

MACHINE LEARNING IN COMPUTER VISION

Computational Imaging and Vision

Managing Editor

MAX A. VIERGEVER

Utrecht University, Utrecht, The Netherlands

Editorial Board

RUZENA BAJCSY, *University of Pennsylvania, Philadelphia, USA*

MIKE BRADY, *Oxford University, Oxford, UK*

OLIVIER D. FAUGERAS, *INRIA, Sophia-Antipolis, France*

JAN J. KOENDERINK, *Utrecht University, Utrecht, The Netherlands*

STEPHEN M. PIZER, *University of North Carolina, Chapel Hill, USA*

SABURO TSUJI, *Wakayama University, Wakayama, Japan*

STEVEN W. ZUCKER, *McGill University, Montreal, Canada*

Machine Learning in Computer Vision

by

N. SEBE

*University of Amsterdam,
The Netherlands*

IRA COHEN

HP Research Labs, U.S.A.

ASHUTOSH GARG

Google Inc., U.S.A.

and

THOMAS S. HUANG

*University of Illinois at Urbana-Champaign,
Urbana, IL, U.S.A.*



Springer

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

ISBN-10 1-4020-3274-9 (HB) Springer Dordrecht, Berlin, Heidelberg, New York
ISBN-10 1-4020-3275-7 (e-book) Springer Dordrecht, Berlin, Heidelberg, New York
ISBN-13 978-1-4020-3274-5 (HB) Springer Dordrecht, Berlin, Heidelberg, New York
ISBN-13 978-1-4020-3275-2 (e-book) Springer Dordrecht, Berlin, Heidelberg, New York

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Printed on acid-free paper

All Rights Reserved

© 2005 Springer

No part of this work may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording or otherwise, without written permission from the Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work.

Printed in the Netherlands.

To my parents
Nicu

To Merav and Yonatan
Ira

To my parents
Asutosh

To my students:
Past, present, and future
Tom

Contents

Foreword	xi
Preface	xiii
1. INTRODUCTION	1
1 Research Issues on Learning in Computer Vision	2
2 Overview of the Book	6
3 Contributions	12
2. THEORY: PROBABILISTIC CLASSIFIERS	15
1 Introduction	15
2 Preliminaries and Notations	18
2.1 Maximum Likelihood Classification	18
2.2 Information Theory	19
2.3 Inequalities	20
3 Bayes Optimal Error and Entropy	20
4 Analysis of Classification Error of Estimated (<i>Mismatched</i>) Distribution	27
4.1 Hypothesis Testing Framework	28
4.2 Classification Framework	30
5 Density of Distributions	31
5.1 Distributional Density	33
5.2 Relating to Classification Error	37
6 Complex Probabilistic Models and Small Sample Effects	40
7 Summary	41

3. THEORY:	
GENERALIZATION BOUNDS	45
1 Introduction	45
2 Preliminaries	47
3 A Margin Distribution Based Bound	49
3.1 Proving the Margin Distribution Bound	49
4 Analysis	57
4.1 Comparison with Existing Bounds	59
5 Summary	64
4. THEORY:	
SEMI-SUPERVISED LEARNING	65
1 Introduction	65
2 Properties of Classification	67
3 Existing Literature	68
4 Semi-supervised Learning Using Maximum Likelihood Estimation	70
5 Asymptotic Properties of Maximum Likelihood Estimation with Labeled and Unlabeled Data	73
5.1 Model Is Correct	76
5.2 Model Is Incorrect	77
5.3 Examples: Unlabeled Data Degrading Performance with Discrete and Continuous Variables	80
5.4 Generating Examples: Performance Degradation with Univariate Distributions	83
5.5 Distribution of Asymptotic Classification Error Bias	86
5.6 Short Summary	88
6 Learning with Finite Data	90
6.1 Experiments with Artificial Data	91
6.2 Can Unlabeled Data Help with Incorrect Models? Bias vs. Variance Effects and the Labeled-unlabeled Graphs	92
6.3 Detecting When Unlabeled Data Do Not Change the Estimates	97
6.4 Using Unlabeled Data to Detect Incorrect Modeling Assumptions	99
7 Concluding Remarks	100

