Computer Vision

Models, Learning, and Inference

This modern treatment of computer vision focuses on learning and inference in probabilistic models as a unifying theme. It shows how to use training data to learn the relationships between the observed image data and the aspects of the world that we wish to estimate, such as the 3D structure or the object class, and how to exploit these relationships to make inferences about the world from new image data.

With minimal prerequisites, the book starts from the basics of probability and model fitting and works up to real examples that the reader can implement and modify to build useful vision systems. Primarily meant for advanced undergraduate and graduate students, the detailed methodological presentation will also be useful for practitioners of computer vision.

- Covers cutting-edge techniques, including graph cuts, machine learning, and multiple view geometry.
- A unified approach shows the common basis for solutions of important computer vision problems, such as camera calibration, face recognition, and object tracking.
- More than 70 algorithms are described in sufficient detail to implement.
- More than 350 full-color illustrations amplify the text.
- The treatment is self-contained, including all of the background mathematics.
- Additional resources at www.computervisionmodels.com.

Dr. Simon J. D. Prince is a faculty member in the Department of Computer Science at University College London. He has taught courses on machine vision, image processing, and advanced mathematical methods. He has a diverse background in biological and computing sciences and has published papers across the fields of computer vision, biometrics, psychology, physiology, medical imaging, computer graphics, and HCI.

Computer Vision

Models, Learning, and Inference

Simon J. D. Prince

University College London

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

32 Avenue of the Americas, New York, NY 10013-2473, USA

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107011793

First published 2012
Reprinted 2013, 2014

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication data
Prince, Simon J. D. (Simon Jeremy Damion), 1972–
Computer vision : models, learning, and inference / Simon J. D. Prince.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-107-01179-3 (hardback)
1. Computer vision. I. Title.
TA1634.P75 2012
006.3′7–dc23 2012008187

ISBN 978-1-107-01179-3 Hardback

Additional resources for this publication at www.computervisionmodels.com

This book is dedicated to Richard Eagle, without whom it would never have been started, and to Lynfa Stroud, without whom it would never have been finished.

Contents

V Models for geometry

VI Models for vision

Acknowledgments

I am incredibly grateful to the following people who read parts of this book and gave me feedback: Yun Fu, David Fleet, Alan Jepson, Marc'Aurelio Ranzato, Gabriel Brostow, Oisin Mac Aodha, Xiwen Chen, Po-Hsiu Lin, Jose Tejero Alonso, Amir Sani, Oswald Aldrian, Sara Vicente, Jozef Doboš, Andrew Fitzgibbon, Michael Firman, Gemma Morgan, Daniyar Turmukhambetov, Daniel Alexander, Mihaela Lapusneanu, John Winn, Petri Hiltunen, Jania Aghajanian, Alireza Bossaghzadeh, Mikhail Sizintsev, Roger De Souza-Eremita, Jacques Cali, Roderick de Nijs, James Tompkin, Jonathan O'Keefe, Benedict Kuester, Tom Hart, Marc Kerstein, Alex Borés, Marius Cobzarenco, Luke Dodd, Ankur Agarwal, Ahmad Humayun, Andrew Glennerster, Steven Leigh, Matteo Munaro, Peter van Beek, Hu Feng, Martin Parsley, Jordi Salvador Marcos, Josephine Sullivan, Steve Thompson, Laura Panagiotaki, Damien Teney, Malcolm Reynolds, Francisco Estrada, Peter Hall, James Elder, Paria Mehrani, Vida Movahedi, Eduardo Corral Soto, Ron Tal, Bob Hou, Simon Arridge, Norberto Goussies, Steve Walker, Tracy Petrie, Kostantinos Derpanis, Bernard Buxton, Matthew Pediaditis, Fernando Flores-Mangas, Jan Kautz, Alastair Moore, Yotam Doron, Tahir Majeed, David Barber, Pedro Quelhas, Wenchao Zhang, Alan Angold, Andrew Davison, Alex Yakubovich, Fatemeh Jamali, David Lowe, Ricardo David, Jamie Shotton, Andrew Zisserman, Sanchit Singh, Vincent Lepetit, David Liu, Marc Pollefeys, Christos Panagiotou, Ying Li, Shoaib Ehsan, Olga Veksler, Modesto Castrillón Santana, Axel Pinz, Matteo Zanotto, Gwynfor Jones, Brian Jensen, Mischa Schirris, Jacek Zienkiewicz, Etienne Beauchesne, Erik Sudderth, Giovanni Saponaro, Moos Hueting, Phi Hung Nguyen, Tran Duc Hieu, Simon Julier, Oscar Plag, Thomas Hoyoux, Abhinav Singh, Dan Farmer, Samit Shah, Martijn van der Veen, Gabriel Brostow, Marco Brambilla, Sebastian Stabinger, Tamaki Toru, Stefan Stavref, Xiaoyang Tan, Hao Guan, William Smith, Shanmuganathan Raman, Mikhail Atroshenko, Xiaoyang Tan, Jonathan Weill, Shotaro Moriya and Alessandro Gentilini. This book is much better because of your selfless efforts!

