Frontiers in Computer Vision: NSF White Paper. November, 2010.

PIs: Alan Yuille ^1 and Aude Oliva ^2

^1 UCLA Dept. Statistics. Joint appts: Computer Science and Psychology. Email: yuille@stat.ucla.edu
^2 MIT Department of Brain and Cognitive Science. Email: oliva@MIT.EDU

A. Introduction.

Computer vision started with the goal of building machines that can see like humans and perform perception for robots, but it has become much broader than that. Applications such as image database search in the world wide web, computational photography, biological imaging, vision for graphics, GIS, biometrics, vision for nanotechnology, were unanticipated and other applications keep arising as computer vision technology develops. Areas such as document analysis and medical image analysis have developed rapidly and have their own conferences. As our computers achieve even a crude understanding of video imagery, computer vision will profoundly change our lives as visual sensors becomes increasingly ubiquitous and enable us to transcend current human limitations. Rapid developments in supportive technologies -- such as digital cameras and computers -- ensure that computer vision systems will become increasingly more capable and affordable. Moreover, the field of robotics itself has enormous potential to revolutionize manufacturing, to provide service by assistive robots, to perform medical surgery -- applications which all require perceptual input from computer vision systems. In addition, there are many applications to defense, homeland security, and the intelligence community.

But various factors currently prevent computer vision from fully reaching its potential. Firstly, computer vision remains a fragmented and dispersed field which is partially due to its interdisciplinary nature and rapid growth. There is vast duplication of effort, and not enough building on other people's work. Secondly, computer vision lacks the name recognition of related endeavors such as Artificial Intelligence and Robotics. This is partly because it is easy to underestimate the difficulty of computer vision. As humans, we simply open our eyes and seem to effortlessly recognize objects and the structures of scenes. But this apparent ease is highly misleading and reflects instead the enormous amount of neuronal resources -- at least half the cortex -- which is involved in performing these visual tasks. Thirdly, the relations between the academic computer vision community and industry is undeveloped. Both communities work largely independently with little understand of the needs, or achievements, of each other.

Our proposed activity is to organize a workshop to address these issues and to explore frontiers of computer vision. The goal of this workshop is to help articulate a national computer vision agenda and provide a roadmap (similar to recent presentations to congress on robotics). The goals of the roadmapping effort are: (1) to identify the future impact of computer vision on the economic, social, and security needs of the nation; (2) to outline the scientific and technological challenges to address; and (3) to draft a roadmap to address those challenges and realize the benefits.

B. Intellectual Merit and Broad Impact

Computer Vision is a rapidly developing technology with an enormous number of potential applications. It is a very active field with a rapidly growing research community. Major companies (e.g. Microsoft, Google) have large research/development groups and there is a growing number of start-up companies – (see David Lowe's computer vision industry webpage http://www.cs.ubc.ca/~lowe/vision.html). Recent applications suggest that computer vision is finally reaching a level of maturity to enable it to fulfill its promise.

For example, a recent report to Congress (Robotics Roadmap) eloquently demonstrated the importance and promise of robotics industries and stresses the need for computer vision to provide perception (e.g. object recognition, depth estimation). But robotics is only one of many application areas for computer vision – which also include, to only cite a few, image search (e.g. Microsoft, Google), computational photography, reconstruction of three-dimensional scenes, surveillance, inspection, medical image analysis, image enhancement and denoising, aids for the visually impaired.

But how can computer vision build on its successes and overcome its remaining challenges? How can computer vision build on the success and enthusiasm of its growing participants? How can the academic community make connections to industry? How can computer vision best interact with related fields such as Machine Learning, Robotics, Artificial Intelligence, Neuroscience, and Cognitive Science? How can the importance and promise of computer vision be communicated to the general public?

This is a critical time since the research field of computer vision is rapidly expanding -- the expansion is largest in Asia and Europe despite the pioneering role of the US. But academic research is proceeding in an unstructured manner and often with little interaction with industry. The rapid growth of the field and its interdisciplinary nature -- computer vision research is performed in Computer Science, Engineering, Mathematics, Statistics, Psychology, and Neuroscience departments -- means that vision research is often fragmented. In general, the field as a whole would greatly benefit from making closer contacts with real world problems, foster scholarship, and communication of knowledge, datasets, and computer code within computer vision and to related disciplines and the broader community. How to develop a community that encourages long term research based on real world issues and avoids short term "two percent" and 'sound-byte' research?

This meeting will bring together experts in computer vision and related disciplines from academia and industry in order to address these issues. We aim to develop and promote a unified agenda for computer vision research and development between US agencies, universities, and industries (while recognizing that research thrives in a flexible environment). We seek to address issues such as what are the open computer vision tasks – e.g. object recognition, human activities recognition, scene understanding, what are the technical and scientific barriers we must overcome in order to solve these tasks, and what strategies – scientific, organizational, funding – are more likely to lead to greatest progress in addressing these challenges.

C. A Historical Perspective

To address these issues, we start with a historical perspective which includes a review of the only two previous workshops which directly addressed the future of computer vision in 1978 and 1991 respectively.

