brief overview is provided as well as a description of the discrete wavelet transform. (For a comprehensive treatment of the theory and applications of wavelets, see Wickerhauser 1994.) Some elementary concepts in information theory, such as entropy and conditional entropy, are used in chapters 4 and 9 and are briefly covered in a section of chapter 4. (For a comprehensive treatment of information theory see Cover and Thomas 1991.)

Computer vision is a fascinating subject. On one hand, there is the satisfaction of developing an algorithm that takes in an image from the web or the local webcam and in less than a second finds all the faces. On the other hand are the amazing capabilities of the human visual system that we experience at every moment of our lives. The computer algorithms are nowhere near to achieving these capabilities. Thus, every once in a while, the face detector will miss a face and quite often will select some part of a bookshelf or a tree as being a face. The visual system makes no such mistakes—the ground truth is unequivocal and brutally confronts us at every step of the way. Thus we need to stay humble on one hand and constantly challenged on the other. It is hoped that the reader will become engaged by this challenge and contribute to this exciting field.

Acknowledgments

A large part of the work presented in this book is a result of a long interaction with Donald Geman, to whom I owe the greatest debt. I am referring not only to particular algorithms we developed jointly, but also to endless conversations and exchanges about the subject of computer vision, which have been crucial in the formation of the views presented here. I am deeply thankful to Ulf Grenander for first introducing me to image analysis and deformable-template models. The book as a whole is influenced by his philosophy and also by my interactions with the Pattern Analysis group in the Division of Applied Mathematics at Brown University: Basilis Gidas, Don McClure, David Mumford, and in particular Stuart Geman who, through scattered conversations over the years, has provided invaluable input.

The work on neural network architectures would not have been possible without the recent interaction with Massimo Mascaro. I am indebted to my father Daniel Amit for prodding me to explore the connections between computer vision algorithms and the biological visual system, and for many helpful discussions along the way. Kenneth Wilder contributed immensely to the accompanying software, which would have been unintelligible otherwise. Many thanks to Mauro Piccioni, Kevin Manbeck, Michael Miller, Augustine Kong, Bruno Jedynak, Alejandro Murua, and Gilles Blanchard who have been supportive and stimulating collaborators. I am grateful to the Department of Statistics at the University of Chicago for being so supportive over the past ten years, and to the Army Research Office for their financial support.

2D Object Detection and Recognition

1 Introduction

The goal of computer vision is to develop algorithms that take an image as input and produce a symbolic interpretation describing which objects are present, at what pose, and some information on the three-dimensional spatial relations between the objects. This involves issues such as learning object models, classifiers to distinguish between objects, and developing efficient methods to analyze the scene, given these learned models. Our visual system is able to carry out such tasks effortlessly and very quickly. We can detect and recognize objects from a library of thousands if not tens of thousands in very complex scenes. However, the goal of developing computer algorithms for these tasks is still far from our grasp. Furthermore, there is still no dominant and accepted paradigm within which most researchers are working. There are a number of major trends, briefly described below, relative to which the work in this book is placed.

1.1 Low-Level Image Analysis and Bottom-up Segmentation

Image segmentation is a dominant field of research in the computer vision and image analysis communities. The goal is to extract boundaries of objects or identify regions defined by objects, with no prior knowledge of what these objects are.

The guiding philosophy is that only through such low-level processing is there any chance of identifying more-restricted regions in the scene for further high-level processing, such as recognition. Because these algorithms operate with no higher-level information about the objects, they are referred to as *low-level image analysis*. Another commonly used term is bottom-up image processing.

Many of the early ideas that guided much of the subsequent research can be found in Duda and Hart (1973) and Marr (1982). Motivated by the connections established

by Marr and Hilderith (1980) between edge detection algorithms and computations carried out in the primary visual cortex, a significant body of work in computer vision has been devoted to the specific use of edge detection for segmentation. An edge detector is used to identify all edges in the image, after which some type of local rule tells how to group the edges into continuous contours that provide continuous outlines of the objects. Other approaches to segmentation are region based. Regions with similar characteristics are identified, typically through local region-growing techniques. A detailed description of a variety of such approaches can be found in Haralick and Shapiro (1992).

A statistical formulation of the segmentation problem from a Bayesian point of view was introduced in Geman and Geman (1984), combining region and edge information. An extensive review of such statistical approaches can be found in Geman (1990). The statistical model introduces global information in that the full segmentation is assigned a cost or posterior probability, in terms of the "smoothness" of the different regions and their contours. The various algorithms proposed to optimize this global cost are quite computationally intensive. Other approaches to bottom-up image segmentation currently being proposed can be found in Elder and Zucker (1996); Parida, Geiger, and Hummel (1998); Ishikawa and Geiger (1998); and Shi and Malik (2000).

However, there are some persistent problems with the notion of determining a segmentation of an image without any models of the objects that are expected to be present. First, there is no agreement as to what a good segmentation really is. Furthermore, continuous contours are very difficult to determine in terms of local edges detected in an image. Using local-edge information alone, it is very difficult to actually trace the contour of an object—for example, various noise effects and occlusion can eliminate some of the edges along the contour. A local procedure for aggregating or grouping edges would encounter spurious bifurcations or terminations. Homogeneous regions are difficult to define precisely, and at times, lighting conditions create artificial regions that may cause an object to be split or merged with parts of the background.

As a result, people have tried to incorporate a priori information regarding specific objects in order to assist in identifying their instantiations. This involves more-specific modeling and more-restricted goals in terms of the algorithms. Instead of an initial segmentation that provides the outlines of *all* the objects of interest, which then need to be classified, one tries to directly detect specific objects with specific models. Because shape information is incorporated into the model, one hopes to avoid the pitfalls of the bottom-up approach and really identify the instantiation of these objects. This approach, called *high-level image analysis,* is the main theme of chapters 3–8.

It should be emphasized that all high-level models use some form of low-level processing of the data, and often an initial edge-detection procedure is performed. However, such processing is always geared toward some predefined goal of detecting a specific object or class of objects, and hence are presented only within the context of the entire algorithm. In that sense, there is no meaning to the notion of "good" edge detection, or a "good" image segmentation divorced from the outcome of the high-level algorithm.

1.2 Object Detection with Deformable-Template Models

The need to introduce higher-level object models has been addressed in a somewhat disjointed manner in the statistics community on one hand and in the computer-vision community on the other. In this section, we briefly discuss the former, which is the point of origin for the work in this manuscript.