5. ALGORITHM:	
MAXIMUM LIKELIHOOD MINIMUM ENTROPY HMM	103
1 Previous Work	103
2 Mutual Information, Bayes Optimal Error, Entropy, and Conditional Probability	105
3 Maximum Mutual Information HMMs	107
3.1 Discrete Maximum Mutual Information HMMs	108
3.2 Continuous Maximum Mutual Information HMMs	110
3.3 Unsupervised Case	111
4 Discussion	111
4.1 Convexity	111
4.2 Convergence	112
4.3 Maximum A-posteriori View of Maximum Mutual Information HMMs	112
5 Experimental Results	115
5.1 Synthetic Discrete Supervised Data	115
5.2 Speaker Detection	115
5.3 Protein Data	117
5.4 Real-time Emotion Data	117
6 Summary	117
6. ALGORITHM:	
MARGIN DISTRIBUTION OPTIMIZATION	119
1 Introduction	119
2 A Margin Distribution Based Bound	120
3 Existing Learning Algorithms	121
4 The Margin Distribution Optimization (MDO) Algorithm	125
4.1 Comparison with SVM and Boosting	126
4.2 Computational Issues	126
5 Experimental Evaluation	127
6 Conclusions	128
7. ALGORITHM:	
LEARNING THE STRUCTURE OF BAYESIAN NETWORK CLASSIFIERS	129
1 Introduction	129
2 Bayesian Network Classifiers	130
2.1 Naive Bayes Classifiers	132
2.2 Tree-Augmented Naive Bayes Classifiers	133

3	Switching between Models: Naive Bayes and TAN Classifiers	138
4	Learning the Structure of Bayesian Network Classifiers: Existing Approaches	140
4.1	Independence-based Methods	140
4.2	Likelihood and Bayesian Score-based Methods	142
5	Classification Driven Stochastic Structure Search	143
5.1	Stochastic Structure Search Algorithm	143
5.2	Adding VC Bound Factor to the Empirical Error Measure	145
6	Experiments	146
6.1	Results with Labeled Data	146
6.2	Results with Labeled and Unlabeled Data	147
7	Should Unlabeled Data Be Weighed Differently?	150
8	Active Learning	151
9	Concluding Remarks	153
8.	APPLICATION: OFFICE ACTIVITY RECOGNITION	157
1	Context-Sensitive Systems	157
2	Towards Tractable and Robust Context Sensing	159
3	Layered Hidden Markov Models (LHMMs)	160
3.1	Approaches	161
3.2	Decomposition per Temporal Granularity	162
4	Implementation of SEER	164
4.1	Feature Extraction and Selection in SEER	164
4.2	Architecture of SEER	165
4.3	Learning in SEER	166
4.4	Classification in SEER	166
5	Experiments	166
5.1	Discussion	169
6	Related Representations	170
7	Summary	172
9.	APPLICATION: MULTIMODAL EVENT DETECTION	175
1	Fusion Models: A Review	176
2	A Hierarchical Fusion Model	177
2.1	Working of the Model	178
2.2	The Duration Dependent Input Output Markov Model	179

<i>Contents</i>	ix
3 Experimental Setup, Features, and Results	182
4 Summary	183
10. APPLICATION: FACIAL EXPRESSION RECOGNITION	187
1 Introduction	187
2 Human Emotion Research	189
2.1 Affective Human-computer Interaction	189
2.2 Theories of Emotion	190
2.3 Facial Expression Recognition Studies	192
3 Facial Expression Recognition System	197
3.1 Face Tracking and Feature Extraction	197
3.2 Bayesian Network Classifiers: Learning the “Structure” of the Facial Features	200
4 Experimental Analysis	201
4.1 Experimental Results with Labeled Data	204
4.1.1 Person-dependent Tests	205
4.1.2 Person-independent Tests	206
4.2 Experiments with Labeled and Unlabeled Data	207
5 Discussion	208
11. APPLICATION: BAYESIAN NETWORK CLASSIFIERS FOR FACE DETECTION	211
1 Introduction	211
2 Related Work	213
3 Applying Bayesian Network Classifiers to Face Detection	217
4 Experiments	218
5 Discussion	222
References	225
Index	237

Foreword

It started with *image processing* in the sixties. Back then, it took ages to digitize a Landsat image and then process it with a mainframe computer. Processing was inspired on the achievements of signal processing and was still very much oriented towards programming.

In the seventies, *image analysis* spun off combining image measurement with statistical pattern recognition. Slowly, computational methods detached themselves from the sensor and the goal to become more generally applicable.

In the eighties, model-driven *computer vision* originated when artificial intelligence and geometric modelling came together with image analysis components. The emphasis was on precise analysis with little or no interaction, still very much an art evaluated by visual appeal. The main bottleneck was in the amount of data using an average of 5 to 50 pictures to illustrate the point.

At the beginning of the nineties, vision became available to many with the advent of sufficiently fast PCs. The Internet revealed the interest of the general public in images, eventually introducing *content-based image retrieval*. Combining independent (informal) archives, as the web is, urges for interactive evaluation of approximate results and hence weak algorithms and their combination in weak classifiers.

In the new century, the last analog bastion was taken. In a few years, sensors have become all digital. Archives will soon follow. As a consequence of this change in the basic conditions datasets will overflow. Computer vision will spin off a new branch to be called something like *archive-based* or *semantic vision* including a role for formal knowledge description in an ontology equipped with detectors. An alternative view is *experience-based* or *cognitive vision*. This is mostly a data-driven view on vision and includes the elementary laws of image formation.

