Computer vision: models, learning and inference

Simon J.D. Prince

July 7, 2012

Please mail errata to s.prince@cs.ucl.ac.uk.

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

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, Erik Sudderth, Etienne Beauchesne, Moos Hueting, Giovanni Saponaro, Phi Hung Nguyen, Tran Duc Hieu, Simon Julier, Oscar Plag and Thomas Hoyoux. 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 nearby some soluble one. 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, providing 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 200cm ± 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 to 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
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.

Chapter 1

Introduction

The goal of computer vision is to extract useful information from images. This has proved a surprisingly challenging task; it has occupied thousands of intelligent and creative minds over the last four decades, and despite this we are still far from being able to build a general-purpose "seeing machine."

Part of the problem is the complexity of visual data. Consider the image in figure 1.1. There are hundreds of objects in the scene. Almost none of these are presented in a "typical" pose. Almost all of them are partially occluded. For a computer vision algorithm, it is not even easy to establish where one object ends and another begins. For example, there is almost no change in the image intensity at the boundary between the sky and the white building in the background. However, there is a pronounced change in intensity on the back window of the SUV in the foreground, although there is no object boundary or change in material here.

We might have grown despondent about our chances of developing useful computer vision algorithms if it were not for one thing: we have concrete proof that vision is possible because our own visual systems make light work of complex images such as figure 1.1. If I ask you to count the trees in this image, or to draw me a sketch of the street layout, you can do this easily. You might even be able to pinpoint where this photo was taken on a world map by extracting subtle visual clues such as the ethnicity of the people, the types of cars and trees, and the weather.

So, computer vision is not impossible, but it is very challenging; perhaps this was not appreciated at first because what we perceive when we look at a scene is already highly processed. For example, consider observing a lump of coal in bright sunlight and then moving to a dim indoor environment and looking at a piece of white paper. The eye will receive far more photons per unit area from the coal than from the paper, but we nonetheless perceive the coal as black and the paper as white. The visual brain performs many tricks of this kind, but unfortunately when we build vision algorithms we do not have the benefit of this preprocessing.

Nonetheless, there has been remarkable recent progress in our understanding of computer vision, and the last decade has seen the first large scale deployments of consumer computer vision technology. For example, most digital cameras now have embedded algorithms for face detection, and at the time of writing the Microsoft Kinect (a peripheral that allows real-time tracking of the human body) holds the

Figure 1.1 A visual scene containing many objects, almost all of which are partially occluded. The red circle indicates a part of the scene where there is almost no brightness change to indicate the boundary between the sky and the building. The green circle indicates a region in which there is a large intensity change but this is due to irrelevant lighting effects; there is no object boundary or change in the object material here.

Guinness World Record for being the fastest-selling consumer electronics device ever. The principles behind both of these applications and many more are explained in this book.

There are a number of reasons for the rapid recent progress in computer vision. The most obvious is that the processing power, memory, and storage capacity of computers has vastly increased; before we disparage the progress of early computer vision pioneers, we should pause to reflect that they would have needed specialized hardware to hold even a single high-resolution image in memory. Another reason for the recent progress in this area has been the increased use of machine learning. The last 20 years have seen exciting developments in this parallel research field, and these are now deployed widely in vision applications. Not only has machine learning provided many useful tools, it has also helped us understand existing algorithms and their connections in a new light.

The future of computer vision is exciting. Our understanding grows by the day, and it is likely that artificial vision will become increasingly prevalent in the next decade. However, this is still a young discipline. Until recently, it would have been unthinkable to even try to work with complex scenes such as that in figure 1.1. As Szeliski (2010) puts it, “It may be many years before computers can name and outline all of the objects in a photograph with the same skill as a two year old child.” However, this book provides a snapshot of what we have achieved and the principles behind these achievements.

Organization of the book

The structure of this book is illustrated in figure 1.2. It is divided into six parts.

The first part of the book contains background information on probability. All the models in this book are expressed in terms of probability, which is a useful language for describing computer vision applications. Readers with a rigorous background in engineering mathematics will know much of this material already but should skim these chapters to ensure they are familiar with the notation. Those readers who do not have this background should read these chapters carefully. The ideas are relatively simple, but they underpin everything else in the rest of the book. It may be frustrating to be forced to read fifty pages of mathematics before the first mention of computer vision, but please trust me when I tell you that this material will provide a solid foundation for everything that follows.

The second part of the book discusses machine learning for machine vision. These chapters teach the reader the core principles that underpin all of our methods to extract useful information from images. We build statistical models that relate the image data to the information that we wish to retrieve. After digesting this material, the reader should understand how to build a model to solve almost any vision problem, although that model may not yet be very practical.