High-level object models, under the name *deformable-template models,* were introduced in the statistics community in Grenander (1970, 1978). A statistical model is constructed that describes the variability in object instantiation in terms of a prior distribution on deformations of a template. The template is defined in terms of generators and bonds between subsets of generators. The generators and the bonds are labeled with variables that define the deformation of the template. In addition, a statistical model of the image data, given a particular deformation of the template, is provided. The data model and the prior are combined to define a posterior distribution on deformations given the image data. The model proposed by Fischler and Elschlager (1973) is closely related, although not formulated in statistical terms, and is quite ahead of its time in terms of the proposed computational tools. Much of the theory relating to these models is presented in Grenander (1978) and revisited in Grenander (1993). Some applications are presented in the latter part of Grenander (1993). The subject matter has been mostly nonrigid objects in particular objects that occur in biological and medical images.

The actual applications described in Grenander (1993) assume that the basic pose parameters, such as location and scale, are roughly known—namely, the detection process is *initialized* by the user. The models involve large numbers of generators with "elastic" types of constraints on their relative locations. Because deformation space—the space of bond values—is high dimensional, there is still much left to be done after location and scale are identified. The algorithms are primarily based on relaxation techniques for maximizing the posterior distributions. These types of elastic models

are described in chapters 3 and 5. Chapter 3 draws primarily on the work presented in Grenander, Chow, and Keenan (1991); Zhu and Yuille (1996); and Chesnaud, Réfrégier, and Boulet (1999). Chapter 5 draws on the work in Amit, Grenander, and Piccioni (1991) and Amit (1994), with some new unpublished material.

Some of these ideas were developed in parallel using nonstatistical formulations. In Kass, Witkin, and Terzopoulos (1987) and Terzopolous and colleagues (1987), the idea of 1D deformable contours was introduced, as well as ideas of elastic constraints on deformations, and Bajcsy and Kovacic (1988) introduced the idea of image deformation as an extension of older work on image sequence analysis by Horn and Schunck (1981) and Nagel (1983). In these models, a regularizing term takes the place of the prior, and the statistical model for the data takes the form of a cost function on the fit of the deformed model to the data.

In much of the above-mentioned work, the gray-level distributions are modeled directly. This can be problematic in achieving photometric invariance, invariance to variations in lighting, gray-scale maps, and so on. At the single pixel level, the distributions can be rather complex due to variable lighting conditions. Furthermore, the gray-level values have complex interactions requiring complex distributions in high-dimensional spaces. The options are then to use very simple models, which are computationally tractable but lacking photometric invariance, or to introduce complex models, which entail enormous computational cost.

An alternative is to transform the image data to variables that are photometric invariant—perhaps at the cost of reducing the information content of the data. However, it is then easier to formulate credible models for the transformed data. The deformable curve model in chapter 4 and the Bernoulli deformable image model in section 5.4 employ transforms of the image data into vectors of simple binary variables. One then models the distribution of the binary variables, given a particular deformation rather than the gray-level values. The material in chapter 4 draws primarily from the work in Petrocelli, Elion, and Manbeck (1992) and from Geman and Jedynak (1996).

All the algorithms mentioned above suffer from a similar drawback. Some form of initialization provided by the user is necessary. However, the introduction of binary features of varying degrees of complexity allows us to formulate simpler and sparser models with more-transparent constraints on the instantiations. Using these models, the initialization problem can be solved with no user intervention and in a very efficient way. Such models are discussed in chapters 6, 7, and 8, based on work in Amit, Geman, and Jedynak (1998), Amit and Geman (1999), and Amit (2000).

These ideas do fit within the theoretical pattern-analysis paradigm proposed in Grenander (1978). However, the emphasis on image data reduction does depart from

Grenander's philosophy, which emphasizes image synthesis and aims at constructing prior distributions and data models, which, if synthesized, would produce realistic images. This image-synthesis philosophy has also been adopted by people studying compositional models, as in Bienenstock, Geman, and Potter (1997) and Geman, Potter, and Chi (1998), and by people studying generative models, such as Mumford (1994), Revow, Williams, and Hinton (1996), and Zhu and Mumford (1997). Providing a comprehensive statistical model for the image ensemble is not only a very hard task, it is not at all clear that it is needed. There is a large degree of redundancy in the gray-level intensity maps recorded in an image, which may not be all that important for interpreting the symbolic contents of the image.

1.3 Detection of Rigid Objects

In the computer-vision community, the limitations of straightforward bottom-up segmentation also led to the introduction of object models that enter into the detection and recognition process. Most of the work has concentrated around rigid 3D objects (see Grimson 1990; Haralick and Shapiro 1992; Ullman 1996). These objects lend themselves to precise 3D modeling, and the main type of deformations considered are linear or projective.

Lists of features at locations on the object at reference pose are deduced analytically from the 3D description. The spatial arrangements of these features in the image are also predicted through analytic computations, using projective 3D geometry and local properties of edge detectors. Typical features that are used in modeling are oriented edges, straight contour segments—lines of various lengths, high curvature points, corners, and curved contours. Two complementary techniques for detection are searches of correspondence space and searches through pose space.

1.3.1 Searching Correspondence Space

One systematically searches for arrangements of local features in the image consistent with the arrangements of features in the model. The matches must satisfy certain constraints. Unary constraints involve the relationship between the model feature and the image feature. Binary constraints involve the relationship between a pair of model features and a pair of image features. Higher-order constraints can also be introduced. Various heuristic tree-based techniques are devised for searching all possible matchings to find the optimal one, as detailed in Grimson (1990). Invariance of the detection algorithm to pose is incorporated directly in the binary constraints.

In Haralick and Shapiro (1992), this problem is called the *inexact consistent labeling problem,* and various graph theory heuristics are employed.

Similar to the search of correspondence space, or the inexact consistent labeling problem, is the dynamic programming algorithm presented in chapter 7, which is based on work in Amit and Kong (1996) and Amit (1997). The constraints in the models are invariant to scale and to some degree of rotation, as well as *nonlinear* deformations. Detection is achieved under significant deformations of the model beyond simple linear or projective transformations. The full graph of constraints is pruned to make it decomposable, and hence amenable to optimization using dynamic programming, in a manner very similar to the proposal in Fischler and Elschlager (1973). The local features employed are highly invariant to photometric transformations but have a much lower density than typical edge features.