I am also especially grateful to Sven Dickinson, who hosted me at the University of Toronto for nine months during the writing of this book; Stephen Boyd, who let me use his beautiful LaTeX template; and Mikhail Sizintsev, for his help in summarizing the bewildering literature on dense stereo vision. I am extremely indebted to Gabriel Brostow, who read the entire draft and spent hours of his valuable time discussing it with me. Finally, I am grateful to Bernard Buxton, who taught me most of this material in the first place and has supported my career in computer vision and every stage.

Foreword

I was very pleased to be asked to write this foreword, having seen snapshots of the development of this book since its inception. I write this having just returned from BMVC 2011, where I found that others had seen draft copies, and where I heard comments like "What amazing figures!", "It's so comprehensive!", and "He's so Bayesian!".

But I don't want you to read this book just because it has amazing figures and provides new insights into vision algorithms of every kind, or even because it's "Bayesian" (although more on that later). I want you to read it because it makes clear the most important distinction in computer vision research: the difference between "model" and "algorithm." This is akin to the distinction that Marr made with his three-level computational theory, but Prince's two-level distinction is made beautifully clear by his use of the language of probability.

Why is this distinction so important? Well, let us look at one of the oldest and apparently easiest problems in vision: separating an image into "figure" and "ground." It is still common to hear students new to vision address this problem just as the early vision researchers did, by reciting an algorithm: first I'll use PCA to find the dominant color axis, then I'll generate a grayscale image, then I'll threshold that at some value, then I'll clean up the holes using morphological operators. Trying their recipe on some test images, the novice discovers that real images are rather more complicated, so new steps are added: I'll need some sort of adaptive threshold, I can get that by blurring the edge map and locally computing maxima.

However, as most readers will already know, such recipes are extremely brittle, meaning that the various "magic numbers" controlling each step all interact, making it impossible to find a set of parameters that works for all images (or even a useful subset). The root of this problem is that the objective of the algorithm has never been defined. What do we *mean* by figure and ground separation? Can we specify what we mean mathematically?

When vision researchers began to address these problems, the language of statistics and Markov random fields allowed a clean distinction between the objective and the algorithm to be drawn. We write down not the steps to solve the problem but the problem itself, for example, as a function to be minimized. In the language of this book, we write down formulae for all the probability distributions that define the problem and then perform operations on those distributions in order to provide answers. This book shows how this can be done for a huge variety of vision problems and how doing so provides more robust solutions that are much easier to reason about.

This is not to say that one can just write down the model and ask others to solve for its parameters because the space of possible models is so much vaster than the space of ones in which the solution is tractable. Thus, one always has at the back of one's mind a collection of models known to be soluble, and one always tries to find a model for one's problem, which is near some soluble problem. At that stage, one may well think in

terms of strategies such as "I can probably generalize alpha expansion a bit to solve for the discrete parameters, and then I can use a Gauss-Newton method for the continuous ones, and that will probably be slow, but it will tell me if it's worth trying to invent a faster combined algorithm." Such strategies are common and can be helpful, provided one always retains an idea of the model underlying them.

However, even armed with the attitudes this book will engender, experienced researchers today can fall into the trap of failing to distinguish model and algorithm. They find themselves thinking thoughts like: "I'll fit a mixture of Gaussians to the color distribution. Then I'll model the mixture weights as an MRF and use graph cuts to update them. Then I'll go back to step 1 and repeat." The good news is that often such recipes can be turned back into models. Even if the only known way of fitting the model is to use the recipe you just thought of, the discipline of thinking of it as a model allows you to reason about it, to make use of alternative techniques, and ultimately to do better research. Reading this book is a sure way to improve your ability to make that jump.