The third part of the book introduces graphical models for computer vision. Graphical models provide a framework for simplifying the models that relate the image data to the properties we wish to estimate. When both of these quantities are high dimensional, the statistical connections between them become impractically complex; we can still define models that relate them, but we may not have the training data or computational power to make them useful. Graphical models provide a principled way to assert sparseness in the statistical connections between the data and the world properties.

The fourth part of the book discusses image preprocessing. This is not necessary to understand most of the models in the book, but that is not to say that it is unimportant. The choice of preprocessing method is at least as critical as the choice of model in determining the final performance of a computer vision system. Although image processing is not the main topic of this book, this section provides a compact summary of the most important and practical techniques.

The fifth part of the book concerns geometric computer vision; it introduces the projective pinhole camera – a mathematical model that describes where a given point in the 3D world will be imaged in the pixel array of the camera. Associated with this model are a set of techniques for finding the position of the camera relative to a scene and for reconstructing 3D models of objects.

Finally, in the sixth part of the book, we present several families of vision models

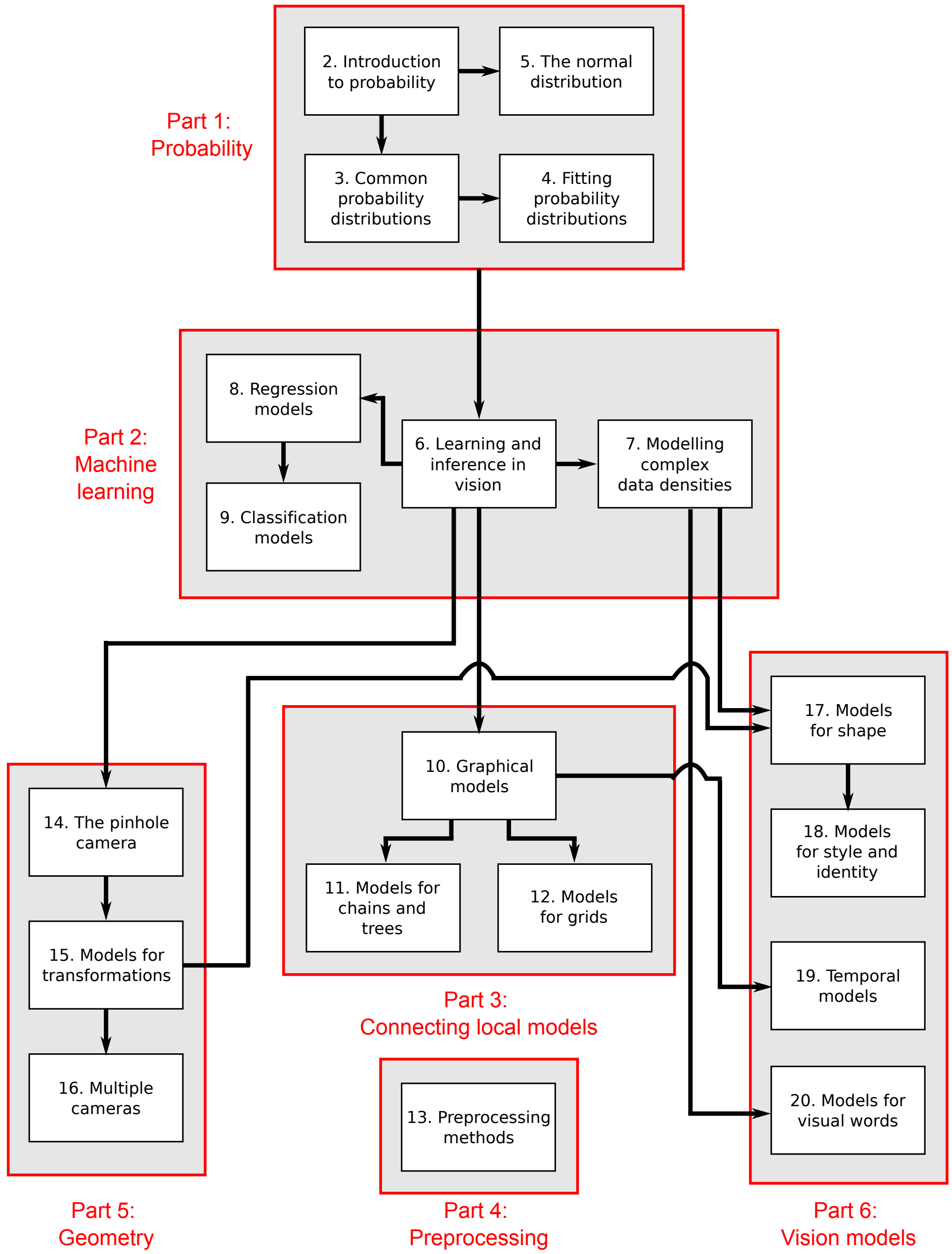

Figure 1.2 Chapter dependencies. The book is organized into six sections. The first section is a review of probability and is necessary for all subsequent chapters. The second part concerns machine learning and inference. It describes both generative and discriminative models. The third part concerns graphical models: visual representations of the probabilistic dependencies between variables in large models. The fourth part describes preprocessing methods. The fifth part concerns geometry and transformations. Finally, the sixth part presents several other important families of vision models.

that build on the principles established earlier in the book. These models address some of the most central problems in computer vision including face recognition, tracking, and object recognition.

The book concludes with several appendices. There is a brief discussion of the notational conventions used in the book, and compact summaries of linear algebra and optimization techniques. Although this material is widely available elsewhere, it makes the book more self-contained and is discussed in the same terminology as the main text.

At the end of every chapter is a brief notes section. This provides details of the related research literature. It is heavily weighted toward the most useful and recent papers and does not reflect an accurate historical description of each area. There are also a number of exercises for the reader at the end of each chapter. In some cases, important but tedious derivations have been excised from the text and turned into problems to retain the flow of the main argument. Here, the solution will be posted on the main book website (http://www.computervisionmodels.com). A series of applications are also presented at the end of each chapter (apart from chapters 1-5 and chapter 10 which contain only theoretical material). Collectively, these represent a reasonable cross-section of the important vision papers of the last decade.