This book comes right on time. The general trend is easy to see. The methods of computation went from dedicated to one specific task to more generally applicable building blocks, from detailed attention to one aspect like filtering

to a broad variety of topics, from a detailed model design evaluated against a few data to abstract rules tuned to a robust application.

From the source to consumption, images are now all digital. Very soon, archives will be overflowing. This is slightly worrying as it will raise the level of expectations about the accessibility of the pictorial content to a level compatible with what humans can achieve.

There is only one realistic chance to respond. From the trend displayed above, it is best to identify basic laws and then to learn the specifics of the model from a larger dataset. Rather than excluding interaction in the evaluation of the result, it is better to perceive interaction as a valuable source of instant learning for the algorithm.

This book builds on that insight: that the key element in the current revolution is the use of machine learning to capture the variations in visual appearance, rather than having the designer of the model accomplish this. As a bonus, models learned from large datasets are likely to be more robust and more realistic than the brittle all-design models.

This book recognizes that machine learning for computer vision is distinctively different from plain machine learning. Loads of data, spatial coherence, and the large variety of appearances, make computer vision a special challenge for the machine learning algorithms. Hence, the book does not waste itself on the complete spectrum of machine learning algorithms. Rather, this book is focussed on machine learning for pictures.

It is amazing so early in a new field that a book appears which connects theory to algorithms and through them to convincing applications.

The authors met one another at Urbana-Champaign and then dispersed over the world, apart from Thomas Huang who has been there forever. This book will surely be with us for quite some time to come.

Arnold Smeulders
University of Amsterdam
The Netherlands
October, 2004

Preface

The goal of computer vision research is to provide computers with human-like perception capabilities so that they can sense the environment, understand the sensed data, take appropriate actions, and learn from this experience in order to enhance future performance. The field has evolved from the application of classical pattern recognition and image processing methods to advanced techniques in image understanding like model-based and knowledge-based vision.

In recent years, there has been an increased demand for computer vision systems to address “real-world” problems. However, much of our current models and methodologies do not seem to scale out of limited “toy” domains. Therefore, the current state-of-the-art in computer vision needs significant advancements to deal with real-world applications, such as navigation, target recognition, manufacturing, photo interpretation, remote sensing, etc. It is widely understood that many of these applications require vision algorithms and systems to work under partial occlusion, possibly under high clutter, low contrast, and changing environmental conditions. This requires that the vision techniques should be robust and flexible to optimize performance in a given scenario.

The field of machine learning is driven by the idea that computer algorithms and systems can improve their own performance with time. Machine learning has evolved from the relatively “knowledge-free” general purpose learning system, the “perceptron” [Rosenblatt, 1958], and decision-theoretic approaches for learning [Blockeel and De Raedt, 1998], to symbolic learning of high-level knowledge [Michalski et al., 1986], artificial neural networks [Rowley et al., 1998a], and genetic algorithms [DeJong, 1988]. With the recent advances in hardware and software, a variety of practical applications of the machine learning research is emerging [Segre, 1992].

Vision provides interesting and challenging problems and a rich environment to advance the state-of-the-art in machine learning. Machine learning technology has a strong potential to contribute to the development of flexible

and robust vision algorithms, thus improving the performance of practical vision systems. Learning-based vision systems are expected to provide a higher level of competence and greater generality. Learning may allow us to use the experience gained in creating a vision system for one application domain to a vision system for another domain by developing systems that acquire and maintain knowledge. We claim that learning represents the next challenging frontier for computer vision research.

More specifically, machine learning offers effective methods for computer vision for automating the model/concept acquisition and updating processes, adapting task parameters and representations, and using experience for generating, verifying, and modifying hypotheses. Expanding this list of computer vision problems, we find that some of the applications of machine learning in computer vision are: segmentation and feature extraction; learning rules, relations, features, discriminant functions, and evaluation strategies; learning and refining visual models; indexing and recognition strategies; integration of vision modules and task-level learning; learning shape representation and surface reconstruction strategies; self-organizing algorithms for pattern learning; biologically motivated modeling of vision systems that learn; and parameter adaptation, and self-calibration of vision systems. As an eventual goal, machine learning may provide the necessary tools for synthesizing vision algorithms starting from adaptation of control parameters of vision algorithms and systems.

The goal of this book is to address the use of several important machine learning techniques into computer vision applications. An innovative combination of computer vision and machine learning techniques has the promise of advancing the field of computer vision, which will contribute to better understanding of complex real-world applications. There is another benefit of incorporating a learning paradigm in the computational vision framework. To mature the laboratory-grown vision systems into real-world working systems, it is necessary to evaluate the performance characteristics of these systems using a variety of real, calibrated data. Learning offers this evaluation tool, since no learning can take place without appropriate evaluation of the results.

