Computer Vision needs a Core and Foundations

A.L. Yuille
Dept. of Statistics, University of California at Los Angeles
Dept. of Brain and Cognitive Engineering, Korea University, Seoul, Korea
yuille@stat.ucla.edu

Abstract

I argue that computer vision needs a core of techniques and foundational research to enable it to build on its current successes and achieve its enormous potential.

"How do I know what papers to read in computer vision? There are so many. And they are so different." Graduate Student. Xi'An. China. November, 2011.

1. Introduction

Computer vision is starting to become practical and successful. Attendance at conferences keeps increasing and the field is vibrant and active. Major companies such as Google and Microsoft have vision groups and there are a growing number of start-ups. Practical applications like face detection and recognition, Kinect, Google Goggles, "Build Rome in a Day", and iPhoto's "Faces" should help computer vision become a household name. Computer vision has started outperforming humans on certain restricted real world tasks such as circuit board inspection and face recognition under controlled conditions. There has also been much progress in traditional application areas like robotics and medical imaging. Moreover, new application areas keep arising such as cosmetic surgery, augmented reality, and vision for the blind. There are growing opportunities for computer vision to provide "outreach" to non-traditional areas such as astronomy, nanotechnology, novel brain imaging techniques, scientific analysis, and many more. The technology that computer vision relies on – computers, the internet, and cameras – keeps improving in quality and its cost keeps decreasing. The computer vision community has grown immensely, particularly since in the early years of this century, has spread far beyond its birthplace in North America and is strongly represented in Europe and Asia.

But from my perspective there are some things lacking which would help make computer vision even more successful. My opinions are based on my long association with the subject and also my experience in different but related disciplines (including cognitive science, the study of biological vision, medical image processing, psychiatric diagnosis from fMRI, machine learning, and a start-up company involving computer vision and other technologies for the visually impaired). These interests led me together with Aude Oliva to organize the recent Frontiers of Computer Vision workshop at MIT sponsored by the National Science Foundation and the Army Research Labs (http://www.frontiersincomputervision.com/). A report based on this workshop is currently in preparation (the views expressed in this article are influenced by the discussions at the Frontiers workshop but represent my personal opinions only).

1.1. How has computer vision changed since 1991?

The last NSF-sponsored workshop on the state of computer vision took place in 1991 (organized by Anil Jain and Shahriar Negahdaripour) and took place following CVPR 1991. The report "Challenges in Computer Vision Research; Future Directions of Research" makes interesting reading (it is available from the Frontier's webpage referred to above).

In 1991 the field of computer vision was fairly small (290 people attended CVPR 1991 while these days CVPR attracts over 1500 people) and was largely dominated by researchers from North America. The discussions at the meeting were partially a counterbalance to some of the more ambitious big picture theories which were frequent in computer vision in the 1980's. Practical progress was limited partly due to the limitations of technology (e.g., few people worked on estimating motion flow because of the slowness of current computers). The most sensible prediction was probably Ted Adelson's comment that progress in computer vision is beginning to happen because researchers are learning to borrow and adapt tools from related disciplines (e.g., Kalman filters, geometry, learning) and apply them. In addition, researchers started concentrating on simpler achievable problems and paid less attention to big picture theories and the more fundamental problems of vision.

The biggest changes since 1991 have been the expansion of the field, including many researchers from Europe

and Asia, the use of large image datasets for learning and evaluation, and the growing number of success stories. In addition, there has been a steady growth in the techniques which computer vision has adapted or developed. Much of this progress, of course, was only possible because of improvements in cameras, computers, the web, and related technology. The issue of datasets is slightly controversial. At their best (e.g., the Pascal Challenge): (i) they helped drive the field forward by proposing difficult challenges, (ii) they contributed to the rapid growth in importance of learning methods, and (iii) they helped benchmark and rank computer vision techniques. This did not surprise speech researchers ("we made no progress at all until we had datasets" as a leading speech expert once told me). But there are concerns about the datasets. Datasets can be small and unrepresentative of the enormous space of natural images so results obtained on them may not generalize to realistic situations. They have also led to a style of research which sees success on datasets as the primary goal – e.g., so that a small two percent percentage improvement is seen as more important than a novel idea. This risks focussing research much too narrowly. For example, my group has been very successful on the Pascal Challenge for object detection (thanks to the efforts of Leo Zhu and Yuanhao Chen) but, frustratingly, we have also had novel work rejected because it "was not tested on Pascal"! But arguably, the rise of datasets and learning has been the biggest difference since 1991 and they have contributed greatly to successes such as face and text detection.