1.3.2 Searching Pose Space

Searching *pose space* can be done through brute force by applying each possible pose to the model and evaluating the fit to the data. This can be computationally expensive, but we will see in chapter 8 that brute force is useful and efficient as long as it is applied to *very simple structures,* and with the appropriate data models involving binary features with relatively low density in the image.

In some cases, searching parts of pose space can be achieved through optimization techniques such as gradient-descent methods or dynamic programming. This is precisely the nature of the deformable models presented in chapters 3–5. Note, however, that here objects are not assumed rigid and hence require many more pose parameters. These methods all face the issue of initialization.

A computational tool that repeatedly comes up as a way to quickly identify the most important parameters of pose, such as location and scale, is the Hough transform, originally proposed by Hough (1962) and subsequently generalized by Ballard (1981). The Hough transform is effectively also a "brute force" search over all pose space. Because the structures are very simple, the search can be efficiently implemented. The outcome of this computation provides an initialization to the correspondence space search or a more refined pose space search (see Grimson 1990 and Ullman 1996) or, in our case, the more complex deformable template models. In Grimson (1990), a careful analysis of the combinatorics of the Hough transform is carried out in terms of the statistics of the local features. A very appealing and efficient alternative to the Hough transform has recently been proposed in Fleuret and Geman (2001), where a coarse-to-fine *cascade* of detectors is constructed for a treelike decomposition of pose space into finer and finer bins.

The Hough transform, as a method of jump-starting more intensive algorithms, is intuitively very appealing but did not take off as a dominant paradigm in computer vision partly because of the combinatoric problems analyzed in Grimson (1990). Testifying to this is the fact that a significant body of work in the same community did not use this approach for face detection (see, for example, Rowley, Baluja, and Kanade 1998; Sung and Poggio 1998). One reason may be the use of predesigned local features. In chapter 6, we introduce a hierarchy of local-edge arrangements of increasing complexity. Despite being more complex than simple edges, these local features are still very stable on object and quite rare in the background. The features in the model are obtained through training and do not necessarily have a clear semantic interpretation. Sparse object models are then defined as flexible arrangements of a small number of these local features. The construction and training of sparse object models in terms of these local features, and the statistical properties of these features on object and on background are also described in chapter 6.

In chapter 8, an efficient algorithm for detecting such models is presented, where the first step of identifying candidate locations is obtained using the Hough transform. This material is based on work in Amit, Geman, and Jedynak (1998), Amit and Geman (1999), and Amit (2000). The work in Burl, Leung, and Perona (1995) and Burl, Weber, and Perona (1998), is very similar in spirit; however, the features and the statistical models are more complex, and the computation of the detection more intensive.

The dominant view in the computer-vision community is that some form of bottom-up processing involving segmentation is still necessary to jump-start the detection and recognition tasks. In Ullman (1996), a case for this is made in terms of biological processes in the visual system. The point of view put forward here is that one can go a long way with a combination of model-driven detections followed by more-refined processing involving classification and obtaining more-detailed instantiations. This is one of the main conclusions of chapter 6, where we study the statistics of the particular local features employed in the sparse models; of chapter 8, where we implement a version of the Hough transform for initial detection of candidate locations; and of chapter 10, where some ideas on combining object detection and recognition for complex scene analysis are explored.

1.3.3 Rigid versus Nonrigid Objects

Much of the work on object detection is centered around rigid objects. This has led, for example, to detailed analysis of the specific pose space associated with 2D and 3D rigid transformations and their projections (see, for example, Arbter and colleagues 1990, Mundy and Zisserman 1992). There is also an emphasis on complete planar

rotation invariance. The rigid nature of the objects has lead to reliance on "predefined features" with labels such as lines, corners, junctions and so on. In recent years, a view-based approach has become widely accepted in which 3D object detection and recognition are treated as 2D problems depending on the particular views of the objects (see Ullman 1996; Riesenhuber and Poggio 2000).

However, even for 2D views of rigid objects, lines and contours or even corners can be ambiguous in the image domain. Moreover, the visual system can detect and recognize rigid objects even if many of the straight lines present on the real object are deformed in the image. The message of chapters 8 and 10 is that all objects should be studied within one framework, based on 2D views, using *nonrigid* 2D models. Views of the object that are substantially different are considered as *different* 2D objects; however, the flexibility (i.e., geometric invariance) introduced in the nonrigid models implies that a *wide range* of views can still be accommodated by one model. This alleviates to some extent the combinatoric problem of the resulting proliferation of 2D objects that need to be modeled, detected, and recognized. Some additional ideas related to this problem are presented in chapter 10.

1.4 Object Recognition

Recognition of isolated objects has been studied extensively in two main contexts: rigid 2D and 3D objects and character recognition. The latter context offers an important test bed for many ideas. Recent extensive reviews can be found in Plamondon and Srihari (2000) and Nagy (2000). The data sets are abundant, different forms of variability are present—rigid for printed characters and nonrigid for handwritten, and one can work with a limited number of classes, say, only the digits, or with large numbers such as all LaTeX symbols, or Chinese characters.

1.4.1 Deformable-Template Models for Object Recognition

The problem of recognizing an image of an isolated object from among several possible classes can be addressed in a Bayesian framework using the deformable-template models. These models have a natural extension to a statistical model for images of the different object classes, once a prior on object classes is determined. The goal is then to compute the Bayes classifier—namely, the class that maximizes the posterior on class, given the data. The deformation parameters are no longer of direct interest but need to be integrated out in order to obtain the posterior on

class. This type of computation is very expensive, so that in real applications, one deformation is estimated from *each* of the class templates to best fit the data, and classification is then based on various metrics defined on these deformations. Such a procedure is spelled out in detail in Hastie and Simard (1998). Despite the fact that the distance of the data to each template is computed modulo the deformation, this approach still requires quite careful preprocessing and registration of the images to a standard size. The underlying assumption is that the deformations are small. It also requires explicit modeling of the prototype images, extensive computation at the classification stage, and appears impractical with large numbers of shape classes. A deformable-template-based approach to face recognition is presented in Wiskott and colleagues (1997), although not based on a statistical model. There, the data model is not based directly on the pixel intensities but on local features derived from Gabor filters extracted at multiple scales.

In chapter 9, we present an alternative based on elements of the sparse-detection models. The main tool will be binary classification trees (see Breiman and colleagues 1984), where the splits are defined in terms of flexible arrangements of local features of the same nature as those defining the sparse models. Trees provide a natural mechanism for exploring arrangements of increasing complexity.