Generally, learning requires large amounts of data and fast computational resources for its practical use. However, all learning does not have to be on-line. Some of the learning can be done off-line, e.g., optimizing parameters, features, and sensors during training to improve performance. Depending upon the domain of application, the large number of training samples needed for inductive learning techniques may not be available. Thus, learning techniques should be able to work with varying amounts of a priori knowledge and data.

The effective usage of machine learning technology in real-world computer vision problems requires understanding the domain of application, abstraction of a learning problem from a given computer vision task, and the selection

of appropriate representations for the learnable (input) and learned (internal) entities of the system. To succeed in selecting the most appropriate machine learning technique(s) for the given computer vision task, an adequate understanding of the different machine learning paradigms is necessary.

A learning system has to clearly demonstrate and answer the questions like what is being learned, how it is learned, what data is used to learn, how to represent what has been learned, how well and how efficient is the learning taking place and what are the evaluation criteria for the task at hand. Experimental details are essential for demonstrating the learning behavior of algorithms and systems. These experiments need to include scientific experimental design methodology for training/testing, parametric studies, and measures of performance improvement with experience. Experiments that exhibit scalability of learning-based vision systems are also very important.

In this book, we address all these important aspects. In each of the chapters, we show how the literature has introduced the techniques into the particular topic area, we present the background theory, discuss comparative experiments made by us, and conclude with comments and recommendations.

Acknowledgments

This book would not have existed without the assistance of Marcelo Cirelo, Larry Chen, Fabio Cozman, Michael Lew, and Dan Roth whose technical contributions are directly reflected within the chapters. We would like to thank Theo Gevers, Nuria Oliver, Arnold Smeulders, and our colleagues from the Intelligent Sensory Information Systems group at University of Amsterdam and the IFP group at University of Illinois at Urbana-Champaign who gave us valuable suggestions and critical comments. Beyond technical contributions, we would like to thank our families for years of patience, support, and encouragement. Furthermore, we are grateful to our departments for providing an excellent scientific environment.

Chapter 1

INTRODUCTION

Computer vision has grown rapidly within the past decade, producing tools that enable the understanding of visual information, especially for scenes with no accompanying structural, administrative, or descriptive text information. The Internet, more specifically the Web, has become a common channel for the transmission of graphical information, thus moving visual information retrieval rapidly from stand-alone workstations and databases into a networked environment.

Practicality has begun to dictate that the indexing of huge collections of images by hand is a task that is both labor intensive and expensive - in many cases more than can be afforded to provide some method of intellectual access to digital image collections. In the world of text retrieval, text “speaks for itself” whereas image analysis requires a combination of high-level concept creation as well as the processing and interpretation of inherent visual features. In the area of intellectual access to visual information, the interplay between human and machine image indexing methods has begun to influence the development of computer vision systems. Research and application by the image understanding (IU) community suggests that the most fruitful approaches to IU involve analysis and learning of the type of information being sought, the domain in which it will be used, and systematic testing to identify optimal methods.

The goal of computer vision research is to provide computers with human-like perception capabilities so that they can sense the environment, understand the sensed data, take appropriate actions, and learn from this experience in order to enhance future performance. The vision field has evolved from the application of classical pattern recognition and image processing techniques to ad-

vanced applications of image understanding, model-based vision, knowledge-based vision, and systems that exhibit learning capability. The ability to reason and the ability to learn are the two major capabilities associated with these systems. In recent years, theoretical and practical advances are being made in the field of computer vision and pattern recognition by new techniques and processes of learning, representation, and adaptation. It is probably fair to claim, however, that learning represents the next challenging frontier for computer vision.

1. Research Issues on Learning in Computer Vision

In recent years, there has been a surge of interest in developing machine learning techniques for computer vision based applications. The interest derives from both commercial projects to create working products from computer vision techniques and from a general trend in the computer vision field to incorporate machine learning techniques.

Learning is one of the current frontiers for computer vision research and has been receiving increased attention in recent years. Machine learning technology has strong potential to contribute to:

- the development of flexible and robust vision algorithms that will improve the performance of practical vision systems with a higher level of competence and greater generality, and
- the development of architectures that will speed up system development time and provide better performance.

The goal of improving the performance of computer vision systems has brought new challenges to the field of machine learning, for example, learning from structured descriptions, partial information, incremental learning, focusing attention or learning regions of interests (ROI), learning with many classes, etc. Solving problems in visual domains will result in the development of new, more robust machine learning algorithms that will be able to work in more realistic settings.

From the standpoint of computer vision systems, machine learning can offer effective methods for automating the acquisition of visual models, adapting task parameters and representation, transforming signals to symbols, building trainable image processing systems, focusing attention on target object, and learning when to apply what algorithm in a vision system.