Finally, pseudocode for over 70 of the algorithms discussed is available and can be downloaded in a separate document from the associated website. This pseudocode uses the same notation as the book and will make it easy to implement many of the models. I chose not to include this in the main text because it would have decreased the readability. However, I encourage all readers of this book to implement as many of the models as possible. Computer vision is a practical engineering discipline and you can learn a lot by experimenting with real code.

Other books

I am aware that most people will not learn computer vision from this book alone so here is some advice about other books that complement this volume. To learn more about machine learning and graphical models I would recommend *Pattern Recognition and Machine Learning* by Bishop (2006) as a good starting point. There are many books on preprocessing, but my favorite is *Feature Extraction and Image Processing* by Nixon & Aguado (2008). The best source for information about geometrical computer vision is without a doubt *Multiple View Geometry* by Hartley & Zisserman (2004). Finally, for a much more comprehensive overview of the state of the art of computer vision and its historical development, consider *Computer Vision: Algorithms and Applications* by Szeliski (2010).

Part I

Probability

Part I: Probability

We devote the first part of this book (chapters 2–5) to a brief review of probability and probability distributions. Almost all models for computer vision can be interpreted in a probabilistic context, and in this book we will present all the material in this light. The probabilistic interpretation may initially seem confusing, but it has a great advantage: it provides a common notation that will be used throughout the book and will elucidate relationships between different models that would otherwise remain opaque.

So why is probability a suitable language to describe computer vision problems? In a camera, the three-dimensional world is projected onto the optical surface to form the image: a two-dimensional set of measurements. Our goal is to take these measurements and use them to establish the properties of the world that created them. However, there are two problems. First, the measurement process is noisy; what we observe is not the amount of light that fell on the sensor, but a noisy estimate of this quantity. We must describe the noise in these data, and for this we use probability. Second, the relationship between world and measurements is generally many to one: there may be many real-world configurations that are compatible with the same measurements. The chance that each of these possible worlds is present can also be described using probability.

The structure of part I is as follows: in chapter 2, we introduce the basic rules for manipulating probability distributions including the ideas of conditional and marginal probability and Bayes' rule. We also introduce more advanced ideas such as independence and expectation.

In chapter 3, we discuss the properties of eight specific probability distributions. We divide these into four pairs. The first set will be used to describe either the observed data or the state of the world. The second set of distributions model the parameters of the first set. In combination, they allow us to fit a probability model and provide information about how certain we are about the fit.

In chapter 4, we discuss methods for fitting probability distributions to observed data. We also discuss how to assess the probability of new data points under the fitted model and how to take account of uncertainty in the fitted model when we do this. Finally, in chapter 5, we investigate the properties of the multivariate normal distribution in detail. This distribution is ubiquitous in vision applications and has a number of useful properties that are frequently exploited in machine vision.

Readers who are very familiar with probability models and the Bayesian philosophy may wish to skip this part and move directly to part II.