Instead of modeling the posterior distribution on deformations and on class, and then computing the posterior on-line, the trees yield partial posteriors conditional on a smaller number of variables, which are obtained off-line during training. Computation during classification is then very fast. An essential element of our approach is to produce multiple randomized classification trees. Individually, the error rates of these trees can be quite large, but when aggregated, a very powerful classifier emerges. The work in chapter 9 is based on Amit and Geman (1997) and Amit, Geman, and Wilder (1997).

1.4.2 Normalization and Registration

In the literature on statistical pattern recognition, it is common to address geometric and photometric variations by preprocessing and normalization. A "standardized" image is produced prior to classification, involving a sequence of operations that brings all images to the same size and then corrects for translation, slant, and rotation. This is not done using some template or model, because the class of the image is unknown. Classification is then performed by one of the standard pattern recognition procedures, based on the gray-level intensities of the standardized image. (For example, penalized discriminant analysis in Hastie, Buja, and Tibshirani 1995, or multilayer neural networks in Bottou and colleagues 1994, or classification trees in Ho, Hull,

and Srihari 1994.) Difficulties in generalization are often encountered because the normalization is not robust and does not accommodate nonlinear deformations. This deficiency can only be ameliorated with very large training sets.

An alternative is to define a collection of binary features extracted from the data, such as edges, contours, endings junctions, and so on. The feature/location pairs are collected to make a fixed-size feature vector, which is fed into a standard type of classifier. These features may be designed to be more invariant to geometric deformations than the raw gray-level values, using explicit disjunction (or-ing). Otherwise put, a feature detected at a particular location is "spread" to an entire neighborhood. The features are designed to be photometric invariant, so that no gray-level normalization is required. These matters are investigated in chapter 9.

1.4.3 Geometric Invariants

Another approach, which has been explored in the computer-vision literature, is to search for functions invariant to a family of transformations, such as the affine transformations. Discrimination is possible if the functions have different values for different classes (see, for example, Lamdan, Schwartz, and Wolfson 1988; Forsyth and colleagues 1991; Mundy and Zisserman 1992; Binford and Levitt 1993; Reiss 1993). The explicit introduction of geometric invariance is very appealing and provides an element that is missing in the standard pattern-recognition approaches. The problem, however, is that the invariant functions are defined in terms of precisely located distinguished points on the object. This is not very practical on real gray-level images, or for objects that are deformed nonlinearly. This brings us back to the discussion above regarding rigid versus nonrigid objects. Just as in the case of detection, it is useful to consider shape classification for both categories as one problem. Focusing on invariants associated strictly with rigidity can lead to unstable algorithms. Hence the introduction of looser types of functions—flexible arrangements of local features, which are also explored in chapter 9. With these functions, however, full rotation invariance is lost.

1.5 Scene Analysis: Merging Detection and Recognition

The grand goal of computer vision is to enable the computer to detect and recognize multiple objects in a visual scene. We are still very far from achieving this goal. This is not only a function of computational limitations, it is also a result of the lack of

a dominant paradigm agreed upon by most of the research community. As indicated earlier, one paradigm assumes bottom-up image segmentation as a precursor to any high-level processing. At the other extreme are compositional and generative models (see Mumford 1994; Hinton and colleagues 1995; Bienenstock, Geman, and Potter 1997; Geman, Potter, and Chi 1998), where people attempt to provide a comprehensive statistical model of entire scenes, both from the point of view of the components generating the scene—namely, priors on complex scenes—and complex data models of the images, given a particular configuration of objects. The view is that local ambiguities can be resolved only in the framework of a comprehensive explanation of the data. These models appear diametrically opposed to the segmentation models. The only way to unambiguously determine the boundary of an object is by identifying the object, its pose, all the objects in its neighborhood, and their respective positions. Conceptually, these models are appealing in their attempt to pose the scene analysis problem in a comprehensive Bayesian framework. Regrettably, they are extremely challenging on all levels: formulating the prior models, the data models, estimating the relevant parameters, and ultimately computing the optimal interpretation given the image data.

Chapter 10 is an initial exploration into a possible middle path between these two extremes, based on a combination of the sparse-detection models described in chapter 8 and the recognition algorithms of chapter 9. Local features in the image are not grouped in a bottom-up manner as in standard segmentation; rather, the grouping is an outcome of the detection of a particular model, and comes with an estimated pose and some additional instantiation parameters. These model-driven groupings of local features can be viewed as elements of a compositional model. Recent implementations of compositional models have used very gradual compositions, from edgelets to lines or curves to small combinations of these, and so on. The compositions proposed here are very coarse and there is a direct jump from the local to the global model.

If the detection models are created to be less specific, either by directly training on a collection of classes or by training on one class and then using lower thresholds, they define object clusters, as opposed to being dedicated to one particular class. This means that classification must follow detection. Detecting instances of several coarse models and subsequently classifying them is a very efficient way to obtain a relabeling of the image into detections (involving some pose parameters) and class labels. This labeling in no way provides a final scene interpretation. There could be multiple labels at the same location, overlapping detections, and so on. From the point of view of the compositional and generative models, this can be taken as a crude first pass, which provides the higher-level models with multiple possible scene interpretations for evaluation. In chapter 10, we discuss possible strategies for generating this basic

map of labeled detections. How to then analyze this information and produce coherent scene interpretations is beyond the scope of this book.

1.6 Neural Network Architectures

There have been a number of attempts to formulate parallel network architectures for higher-level vision tasks such as detection. Some examples are the work in Fukushima (1986) and Fukushima and Wake (1991), Olshausen, Anderson, and Van Essen (1993), and recent models, such as Riesenhuber and Poggio (1999). Each of these models touches upon certain important aspects of the problem. In Fukushima and Wake (1991), hard wiring of invariance is achieved through or-ing or *spreading,* which is an important component of the algorithms described in this book. However, the proposed network depends too heavily on a long sequence of processing layers and on learning more and more complex features. In Olshausen, Anderson, and Van Essen (1993), mechanisms for shifting data from the periphery to the center for further processing are studied, but the training of classifiers or implementing object detection as a component of visual selection are not discussed. In Riesenhuber and Poggio (1999), invariant recognition is achieved through a combination of or-ing as in Fukushima and Wake (1991), generalized to continuous variables through a MAX (maximization) operation, and predefined pair-wise conjunctions of features. The MAX operation is taken to an extreme where all information on the relative locations of features is lost at the highest stage. This is problematic when dealing with even simple data sets such as the NIST (National Institute of Standards and Technology) handwritten character data set. Moreover, this approach cannot produce accurate location information in a detection problem. In chapter 11, we explore how object representations and classifiers, trained using the principles of Hebbian learning, in a central memory module, are able to drive visual selection over the entire scene and at a wide range of scales, as well as classify isolated objects or those present at a selected location. The material ties together the work in Amit (2000) and Amit and Mascaro (2001) into one comprehensive system.