From the standpoint of machine learning systems, computer vision can provide interesting and challenging problems. As examples consider the following: learning models rather than handcrafting them, learning to transfer experience gained in one application domain to another domain, learning from large sets of images with no annotation, designing evaluation criteria for the quality

of learning processes in computer vision systems. Many studies in machine learning assume that a careful trainer provides internal representations of the observed environment, thus paying little attention to the problems of perception. Unfortunately, this assumption leads to the development of brittle systems with noisy, excessively detailed, or quite coarse descriptions of the perceived environment.

Esposito and Malerba [Esposito and Malerba, 2001] listed some of the important research issues that have to be dealt with in order to develop successful applications:

- *Can we learn the models used by a computer vision system rather than handcrafting them?*

In many computer vision applications, handcrafting the visual model of an object is neither easy nor practical. For instance, humans can detect and identify faces in a scene with little or no effort. This skill is quite robust, despite large changes in the visual stimulus. Nevertheless, providing computer vision systems with models of facial landmarks or facial expressions is very difficult [Cohen et al., 2003b]. Even when models have been handcrafted, as in the case of page layout descriptions used by some document image processing systems [Nagy et al., 1992], it has been observed that they limit the use of the system to a specific class of images, which is subject to change in a relatively short time.

- *How is machine learning used in computer vision systems?*

Machine learning algorithms can be applied in at least two different ways in computer vision systems:

- to improve perception of the surrounding environment, that is, to improve the transformation of sensed signals into internal representations, and
- to bridge the gap between the internal representations of the environment and the representation of the knowledge needed by the system to perform its task.

A possible explanation of the marginal attention given to learning internal representations of the perceived environment is that feature extraction has received very little attention in the machine learning community, because it has been considered application-dependent and research on this issue is not of general interest. The identification of required data and domain knowledge requires the collaboration with a domain expert and is an important step of the process of applying machine learning to real-world problems.

Only recently, the related issues of feature selection and, more generally, data preprocessing have been more systematically investigated in machine learning. Data preprocessing is still considered a step of the knowledge discovery process and is confined to data cleaning, simple data transformations (e.g., summarization), and validation. On the contrary, many studies in computer vision and pattern recognition focused on the problems of feature extraction and selection. Hough transform, FFT, and textural features, just to mention some, are all examples of features widely applied in image classification and scene understanding tasks. Their properties have been well investigated and available tools make their use simple and efficient.

- *How do we represent visual information?*

In many computer vision applications, feature vectors are used to represent the perceived environment. However, relational descriptions are deemed to be of crucial importance in high-level vision. Since relations cannot be represented by feature vectors, pattern recognition researchers use graphs to capture the structure of both objects and scenes, while people working in the field of machine learning prefer to use first-order logic formalisms. By mapping one formalism into another, it is possible to find some similarities between research done in pattern recognition and machine learning. An example is the spatio-temporal decision tree proposed by Bischof and Caelli [Bischof and Caelli, 2001], which can be related to logical decision trees induced by some general-purpose inductive learning systems [Blockeel and De Raedt, 1998].

- *What machine learning paradigms and strategies are appropriate to the computer vision domain?*

Inductive learning, both supervised and unsupervised, emerges as the most important learning strategy. There are several important paradigms that are being used: conceptual (decision trees, graph-induction), statistical (support vector machines), and neural networks (Kohonen maps and similar auto-organizing systems). Another emerging paradigm, which is described in detail in this book, is the use of probabilistic models in general and probabilistic graphical models in particular.

- *What are the criteria for evaluating the quality of the learning processes in computer vision systems?*

In benchmarking computer vision systems, estimates of the predictive accuracy, recall, and precision [Huijsman and Sebe, 2004] are considered the main parameters to evaluate the success of a learning algorithm. How-

ever, the comprehensibility of learned models is also deemed an important criterion, especially when domain experts have strong expectations on the properties of visual models or when understanding of system failures is important. Comprehensibility is needed by the expert to easily and reliably verify the inductive assertions and relate them to their own domain knowledge. When comprehensibility is an important issue, the conceptual learning paradigm is usually preferred, since it is based on the comprehensibility postulate stated by Michalski [Michalski, 1983]:

The results of computer induction should be symbolic descriptions of given entities, semantically and structurally similar to those a human expert might produce observing the same entities. Components of these descriptions should be comprehensible as single “chunks” of information, directly interpretable in natural language, and should relate quantitative and qualitative concepts in an integrated fashion.