So what is this language of probabilities that will allow us to become better researchers? Well, let me provide my "Engineer's view of Bayes' theorem." It is common to hear a distinction between "Bayesians" and "Frequentists," but I think many engineers have a much more fundamental problem with Bayes: Bayesians must lie. Their estimates, biased toward the prior mean, are deliberately different from the most probable reading of their sensor. Consider the example of an "I speak your height" machine whose sensor has a uniformly distributed ± 1 cm error. You receive £1 every time you correctly predict someone's height to within 1 cm. Bayesian principles suggest that if your sensor reads 200 cm ± 1 cm, you should report 199 cm; you will make more money than guessing the actual sensor reading, because more 199 cm people will appear than those of 200 cm. So I as an engineer believe in Bayes as a way of getting better answers, and thus very much welcome this book's pragmatic (but much more subtle than mine) embrace of Bayes. I wonder if it might even be considered a book on statistics with vision examples rather than a book on vision built on probability.

But it would be wrong to finish this foreword without mentioning the figures. They really are good, not because they're beautiful (they often are), but because they provide crucial insights into the workings of even the most basic of algorithms and ideas. The illustrations in Chapters 2–4 are fundamental to the understanding of modern Bayesian inference, and yet I doubt that there are more than a handful of researchers who have ever seen them all. Later figures express extremely complex ideas more clearly than I have ever seen, as well as representing fabulously "clean" implementations of fundamental algorithms, which really show us how the underlying models influence our capabilities.

Finally I believe it is worth directly comparing this book to the recent textbook by my colleague Richard Szeliski. That book too is marked by an enormously comprehensive view of computer vision, by excellent illustration, by insightful notation, and intellectual synthesis of large groups of existing ideas. But in a real sense the two books operate at opposite ends of the pedagogical spectrum: Szeliski is a comprehensive summary of the state of the art in computer vision, the frontier of our knowledge and abilities, while this book addresses the fundamentals of how we make progress in this challenging and exciting field. I look forward to many decades with both on my shelf, or indeed, I suspect, open on my desktop.

Andrew Fitzgibbon
Microsoft Research, Cambridge
September 2011

Preface

There are already many computer vision textbooks, and it is reasonable to question the need for another. Let me explain why I chose to write this volume.

Computer vision is an engineering discipline; we are primarily motivated by the real-world concern of building machines that see. Consequently, we tend to categorize our knowledge by the real-world problem that it addresses. For example, most existing vision textbooks contain chapters on object recognition and stereo vision. The sessions at our research conferences are organized in the same way. The role of this book is to question this orthodoxy: Is this really the way that we should organize our knowledge?

Consider the topic of object recognition. A wide variety of methods have been applied to this problem (e.g., subspace models, boosting methods, bag of words models, and constellation models). However, these approaches have little in common. Any attempt to describe the grand sweep of our knowledge devolves into an unstructured list of techniques. How can we make sense of it all for a new student? I will argue for a different way to organize our knowledge, but first let me tell you how I see computer vision problems.

We observe an image and from this we extract *measurements*. For example, we might use the RGB values directly or we might filter the image or perform some more sophisticated preprocessing. The *vision problem* or *goal* is to use the measurements to infer the *world state*. For example, in stereo vision we try to infer the depth of the scene. In object detection, we attempt to infer the presence or absence of a particular class of object.

To accomplish the goal, we build a *model*. The model describes a family of statistical relationships between the measurements and the world state. The particular member of that family is determined by a set of *parameters*. In *learning* we choose these parameters so they accurately reflect the relationship between the measurements and the world. In *inference* we take a new set of measurements and use the model to tell us about the world state. The methods for learning and inference are embodied in *algorithms*. I believe that computer vision should be understood in these terms: the goal, the measurements, the world state, the model, the parameters, and the learning and inference algorithms.

We could choose to organize our knowledge according to any of these quantities, but in my opinion what is most critical is the model itself – the statistical relationship between the world and the measurements. There are three reasons for this. First, the model type often transcends the application (the same model can be used for diverse vision tasks). Second, the models naturally organize themselves neatly into distinct families (e.g., regression, Markov random fields, camera models) that can be understood in relative isolation. Finally, discussing vision on the level of models allows us to draw

connections between algorithms and applications that initially appear unrelated. Accordingly, this book is organized so that each main chapter considers a different family of models.

On a final note, I should say that I found most of the ideas in this book very hard to grasp when I was first exposed to them. My goal was to make this process easier for subsequent students following the same path; I hope that this book achieves this and inspires the reader to learn more about computer vision.