Bibliography

Abbott, L. F. and Nelson, S. B. (2000). Synaptic plasticity: taming the beast. *Nat. Neurosci., 3* (suppl), 1178–1183.

Amit, D. J. (1989). *Modelling brain function: The world of attractor neural networks*. Cambridge: Cambridge University Press.

Amit, D. J. and Brunel, N. (1995). Learning internal representations in an attractor neural network with analogue neurons. *Network, 6,* 261.

Amit, D. J. and Brunel, N. (1997). Model of global spontaneous activity and local structured (learned) delay activity during delay periods in cerebral cortex. *Cereb. Cortex, 7,* 237–252.

Amit, D. J. and Fusi, S. (1994). Dynamic learning in neural networks with material synapses. *Neural Computation, 6,* 957.

Amit, Y. (1994). A non-linear variational problem for image matching. *SIAM J. Sci. Computing, 15*(1), 207–224.

Amit, Y. (1997). Graphical shape templates for automatic anatomy detection: application to MRI brain scans. *IEEE Trans. Med. Imaging, 16,* 28–40.

Amit, Y. (2000). A neural network architecture for visual selection. *Neural Computation, 12,* 1059–1082.

Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation, 9,* 1545–1588.

Amit, Y. and Geman, D. (1999). A computational model for visual selection. *Neural Computation, 11,* 1691–1715.

Amit, Y. and Kong, A. (1996). Graphical templates for model registration. *IEEE Pattern Anal. Machine Intell., 18,* 225–236.

Amit, Y. and Mascaro, M. (2001). Attractor networks for shape recognition. *Neural Computation, 13,* 1415–1442.

Amit, Y. and Blanchard, G. (2001). Multiple randomized classifiers: MRCL. Technical report, Dept. of Statistics, University of Chicago.

Amit, Y., Geman, D., and Jedynak, B. (1998). Efficient focusing and face detection. In H. Wechsler and J. Phillips, eds., *Face recognition: From theory to applications,* NATO ASI Series F. Berlin: Springer-Verlag.

Amit, Y., Geman, D., and Wilder, K. (1997). Joint induction of shape features and tree classifiers. *IEEE Trans. Pattern Anal. Machine Intell., 19,* 1300–1306.

Amit, Y., Grenander, U., and Piccioni, M. (1991). Structural image restoration through deformable template. *J. Am. Stat. Assoc., 86*(414), 376–387.

Arbter, K., Snyder, W. E., Burkhardt, H., and Hirzinger, G. (1990). Application of affine invariant fourier descriptors to recognition of 3D objects. *IEEE Trans. Pattern Anal. Machine Intell., 12,* 640–647.

Atkinson, R. (1975). Mnemotechnics in second-language learning. *Am. Psychol., 30,* 821–828.

Bajcsy, R. and Kovacic, S. (1988). Multiresolution elastic matching. *Comput. Vis. Graphics Image Processing, 46,* 1–21.

Ballard, D. H. (1981). Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognit., 13,* 111–122.

Bartlett, M. S. and Sejnowski, T. J. (1998). Learning viewpoint-invariant face representations from visual experience in an attractor network. *Network: Comput. Neural Syst., 9,* 399–417.

Bertele, U. and Brioschi, F. (1969). A new algorithm for the solution of the secondary optimization problem in non-serial dynamic programming. *J. Math. Anal. Appl., 27,* 565–57.

Biederman, I. (1995). Visual object recognition. In S. M. Kosslyn and D. N. Osherson, eds., *Visual cognition.* Cambridge: MIT Press, pp. 121–166.

Bienenstock, E., Geman, S., and Potter, D. (1997). Compositionality, MDL priors, and object recognition. In M. C. Mozer, M. I. Jordan, and T. Petsche, eds., *Advances in neural information and processing systems,* vol. 9. Cambridge: MIT Press, pp. 834–844.

Binford, T. O. and Levitt, T. S. (1993). Quasi-invariants: theory and exploitation. In *Proc. Image Understanding Workshop*. Washington, D.C.: pp. 819–828.

Bishop, C. M. (1995). *Neural networks for pattern recognition.* New York: Oxford University Press.

Blake, A. and Issard, M. (1998). *Active Contours*. New York: Springer-Verlag.

Blake, A. and Yuille, A. (1992). *Active Vision*. Cambridge: MIT Press.

Bliss, T. V. P. and Collingridge, G. L. (1993). A synaptic model of memory: long term potentiation in the hippocampus. *Nature, 361,* 31.

Bookstein, L. F. (1991). *Morphometric tools for landmark data: Geometry and biology*. Cambridge: Cambridge University Press.

Bottou, L., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., Jackel, L. D., LeCun, Y., Muller, U. A., Sackinger, E., Simard, P., and Vapnik, V. (1994). Comparison of classifier methods: a case study in handwritten digit recognition. In *Proceedings of the IEEE international conference on pattern recognition,* pp. 77–82.

Breiman, L. (1994). Bagging predictors. Technical report 451, Department of Statistics, University of California, Berkeley.

Breiman, L. (1998). Arcing classifiers (with discussion). *Ann. Stat., 26,* 801–849.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and regression trees*. Belmont, Calif.: Wadsworth.

Brooks, R. A. (1981). Symbolic reasoning among 3D models and 2D images. *Artif. Intell., 17,* 285–348.

Brunel, N., Carusi, F., and Fusi, S. (1998). Slow stochastic Hebbian learning of classes of stimuli in a recurrent neural network. *Network, 9,* 123–152.

Burl, M. C., Leung, T. K., and Perona, P. (1995). Face localization via shape statistics. In M. Bichsel, ed., *Proceedings of the international workshop on automatic face and gesture recognition*, pp. 154–159.