- *When is it useful to adopt several representations of the perceived environment with different levels of abstraction?*

In complex real-world applications, multi-representations of the perceived environment prove very useful. For instance, a low resolution document image is suitable for the efficient separation of text from graphics, while a finer resolution is required for the subsequent step of interpreting the symbols in a text block (OCR). Analogously, the representation of an aerial view of a cultivated area by means of a vector of textural features can be appropriate to recognize the type of vegetation, but it is too coarse for the recognition of a particular geomorphology. By applying abstraction principles in computer programming, software engineers have managed to develop complex software systems. Similarly, the systematic application of abstraction principles in knowledge representation is the keystone for a long term solution to many problems encountered in computer vision tasks.

- *How can mutual dependency of visual concepts be dealt with?*

In scene labelling problems, image segments have to be associated with a class name or a label, the number of distinct labels depending on the different types of objects allowed in the perceived world. Typically, image segments cannot be labelled independently of each other, since the interpretation of a part of a scene depends on the understanding of the whole scene (holistic view). Context-dependent labelling rules will take such concept dependencies into account, so as to guarantee that the final result is globally (and not only locally) consistent [Haralick and Shapiro, 1979]. Learning context-dependent labelling rules is another research issue, since

most learning algorithms rely on the independence assumption, according to which the solution to a multiclass or multiple concept learning problem is simply the sum of independent solutions to single class or single concept learning problems.

Obviously, the above list cannot be considered complete. Other equally relevant research issues might be proposed, such as the development of noise-tolerant learning techniques, the effective use of large sets of unlabeled images and the identification of suitable criteria for starting/stopping the learning process and/or revising acquired visual models.

2. Overview of the Book

In general, the study of machine learning and computer vision can be divided into three broad categories: *Theory* leading to *Algorithms* and *Applications* built on top of theory and algorithms. In this framework, the applications should form the basis of the theoretical research leading to interesting algorithms. As a consequence, the book was divided into three parts. The first part develops the theoretical understanding of the concepts that are being used in developing algorithms in the second part. The third part focuses on the analysis of computer vision and human-computer interaction applications that use the algorithms and the theory presented in the first parts.

The theoretical results in this book originate from different practical problems encountered when using machine learning in general, and probabilistic models in particular, to computer vision and multimedia problems. The first set of questions arise from the high dimensionality of models in computer vision and multimedia. For example, integration of audio and visual information plays a critical role in multimedia analysis. Different media streams (e.g., audio, video, and text, etc.) may carry information about the task being performed and recent results [Brand et al., 1997; Chen and Rao, 1998; Garg et al., 2000b] have shown that improved performance can be obtained by combining information from different sources compared with the situation when a single modality is considered. At times, different streams may carry similar information and in that case, one attempts to use the redundancy to improve the performance of the desired task by cancelling the noise. At other times, two streams may carry complimentary information and in that case the system must make use of the information carried in both channels to carry out the task. However, the merits of using multiple streams is overshadowed by the formidable task of learning in high dimensional which is invariably the case in multi-modal information processing. Although, the existing theory supports the task of learning in high dimensional spaces, the data and model complexity requirements posed are typically not met by the real life systems. Under such scenario, the existing

results in learning theory falls short of giving any meaningful guarantees for the learned classifiers. This raises a number of interesting questions:

- Can we analyze the learning theory for more practical scenarios?
- Can the results of such analysis be used to develop better algorithms?

Another set of questions arise from the practical problem of data availability in computer vision, mainly labeled data. In this respect, there are three main paradigms for learning from training data. The first is known as *supervised learning*, in which all the training data are labeled, i.e., a datum contains both the values of the attributes and the labeling of the attributes to one of the classes. The labeling of the training data is usually done by an external mechanism (usually humans) and thus the name *supervised*. The second is known as *unsupervised learning* in which each datum contains the values of the attributes but does not contain the label. Unsupervised learning tries to find regularities in the unlabeled training data (such as different clusters under some metric space), infer the class labels and sometimes even the number of classes. The third kind is *semi-supervised learning* in which some of the data is labeled and some unlabeled. In this book, we are more interested in the latter.

Semi-supervised learning is motivated from the fact that in many computer vision (and other real world) problems, obtaining unlabeled data is relatively easy (e.g., collecting images of faces and non-faces), while labeling is difficult, expensive, and/or labor intensive. Thus, in many problems, it is very desirable to have learning algorithms that are able to incorporate a large number of unlabeled data with a small number of labeled data when learning classifiers.

Some of the questions raised in semi-supervised learning of classifiers are:

- Is it feasible to use unlabeled data in the learning process?
- Is the classification performance of the learned classifier guaranteed to improve when adding the unlabeled data to the labeled data?
- What is the value of unlabeled data?

The goal of the book is to address all the challenging questions posed so far. We believe that a detailed analysis of the way machine learning theory can be applied through algorithms to real-world applications is very important and extremely relevant to the scientific community.

