Chapter 15

Conclusion

In this book, we have covered a broad range of computer vision topics. Starting with image formation, we have seen how images can be pre-processed to remove noise or blur, segmented into regions, or converted into feature descriptors. Multiple images can be matched and registered, with the results used to estimate motion, track people, reconstruct 3D models, or merge images into more attractive and interesting composites and renderings. Images can also be analyzed to produce semantic descriptions of their content. However, the gap between computer and human performance in this area is still large and is likely to remain so for many years.

Our study has also exposed us to a wide range of mathematical techniques. These include continuous mathematics, such as signal processing, variational approaches, three-dimensional and projective geometry, linear algebra, and least squares. We have also studied topics in discrete mathematics and computer science, such as graph algorithms, combinatorial optimization, and even database techniques for information retrieval. Since many problems in computer vision are inverse problems that involve estimating unknown quantities from noisy input data, we have also looked at Bayesian statistical inference techniques, as well as machine learning techniques to learn probabilistic models from large amounts of training data. As the availability of partially labeled visual imagery on the Internet continues to increase exponentially, this latter approach will continue to have a major impact on our field.

You may ask: why is our field so broad and aren't there any unifying principles that can be used to simplify our study? Part of the answer lies in the expansive definition of computer vision, which is the analysis of images and video, as well as the incredible complexity inherent in the formation of visual imagery. In some ways, our field is as complex as the study of automotive engineering, which requires an understanding of internal combustion, mechanics, aerodynamics, ergonomics, electrical circuitry, and control systems, among other topics. Computer vision similarly draws on a wide variety of sub-disciplines, which makes it challenging to cover in a one-semester course, let alone to achieve mastery during a course of graduate studies. Conversely, the incredible breadth and technical complexity of computer vision problems is what draws many people to this research field.

Because of this richness and the difficulty in making and measuring progress, I have attempted to instill in my students and in readers of this book a discipline founded on principles from engineering, science, and statistics.

R. Szeliski, *Computer Vision: Algorithms and Applications*, Texts in Computer Science,
DOI 10.1007/978-1-84882-935-0_15, © Springer-Verlag London Limited 2011

The engineering approach to problem solving is to first carefully define the overall problem being tackled and to question the basic assumptions and goals inherent in this process. Once this has been done, a number of alternative solutions or approaches are implemented and carefully tested, paying attention to issues such as reliability and computational cost. Finally, one or more solutions are deployed and evaluated in real-world settings. For this reason, this book contains many different alternatives for solving vision problems, many of which are sketched out in the exercises for students to implement and test on their own.

The scientific approach builds upon a basic understanding of physical principles. In the case of computer vision, this includes the physics of man-made and natural structures, image formation, including lighting and atmospheric effects, optics, and noisy sensors. The task is to then invert this formation using stable and efficient algorithms to obtain reliable descriptions of the scene and other quantities of interest. The scientific approach also encourages us to formulate and test hypotheses, which is similar to the extensive testing and evaluation inherent in engineering disciplines.

Lastly, because so much about the image formation process is inherently uncertain and ambiguous, a statistical approach that models both uncertainty in the world (e.g., the number and types of animals in a picture) and noise in the image formation process, is often essential. Bayesian inference techniques can then be used to combine prior and measurement models to estimate the unknowns and to model their uncertainty. Machine learning techniques can be used to create the probabilistic models in the first place. Efficient learning and inference algorithms, such as dynamic programming, graph cuts, and belief propagation, often play a crucial role in this process.

Given the breadth of material we have covered in this book, what new developments are we likely to see in the future? As I have mentioned before, one of the recent trends in computer vision is using the massive amounts of partially labeled visual data on the Internet as sources for learning visual models of scenes and objects. We have already seen data-driven approaches succeed in related fields such as speech recognition, machine translation, speech and music synthesis, and even computer graphics (both in image-based rendering and animation from motion capture). A similar process has been occurring in computer vision, with some of the most exciting new work occurring at the intersection of the object recognition and machine learning fields.

More traditional quantitative techniques in computer vision such as motion estimation, stereo correspondence, and image enhancement, all benefit from better prior models for images, motions, and disparities, as well as efficient statistical inference techniques such as those for inhomogeneous and higher-order Markov random fields. Some techniques, such as feature matching and structure from motion, have matured to where they can be applied to almost arbitrary collections of images of static scenes. This has resulted in an explosion of work in 3D modeling from Internet datasets, which again is related to visual recognition from massive amounts of data.

While these are all encouraging developments, the gap between human and machine performance in semantic scene understanding remains large. 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, we have to remember that human performance is often the result of many years of training and familiarity and often works best in special ecologically important situations. For example, while humans appear to be experts at face recognition, our actual

performance when shown people we do not know well is not that good. Combining vision algorithms with general inference techniques that reason about the real world will likely lead to more breakthroughs, although some of the problems may turn out to be "AI-complete", in the sense that a full emulation of human experience and intelligence may be necessary.

Whatever the outcome of these research endeavors, computer vision is already having a tremendous impact in many areas, including digital photography, visual effects, medical imaging, safety and surveillance, and Web-based search. The breadth of the problems and techniques inherent in this field, combined with the richness of the mathematics and the utility of the resulting algorithms, will ensure that this remains an exciting area of study for years to come.