1.2. The need for a core

Despite this significant progress I have concerns about computer vision which are partially illustrated by the Chinese student's question about which papers he should read. As a dynamic and highly interdisciplinary subject, computer vision has developed by incorporating a large range of techniques borrowed or adapted from engineering, mathematics, physics and statistics. New methods are continually being introduced and are often only known to a subpart of the community. A list initiated at Frontiers on "20 techniques that all computer vision researchers should know" had grown to over 80 the last time I checked the web-site. In addition to these techniques there is also considerable accumulated expertise about images and experience about which methods do and do not work. But as in other disciplines which have to deal with complicated data this type of expertise is rarely articulated precisely and is hard for a newcomer to the field to learn and appreciate it (this knowledge includes filter design, the classes of models that really work, and the "biology of vision" meaning the taxonomy of images and visual tasks – if I am interpreting Jitendra Malik correctly). So the computer vision community has accumulated a large amount of knowledge but there has been little effort to synthesize it. There have, for example, been few attempts to understand the relations between different techniques and to what extend they rely on the same underlying ideas. The speed of progress in computer vision seems to often encourage frenetic activity at the expense of thinking about these issues: a recent visitor to my lab said he was surprised at how much time we spent thinking because he was much more used to hacking up an algorithm as quickly as possible.

These issues also affect the interactions between computer vision and other disciplines. When I work in related field like medical imaging and fMRI analysis I see examples where researchers could save themselves a lot of time, and funding agencies a lot of money, by learning from the experience and knowledge of computer vision researchers. But, like the Chinese student in Xi'An, these researchers would find it hard to discover this knowledge by reading the computer vision literature unless carefully guided by an expert.

In short, vision lacks a core of techniques and concepts that are shared by all researchers in the field. The benefits of a core would include the education of new researchers, communication between different schools of computer vision researchers, dissemination of computer vision knowledge to researcher in related fields, evaluation and reviewing of computer vision research, and interaction with industry. The Frontiers workshop provided several examples which illustrated these issues. For example, there were several discussions where different people argued strongly for apparently different intellectual positions which seemed to me on reflection to be fairly straightforward to reconcile in terms of the underlying concepts. Similarly, there were discussions about the limitations of the current review process in computer vision which, to some extent, can be traced to lack of a core (I've had papers rejected because reviewers did not understand dynamic programming, and I am not alone according to Pedro Felzenszwalb). Many of these problems have been exacerbated by the rapid expansion of computer vision in the last ten years but, in one form or another, they have always existed within the community.

What are the arguments against a core? Leading figures in computer vision have sensibly warned about the dangers of "premature theorizing" and there was certainly a history, perhaps strongest in the 1980's, for big concept theories to fail to live up to their promises and also for mathematically complex ideas being needlessly introduced (e.g., attempts to prove structure from motion theorems using techniques like fibre bundles from differential geometry). A strategy of letting 100 flowers bloom is a good way to start exploring a research area. But after many flowers have been planted it makes sense to see which flowers are successful, what are their similarities and differences, and whether we can find some commonalities or underlying structure.

So I argue that computer vision has reached a stage where there should be an established core set of techniques that should be know by all researchers in the field. This should include shared computer code. Exploiting the web by online courses and by wikipedia articles are attractive ways to help establish and disseminate such a core.

1.3. The need for foundations

In addition to a core, I also argue for the need for foundational work. This should attempt to find common unifying concepts and principles which underly computer vision theories and algorithms, which relates vision theories to those developed in related disciplines, and which will enable us to address and solve the fundamental problems of computer vision. Foundational work would ultimately be incorporated in the core.

Here are a few examples to illustrate what I mean by foundational work. In the 17th century Kepler developed twenty laws of planetary motion (some correct, some redundant, some incorrect) based on experimental study and mathematical analysis. But Newton's foundational work on the laws of motion and gravity showed that Kepler's data could be summarized by three laws which had a deeper fundamental explanation (i.e. they also explained why apples fell from trees). More recent examples can be found in engineering, statistics, and machine learning. Many people designed tracking systems in the 1950s which combined prediction and correction stages but progress improved rapidly after Kalman's formulation of this task. Similarly, Dempster *et al.* showed that many methods used to deal with missing data in Statistics could be elegantly unified in terms of the Expectation Maximization (EM) algorithm. Fundamental work on the theory of learning was done by people such as Vapnik and Valiant which helped provide foundations for machine learning research and pointed out the fundamental relations between the size of datasets and the capacity of the hypothesis sets.