Chapters 2, 3, and 4 provide the theoretical answers to the questions posed above. Chapter 2 introduces the basics of probabilistic classifiers. We argue that there are two main factors contributing to the error of a classifier. Because of the inherent nature of the data, there is an upper limit on the performance of any classifier and this is typically referred to as Bayes optimal error. We start by analyzing the relationship between the Bayes optimal performance of

a classifier and the conditional entropy of the data. The mismatch between the true underlying model (one that generated the data) and the model used for classification contributes to the second factor of error. In this chapter, we develop bounds on the classification error under the hypothesis testing framework when there is a mismatch in the distribution used with respect to the true distribution. Our bounds show that the classification error is closely related to the conditional entropy of the distribution. The additional penalty, because of the mismatched distribution, is a function of the Kullback-Leibler distance between the true and the mismatched distribution. Once these bounds are developed, the next logical step is to see how often the error caused by the mismatch between distributions is large. Our average case analysis for the independence assumptions leads to results that justify the success of the conditional independence assumption (e.g., in naive Bayes architecture). We show that in most cases, almost all distributions are very close to the distribution assuming conditional independence. More formally, we show that the number of distributions for which the additional penalty term is large goes down exponentially fast.

Roth [Roth, 1998] has shown that the probabilistic classifiers can be always mapped to linear classifiers and as such, one can analyze the performance of these under the probably approximately correct (PAC) or Vapnik-Chervonenkis (VC)-dimension framework. This viewpoint is important as it allows one to directly study the classification performance by developing the relations between the performance on the training data and the expected performance on the future unseen data. In Chapter 3, we build on these results of Roth [Roth, 1998]. It turns out that although the existing theory argues that one needs large amounts of data to do the learning, we observe that in practice a good generalization is achieved with a much small number of examples. The existing VC-dimension based bounds (being the worst case bounds) are too loose and we need to make use of properties of the observed data leading to data dependent bounds. Our observation, that in practice, classification is achieved with good margin, motivates us to develop bounds based on margin distribution. We develop a classification version of the Random projection theorem [Johnson and Lindenstrauss, 1984] and use it to develop data dependent bounds. Our results show that in most problems of practical interest, data actually reside in a low dimensional space. Comparison with existing bounds on real datasets shows that our bounds are tighter than existing bounds and in most cases less than 0.5.

The next chapter (Chapter 4) provides a unified framework of probabilistic classifiers learned using maximum likelihood estimation. In a nutshell, we discuss what type of probabilistic classifiers are suited for using unlabeled data in a systematic way with the maximum likelihood learning, namely classifiers known as *generative*. We discuss the conditions under which the assertion that unlabeled data are always profitable when learning classifiers, made in

the existing literature, is valid, namely when the assumed probabilistic model matches reality. We also show, both analytically and experimentally, that unlabeled data can be detrimental to the classification performance when the conditions are violated. Here we use the term ‘reality’ to mean that there exists some true probability distribution that generates data, the same one for both labeled and unlabeled data. The terms are more rigorously defined in Chapter 4.

The theoretical analysis although interesting in itself gets really attractive if it can be put to use in practical problems. Chapters 5 and 6 build on the results developed in Chapters 2 and 3, respectively. In Chapter 5, we use the results of Chapter 2 to develop a new algorithm for learning HMMs. In Chapter 2, we show that conditional entropy is inversely related to classification performance. Building on this idea, we argue that when HMMs are used for classification, instead of learning parameters by only maximizing the likelihood, one should also attempt to minimize the conditional entropy between the query (hidden) and the observed variables. This leads to a new algorithm for learning HMMs - MMIHMM. Our results on both synthetic and real data demonstrate the superiority of this new algorithm over the standard ML learning of HMMs.

In Chapter 3, a new, data-dependent, complexity measure for learning – projection profile – is introduced and is used to develop improved generalization bounds. In Chapter 6, we extend this result by developing a new learning algorithm for linear classifiers. The complexity measure – *projection profile* – is a function of the *margin distribution* (the distribution of the distance of instances from a separating hyperplane). We argue that instead of maximizing the margin, one should attempt to directly minimize this term which actually depends on the margin distribution. Experimental results on some real world problems (face detection and context sensitive spelling correction) and on several UCI data sets show that this new algorithm is superior (in terms of classification performance) over Boosting and SVM.