Burl, M., Weber, M., and Perona, P. (1998). A probabilistic approach to object recognition using local photometry and global geometry. In *Proceedings of the fifth European Conference on Computer Vision, ECCV '98*.

Camion, V. and Younes, L. (2001). Geodesic interpolating splines. In *EEMVCPR*.

Caselles, V., Kimmel, R., and Sapiro, G. (1997). Geodesic active contours. *Int. J. Comput. Vis., 22,* 61–79.

Caselles, V., Kimmel, R., Sapiro, G., and Sbert, C. (1997). Minimal surfaces based object segmentation. *IEEE Trans. Pattern Anal. Machine Intell., 19,* 394–398.

Chalidabhongse, J. and Kuo, C.-C. J. (1997). Fast motion vector estimation using multiresolution-spatio-temporal correlations. *IEEE Trans. Circuits Syst. Video Technol., 7,* 477–488.

Chelazzi, L., Miller, E. K., Duncan, J., and Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex. *Nature, 363,* 345–347.

Chesnaud, C., Réfrégier, P., and Boulet, V. (1999). Statistical region snake-based segmentation adapted to different physical noise models. *IEEE Trans. Pattern Anal. Machine Intell., 21,* 1145–1157.

Christensen, G., Rabbitt, R. D., and Miller, M. I. (1996). Deformable templates using large deformation kinematics. *IEEE Trans. Image Processing, 5,* 1435–1447.

Chuang, G. and Kuo, C. (1996). Wavelet description of planar curves: theory and applications. *IEEE Trans. Image Processing, 5,* 56–70.

Cohen, I. and Cohen, L. D. (1993). Finite element methods for active contour models and balloons for 2D and 3D images. *IEEE Trans. Pattern Anal. Machine Intell., 15,* 1131–1147.

Cohen, I., Cohen, L. D., and Ayache, N. (1992). Using deformable surfaces to segment 3-D images and infer differential structures. *CVGIP: Image Understanding, 56*(2), 242–263.

Cohen, L. D. (1991). On active contour models and balloons. *CVGIP: Image Understanding, 53,* 211–218.

Cootes, T. F. and Taylor, C. J. (1992). Active shape models—smart snakes. In *Proceedings of BMVC.* pp. 267–275.

Cootes, T. F. and Taylor, C. J. (1996). Locating faces using statistical feature detectors. In *Proceedings of the second international workshop on automatic face and gesture recognition*. IEEE Computer Society Press, pp. 204–210.

Cover, T. M. and Thomas, J. A. (1991). *Elements of information theory.* New York: John Wiley.

Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math., 41,* 909–996.

Desimone, R. and Dyuncan, J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci., 18,* 193–222.

Desimone, R., Miller, E. K., Chelazzi, L., and Lueschow, A. (1995). Multiple memory systems in visual cortex. In M. S. Gazzaniga, ed., *The cognitive Neurosciences*. Cambridge: MIT Press, pp. 475–486.

Diederich, S. and Opper, M. (1987). Learning correlated patterns in spin-glass networks by local learning rules. *Phys. Rev. Lett., 58,* 949–952.

Dryden, I. L. and Mardia, K. V. (1998). *Statistical shape analysis.* New York: Wiley.

Duda, R. O. and Hart, P. E. (1973). *Pattern classification and scene analysis.* New York: John Wiley.

Durtewitz, D., Seamans, J. K., and Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nat. Neurosci., 3* (suppl), 1192–1198.

Elder, J. and Zucker, S. W. (1996). Computing contour closure. In *Computer vision—ECCV*. New York: Springer, pp. 399–412.

Figueiredo, M. and Leitao, J. (1992). Bayesian estimation of ventricular contours in angiographic images. *11,* 416–429.

Fischler, M. A. and Elschlager, R. A. (1973). The representation and matching of pictorial structures. *IEEE Trans. Comput., 22,* 67–92.

Fleuret, F. (2000). Détection hiérarchique de visages par apprentissage statistique. Ph.D. thesis, L'úniversite Paris 6.

Fleuret, F. and Geman, D. (2001). Coarse-to-fine visual selection. *Int. J. Comput. Vis., 41,* 85–107.

Forsyth, D., Mundy, J. L., Zisserman, A., Coelho, C., Heller, A., and Rothwell, C. (1991). Invariant descriptors for 3-D object recognition and pose. *IEEE Trans. Pattern Anal. Machine Intell., 13,* 971–991.

Freund, Y. and Shapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci., 55,* 119–139.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *Ann. Stat., 28,* 337–374.

Friston, K., Ashburner, J., Frith, C. D., Poline, J.-B., Heather, J. D., and Frackowiak, R. (1995). Spatial registration and normalization of images. *Hum. Brain Mapping, 3,* 165–189.

Fukushima, K. (1986). A neural network model for selective attention in visual pattern recognition. *Biol. Cybern., 55,* 5–15.

Fukushima, K. and Wake, N. (1991). Handwritten alphanumeric character recognition by the neocognitron. *IEEE Trans. Neural Netw*.

Fusi, S., Annunziato, M., Badoni, D., Salamon, A., and Amit, D. J. (2000). Spike-driven synaptic plasticity: theory, simulation, VLSI implementation. *Neural Comput., 12,* 2227.

Garris, M. D. and Wilkinson, R. A. (1996). *NIST special database 3. Handwritten segmented characters*. Gaithersburg, Md: NIST.

Geiger, D., Gupta, A., Costa, L., and Vlontzos, J. (1995). Dynamic programming for detecting, tracking, and matching defomrable contours. *17,* 294–302.

Geman, D. (1990). Random fields and inverse problems in imaging. In *Lecture notes in mathematics,* no. 1427. Springer Verlag.

Geman, D. and Jedynak, B. (1996). An active testing model for tracking roads from satellite images. *IEEE Trans. Pattern Anal. Machine Intell., 18,* 1–15.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell., 6,* 721–741.

Geman, S., Potter, D. F., and Chi, Z. (1998). Composition systems. Technical report, Division of Applied Mathematics, Brown University.

Gersho, A. and Gray, R. M. (1992). *Vector quantization and signal compression.* Boston: Kluwer Academic.

Gold, S. and Rangarajan, A. (1996). A graduated assignment algorithm for graph matching. *IEEE Trans. Pattern Anal. Machine Intell., 18,* 377–388.

Grenander, U. (1970). A unified approach to pattern analysis. *Adv. Comput., 10,* 175–216.