I argue that computer vision would benefit from more foundational work which would help clarify the core but also address some of the fundamental problems of computer vision.

What ideas and techniques could supply foundations capable of addressing the complexity of computer vision? I'd argue that probabilistic models defined over structured representations, such as graphs and grammars, offers a very promising framework that subsumes most of the work that would be generally be considered to lie within the core (judging by the topics listed on the 20+ techniques every computer vision researcher should know). I'm using the word "representation" in a broad sense to include geometry and "probability models" to include discriminative methods which learn conditional distributions based on image features (e.g., discriminative random fields). The combination of representations and probabilities helps reconcile early divisions in the computer vision community between those who advocated pattern recognition approaches (e.g., Duda and Hart) and others who embraced an Artificial Intelligence perspective and argued for the fundamental role of representation. It also makes links to pattern theory (Grenander, Mumford and Desolneux) which argues for the need to model the patterns in visual, and other stimuli, by a Bayesian framework which enables both analysis and synthesis. Probability on graphs also subsumes methods developed by different research communities, like Hidden Markov Models and Stochastic Context Free Grammars, and shows the relationships between them (both are examples of probability models with hidden variables defined over graphs without closed loops, hence their computations can be performed by dynamic programming inference algorithms). This framework facilitates learning. Indeed almost all work in machine learning can be expressed in these terms.

Moreover, a similar conceptual framework for cognitive science is being developed by Tenenbaum, Griffiths and their collaborators ("Reverse Engineering the Brain"). This framework gives a way to model psychological phenomena which, at first sight, seems very hard to formulate mathematically. It is also able to reconcile apparent dichotomies in existing theories — e.g., the distinction between rule-based and example-based reasoning – by showing that both can be obtained as two extreme aspects of a deeper underlying theory. In addition, this framework applies to most aspects of cognition and artificial intelligence and helps bring out commonalities and relationships between phenomena in natural language, reasoning, induction, and vision. (This framework enabled somebody like myself, a novice in cognitive science but familiar with the theoretical tools in the framework, to perform research on topics like causal learning, reasoning, and animal conditioning).

But are probability models defined over rich representational structures sufficient to address the fundamental problems of vision? Can they address the complexity challenges of images and the world? This remains to be seen. But it is encouraging that this framework is rich enough to include the recent advances in feature design and that compositional approaches offer a possibility of dealing with the complexity of images and the world.

2. Summary

Although computer vision is becoming successful and its research community is growing rapidly it remains a fragmented field which causes problems in teaching, in communicating between different schools, and interacting with other disciplines. I argue that the field should establish a core and encourage more foundational work addressed at the major unsolved problems of vision. There needs to be a

balance between short term research which can pick the low hanging fruit and the more systematic long term research which develops the tools capable of picking the rest. This is part of the process as computer vision evolves into a mature field with industrial applications.

I believe that a core of computer vision should include some of history. It is discouraging at recent computer vision conferences to see how people who made major contributions to the field seem almost forgotten and their work not referenced. Computer vision has a tendency to re-invent and go round in a circle although fortunately, in Andrew Blake's metaphor, this circle is more like a helix because each time the techniques and understanding gets better and progress gets higher.

Finally, as computer vision starts succeeding it has immense possibilities for outreach into non-standard domains by taking advantage of the growing number of novel imaging devices for studying brain activity, nanotechnology, astronomy, high energy physics, and many more. The images of these devices differ from those in conventional "natural images" but still share many similarities. Computer vision has not always embraced novel areas of this type and has sometimes tended to define computer vision too narrowly (I learnt to submit papers on topics like the detection of particles in high energy physics experiments to NIPS and avoid computer vision reviewers). But as computer vision establishes itself as a mature discipline it can embrace its immense potential for outreach provided, of course, it has a core of knowledge and techniques which can be communicated clearly.

Acknowledgments

The ideas here were influenced by discussions with many people at the Frontiers of Computer Workshop. I like to acknowledge feedback from Daniel Kersten and the hospitality of the Department of Brain and Cognitive Engineering at Korea University where this opinion was written funded by Korean Ministry of Education, Science, and Technology, under the National Research Foundation WCU program R31-10008.