Chapter 7 provides a discussion of the implication of the analysis of semi-supervised learning (Chapter 4) when learning Bayesian network classifiers, suggesting and comparing different approaches that can be taken to utilize positively unlabeled data. Bayesian networks are directed acyclic graph models that represent joint probability distributions of a set of variables. The graphs consist of nodes (vertices in the graph) which represent the random variables and directed edges between the nodes which represent probabilistic dependencies between the variables and the casual relationship between the two connected nodes. With each node there is an associated probability mass function when the variable is discrete, or probability distribution function, when the variable is continuous. In classification, one of the nodes in the graph is the class variable while the rest are the attributes. One of the main advantages of Bayesian networks is the ability to handle missing data, thus it is possible to systematically handle unlabeled data when learning the Bayesian network. The

structure of a Bayesian network is the graph structure of the network. We show that learning the graph structure of the Bayesian network is key when learning with unlabeled data. Motivated by this observation, we review the existing structure learning approaches and point out to their potential disadvantages when learning classifiers. We describe a structure learning algorithm, driven by classification accuracy and provide empirical evidence of the algorithm's success.

Chapter 8 deals with automatic recognition of high level human behavior. In particular, we focus on the office scenario and attempt to build a system that can decode the human activities (*phone conversation, face-to-face conversation, presentation mode, other activity, nobody around, and distant conversation*). Although there has been some work in the area of behavioral analysis, this is probably the first system that does the automatic recognition of human activities in real time from low-level sensory inputs. We make use of probabilistic models for this task. Hidden Markov models (HMMs) have been successfully applied for the task of analyzing temporal data (e.g. speech). Although very powerful, HMMs are not very successful in capturing the long term relationships and modeling concepts lasting over long periods of time. One can always increase the number of hidden states but then the complexity of decoding and the amount of data required to learn increases many fold. In our work, to solve this problem, we propose the use of layered (a type of hierarchical) HMMs (LHMM), which can be viewed as a special case of Stacked Generalization [Wolpert, 1992]. At each level of the hierarchy, HMMs are used as classifiers to do the inference. The inferential output of these HMMs forms the input to the next level of the hierarchy. As our results show, this new architecture has a number of advantages over the standard HMMs. It allows one to capture events at different level of abstraction and at the same time is capturing long term dependencies which are critical in the modeling of higher level concepts (human activities). Furthermore, this architecture provides robustness to noise and generalizes well to different settings. Comparison with standard HMM shows that this model has superior performance in modeling the behavioral concepts.

The other challenging problem related to multimedia deals with automatic analysis/annotation of videos. This problem forms the topic of Chapter 9. Although similar in spirit to the problem of human activity recognition, this problem gets challenging because of the limited number of modalities (audio and vision) and the correlation between them being the key in event identification. In this chapter, we present a new algorithm for detecting events in videos, which combines the features with temporal support from multiple modalities. This algorithm is based on a new framework "Duration dependent input/output Markov models (DDIOMM)". Essentially DDIOMM is a time varying Markov model (state transition matrix is a function of the inputs at any given time) and

the state transition probabilities are modified to explicitly take into account the non-exponential nature of the durations of various events being modeled. Two main features of this model are (a) the ability to account for non-exponential duration and (b) the ability to map discrete state input sequences to decision sequences. The standard algorithms modeling the video-events use HMMs which model the duration of events as an exponentially decaying distribution. However, we argue that the duration is an important characteristic of each event and we demonstrate it by the improved performance over standard HMMs in solving real world problems. The model is tested on the audio-visual event *explosion*. Using a set of hand-labeled video data, we compare the performance of our model with and without the explicit model for duration. We also compare the performance of the proposed model with the traditional HMM and observe an improvement in detection performance.

The algorithms LHMM and DDIOMM presented in Chapters 8 and 9, respectively, have their origins in HMM and are motivated by the vast literature on probabilistic models and some psychological studies arguing that human behavior does have a hierarchical structure [Zacks and Tversky, 2001]. However, the problem lies in the fact that we are using these probabilistic models for classification and not purely for inferencing (the performance is measured with respect to the 0–1 loss function). Although one can use arguments related to Bayes optimality, these arguments fall apart in the case of mismatched distributions (i.e. when the true distribution is different from the used one). This mismatch may arise because of the small number of training samples used for learning, assumptions made to simplify the inference procedure (e.g. a number of conditional independence assumptions are made in Bayesian networks) or may be just because of the lack of information about the true model. Following the arguments of Roth [Roth, 1999], one can analyze these algorithms both from the perspective of probabilistic classifiers and from the perspective of statistical learning theory. We apply these algorithms to two distinct but related applications which require machine learning techniques for multimodal information fusion: office activity recognition and multimodal event detection.

Chapters 10 and 11 demonstrate the theory and algorithms of semi-supervised learning (Chapters 4 and 7) to two classification tasks related to human computer intelligent interaction. The first is facial expression recognition from video sequences using non-rigid face tracking results as the attributes. We show that Bayesian networks can be used as classifiers to recognize facial expressions with good accuracy when the structure of the network is estimated from data. We also describe a real-time facial expression recognition system which is based on this analysis. The second application is frontal face detection from images under various illuminations. We describe the task and show that learning Bayesian network classifiers for detecting faces using our