Grenander, U. (1978). *Pattern analysis: Lectures in pattern theory I–III.* New York: Springer-Verlag.

Grenander, U. (1993). *General Pattern Theory.* Oxford: Oxford University Press.

Grenander, U. and Miller, I. M. (1998). Computational anatomy: an emerging discipline. *Q. Appl. Math., LVI*(4), 617–694.

Grenander, U., Chow, Y., and Keenan, D. (1991). *A pattern theoretical study of biological shape.* New York: Springer Verlag.

Grimson, W. E. L. (1990). *Object recognition by computer: The role of geometric constraints.* Cambridge: MIT Press.

Hallinan, P. L., Gordon, G., Yuille, A. L., Giblin, P., and Mumford, D. (1999). *Two- and three-dimensional patterns of the face.* Natick, Mass.: A. K. Peters.

Haralick, R. M. and Shapiro, G. L. (1992). *Computer and robot vision,* vols. 1–2. Reading, Mass.: Addison Wesley.

Hastie, T. and Simard, P. Y. (1998). Metrics and models for handwritten character recognition. *Stat. Sci.,*

Hastie, T., Buja, A., and Tibshirani, R. (1995). Penalized discriminant analysis. *Ann. Stat., 23,* 73–103.

Hebb, D. O. (1949). *The organization of behavior.* New York: Wiley.

Hedgé, J. and Van Essen, D. C. (2000). Selectivity for complex shapes in primate visual area v2. *J. Neurosci.,*

Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. (1995). The wake-sleep algorithm for unsupervised neural networks. *Science, 268,* 1158–1161.

Ho, T. K., Hull, J. J., and Srihari, S. N. (1994). Decision combination in multiple classifier systems. *IEEE Trans. Pattern Anal. Machine Intell., 16,* 66–75.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent selective computational abilities. *Proc. Natl. Acad. Sci. USA., 79,* 2554–2558.

Horn, B. K. P. and Schunck, B. G. (1981). Determining optical flow. *Artif. Intell., 17,* 185–203.

Hough, P. V. C. (1962). Methods and means for recognizing complex patterns. *U.S. Patent, 3069654*.

Huang, T. S. and Tsai, R. Y. (1981). Image sequence analysis: Motion estimation. In T. S. Huang, ed., *Image sequence analysis*. New York: Springer-Verlag.

Hubel, H. D. (1988). *Eye, brain, and vision*. New York: Scientific American Library.

Ishikawa, H. and Geiger, D. (1998). Segmentation by grouping junctions. In *Proceedings of the IEEE computer vision and pattern recognition*.

Ishikawa, H. and Geiger, D. (1999). Mapping image restoration to a graph problem. In *Proceedings of the IEEE-EURASIP workshop on non-linear and signal and image processing*.

Jermyn, I. and Ishikawa, H. (1999). Globally optimal regions and boundaries. In *Proceedings of the seventh IEEE international conference on computer vision (ICCV '99).*

Joshi, S. (1997). Large deformation diffeomorphisms and Gaussian random fields for statistical characterization of brain submanifolds. Ph.D. thesis, Department of Electrical Engineering, Washington University.

Kass, M., Witkin, A., and Terzopoulos, D. (1987). Snakes: active contour models. *Int. J. Comput. Vis.,* 321–331.

Kim, B., Boes, J. L., Frey, K. A., and Meyer, C. R. (1997). Mutual information for automated unwarping of rat brain autoradiographs. *NeuroImage, 5,* 31–40.

Kohonen, T. (1984). *Self-organization and associative memory.* Berlin: Springer Verlag.

Kwok, S. W. and Carter, C. (1990). Multiple decision trees. In R. D. Shachter, T. S. Levitt, L. Kanal, and J. F. Lemmer, eds., *Uncertainty and artificial intelligence.* North-Holland, Amsterdam: Elsevier Science Publishers.

Lamdan, Y., Schwartz, J. T., and Wolfson, H. J. (1988). Object recognition by affine invariant matching. In *IEEE international conference on computer vision and pattern recognition.* pp. 335–344.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE, 86*(11), 2278–2324.

Levenex, P. and Schenk, F. (1997). Olfactory cues potentiate learning of distant visuospatial information. *Neurobiol. Learn. Mem., 68,* 140–153.

Malik, J. and Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms. *J. Opti. Soc. Am. A, 7,* 923–932.

Malladi, R., Sethian, J. A., and Vemuri, B. C. (1995). Shape modeling with front propagation. *IEEE Trans. Pattern Anal. Machine Intell., 17,* 158–176.

Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Machine Intell.,* 674–693.

Markram, H., Lubke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic ap's and epsp's. *Science, 375,* 213.

Marr, D. (1982). *Vision.* W. H. New York: Freeman and Company.

Marr, D. and Hilderith, E. (1980). Theory of edge detection. *Proc. R. Soc. Lond. B Biol. Sci., 207,* 187–217.

Marr, D. and Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond. B Biol. Sci., 200,* 269–294.

Mascaro, M. and Amit, D. J. (1999). Effective neural response function for collective population states. *Network, 10,* 351–373.

Mattia, M. and Del Giudice, P. (1999). Asynchronous simulation of large networks of spiking neurons and dynamical synapses. Submitted for publication to *Neural Computation*.

Meyer, Y. (1990). *Ondelettes et operateurs.* Paris: Herman.

Miller, M., Christensen, G., Amit, Y., and Grenander, U. (1993). A mathematical textbook of deformable neuro-anatomies. *Proc. Nat. Acad. Sci., R90,* 11944–11948.

Minsky, M. and Papert, S. (1969). *Perceptrons.* Cambridge: MIT Press.

Mumford, D. (1994). Pattern theory: a unifying perspective. In *First European congress of mathematics, vol 1.* Birkhausër, pp. 187–224.

Mundy, J. L. and Zisserman, A. (1992). *Geometric invariance in computer vision.* Cambridge: MIT Press.

Nagel, H. H. (1983). Displacement vectors derived from second-order intensity variations in image sequences. *Comput. Vis. Graph. Image Processing, 21,* 85–117.

Nagy, G. (2000). Twenty years of document analysis in PAMI. *IEEE Trans. Pattern Anal. Machine Intell., 22,* 38–62.

Oja, E. (1989). Neural networks, principle components, and subspaces. *Int. J. Neural Syst., 1,* 62–68.

Olshausen, B. A., Anderson, C. H., and Van Essen, D. C. V. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci., 13,* 4700–4719.

Parida, L., Geiger, D., and Hummel, R. (1998). Junctions: detection, classification, and reconstruction. *IEEE Trans. Pattern Anal. Machine Intell., 20,* 687–698.

Petrocelli, R. R., Elion, J. L., and Manbeck, K. M. (1992). A new method for structure recognition in unsubtracted digital angiograms. In *Proceedings of computers in cardiology*. IEEE Computer Society, pp. 207–210.

Plamondon, R. and Srihari, S. N. (2000). On-line and off-line handwritten recognition. *IEEE Trans. Pattern Anal. Machine Intell., 22,* 63–84.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. (1995). *Numerical recipes in C, The art of scientific computing,* 2nd ed. Cambridge: Cambridge University Press.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learn., 1,* 81–106.

Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of speech recognition.* Englewood Cliffs, N.J.: Prentice Hall.

Rangarajan, A., Chui, H., and Bookstein, F. (1997). The softassign procrustes matching algorithm. In J. Duncan and G. Gindi, eds., *Information processing in medical imaging.* Springer, pp. 29–42.

Reiss, T. H. (1993). Recognizing planar objects using invariant image features. In *Lecture notes in computer science,* no. 676. Berlin: Springer Verlag.

Revow, M., Williams, C. K. I., and Hinton, G. E. (1996). Using generative models for handwritten digit recognition. *IEEE Trans. Pattern Anal. Machine Intell., 18,* 592–606.

Rice, J. A. (1995). *Mathematical statistics and data analysis,* 2nd ed. Belmont, Calif.: Duxbury Press.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci., 2,* 1019–1025.

Riesenhuber, M. and Poggio, T. (2000). Models of object recognition. *Nat. Neurosci., 3* (suppl), 1199–1204.

Ripley, B. D. (1994). Neural networks and related methods for classification. *J. R. Stat. Soc. B, 56,* 409–437.

Rojer, A. S. and Schwartz, E. L. (1992). A quotient space Hough transform for scpae-variant visual attention. In G. A. Carpenter and S. Grossberg, eds., *Neural networks for vision and image processing*. Cambridge: MIT Press.

Rolls, E. T. (2000). Memory systems in the brain. *Annu. Rev. Psychol., 51,* 599–630.

Rose, D. J., Tarjan, R. E., and Leuker, G. S. (1976). Algorithmic aspects of vertex elimination on graphs. *Siam J. Comput.,* pp. 266–283.

Rowley, H. A., Baluja, S., and Kanade, T. (1998). Neural network-based face detection. *IEEE Trans. Pattern Anal. Machine Intell., 20,* 23–38.

Sandor, S. E. and Leahy, R. M. (1995). Towards automated labelling of the cerebral cortex using a deformable atlas. In Y. E. A. Bizais, ed., *Information processing in medical imaging*. Netherlands: Kluwer Academic Press, pp. 127–138.

Schapire, R. E., Freund, Y., Bartlett, P., and Lee, W. S. (1998). Boosting the margins: a new explanation for the effectiveness of voting methods. *Ann. Stat., 26*(5), 1651–1686.

Shapiro, L. G. (1980). A structural model of shape. *IEEE Trans. Pattern Anal. Machine Intell., 2,* 111–126.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Machine Intell., 22,* 888–905.

Simard, P. Y., LeCun, Y., Denker, J. S., and Victorri, B. (2000). Transformation invariance in pattern recognition—tangent distance and tangent propagation. *Int. J. Imaging Syst. Technol., 11,* 181–197.

Sung, K. K. and Poggio, T. (1998). Example-based learning for view-based face detection. *IEEE Trans. Pattern Anal. Machine Intell., 20,* 39–51.

Tanaka, K., Saito, H. A., Fukada, Y., and Moriya, M. (1991). Coding visual images of objects and the inferotemporal cortex of the macaque monkey. *J. Neurosci., 66*(1), 170–189.

Tarr, M. and Bülthoff, H. (1998). Image-based object recognition in man, monkey, and machine. *Cognition, 67,* 1–20.

Terzopolous, D., Platt, J., Barr, A., and Fleisher, K. (1987). Elastically deformable models. *Comput. Graph., 21,* 205–214.

Tovee, M. J. (1996). *An introduction to the visual system.* Cambridge: Cambridge University Press.

Trouvé, A. (1998). Diffeomorphism groups and pattern matching in image analysis. *Int. J. Comput. Vis., 28,* 213–221.

Ullman, S. (1996). *High-level vision.* Cambridge: MIT Press.

Van Rullen, R., Gautrais, J., Delorme, A., and Thorpe, S. (1998). Face processing using one spike per neuron. *Biosystems, 48,* 229–239.

Vapnik, V. N. (1995). *The nature of statistical learning theory.* New York: Springer Verlag.

Viola, P. and Jones, M. J. (2001). Robust real time object detection. to appear in *Int. J. Comput. Vis.*

Viola, P. and Wells, W. M. I. (1997). Alignment by maximization of mutual information. *Int. J. Comput. Vis., 24,* 137–154.

von der Heydt, R. (1995). Form analysis in visual cortex. In M. S. Gazzaniga, ed., *The cognitive neurosciences*. Cambridge: MIT Press, pp. 365–382.

Wang, S. C. (1998). A statistical model for computer recognition of sequences of handwritten digits, with applications to zip codes. Ph.D. thesis, Department of Statistics, University of Chicago.

Wang, Y. and Staib, L. H. (2000). Boundary finding with prior shapes and smoothness models. *IEEE Trans. Pattern Anal. Machine Intell., 22,* 738–743.

Wickerhauser, M. V. (1994). *Adapted wavelet analysis from theory to software.* Wellesley, Mass.: IEEE Press.

Wiskott, L., Fellous, J.-M., Kruger, N., and von der Marlsburg, C. (1997). Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern. Anal. Machine Intell., 7,* 775–779.

Zeki, S. (1993). *A Vision of the brain.* Oxford: Blackwell Scientific Publications.

Zhu, S. and Yuille, A. (1996). Region competition: unifying snakes, region growing, energy/ Bayes/MDL for multi-band image segmentation. *IEEE Trans. Pattern Anal. Machine Intell., 18,* 884–900.

Zhu, S. C. and Mumford, D. (1997). Prior learning and Gibbs reaction-diffusion. *IEEE Trans. Pattern Anal. Machine Intell., 19,* 1236–1250.