Research in computer vision started in the 1960's and 70's and the field became firmly established in the 1980's with the founding of the leading journals and conferences (PAMI 1979, CVPR 1983, IJCV 1987, ICCV 1987, and ECCV 1990). By the end of the 1980's, many universities had hired faculty performing research in computer vision, teaching courses, and training graduate students. The number of people involved was still comparatively small, by current standards, and the research was almost entirely being performed in North America.

There have been two national workshops to address the future directions of computer vision which, interestingly, occurred at the start and the end of the 1980's. The first was a workshop held at U. Mass Amherst organized by E. Riseman and A. Hanson which resulted in a volume 'Computer Vision Systems'

published by Academic Press. The second was an NSF-sponsored meeting which took place in Maui in 7-8 June, 1991 with program manager H. Moraff. This workshop was organized by S. Negahdaripour (U. Miami) and A.K. Jain (Michigan State) and resulted in a final report titled "Challenges in Computer Vision Research: Future Directions of Research".

The 1991 report is the starting point for our historical perspective. The report itself consists of 57 pages with a 97 page appendix which consisting largely of questionnaire on future directions filled out by computer vision researchers. The report includes comments from E. Riseman and A. Hanson, the authors of the 1978 report, which gives their perspective. In turn, we have discussed the 1991 report with A.K. Jain to gain the benefit of his experience.

We first briefly summarize the main findings and recommendations of the 1991 report, then we briefly sketch how computer vision has developed in the subsequent twenty years.

The workshop took place over one and a half days. Roughly fifty people attended the workshop with the majority from academia (90%), with a few program managers (5%), and some representatives from industry (5%). All attendees filled out a one page questionnaire before the meeting which asked for suggestions for the workshop and opinions about the state of computer vision and predictions for its future.

Many of the recommendations of the 1991 report remains relevant (although the technical discussions are more dated). The most relevant recommendations are: (i) the need for more experimental validation of models on large datasets, (ii) the sharing of images, algorithms, and models between research groups, (iii) greater interaction between academia and industry, and (iv) the need for complete computer vision systems that perform real world tasks.

Attendees at the workshop were cautious about the future of vision and there was some concern about the shortage of computer vision systems that worked on real images. But most thought that there were grounds for optimism and that there had been steady, but not spectacular progress in the previous ten years. E. Riseman and A. Hanson (writers of the 1978 report) strongly argued that there had been considerable progress stating that researchers in 1991 "know far more about almost all subareas on computer vision than we did in 1980. In 1970, the field was in its infancy and much of the work being done then was 'groping' for suitable 'paradigms'." They also stressed the importance of improved technology by stating, for example, that their workshop in 1978 contained 'no papers on motion analysis (partly because the computational requirements were so staggering)".

There were many predictions about the future. Perhaps the most accurate were the relatively low-key statements about how computer vision would benefit from advances in computers, sensing technologies, and mathematical and computational techniques. For example, E. Adelson stated that advances are coming because "vision people are learning how to appropriately use the tools of applied math and engineering to solve vision problems,…, people are getting better educated in control theory, optimization, signal processing, etc."

What have been the big changes since 1991? Overall there has a lot of activity and continued steady progress. The vision community is much bigger and more optimistic, results on real images are a pre-requisites for most quality publications, computer vision systems have obtained impressive results on problems which seemed impractical only a few years ago. We list some of the more noticeable changes.

(I) The technology has got a lot better and cheaper. Computers are much faster, have far greater memory, and are much cheaper. The internet has developed rapidly and data, algorithms, and reports can be downloaded almost instantaneously. Better sensing devices are available –

e.g., cheap high quality cameras can be attached to cell phones. This has enabled vision researchers to work on large shared datasets, to share code on their webpages, to work with image sequences, and communicate results rapidly.

(II) Computer vision researchers have continued to learn, adapt, develop, and apply tools from mathematics, statistics, computer science, and engineering. Indeed the range of tools available is now so large that specialization is required to keep track of them. Moreover, there has a lot more pragmatic research which has determined, by trial and error, what types of approaches do and do not work.

(III) The research community has grown greatly and the demographic balance has changed. In 1991 computer vision research was largely dominated by the United States with a limited amount of activity in Europe and even less in Asia. During the last twenty years there has been steady growth in the United States, a big expansion in Europe, and recently an enormous expansion in Asia. The expansions in Asia and Europe are largely driven by strong funding for this area. Even within the US the majority of researchers are foreign born.

(IV) Computer vision researchers have developed new tools specific to vision and we list a few examples. (E.g., new filtering methods such as SIFT and HOG, which have been very effective for certain real world tasks). Techniques for detecting and tracking certain types of objects are also well advanced. In addition, there has also been considerable progress in understanding geometry and the ability to reconstruct three dimensional structures from multiple viewpoints.

(V) The use of benchmarked image databases and learning algorithms has become common since 2000. This not only gives objective measures to evaluate and compare different techniques but it also has lead to an enormous growth in learning based methods. These have been successfully applied to a range of problems such as edge detection, region classification, face and text detection, scene understanding, and many more.

(VI) Connections to industry, although still far from ideal, have been strengthened. There are a growing number of start-up companies as well as interest from giants such as Microsoft, Google, Siemens and GE.

(VII) The range of applications of computer vision has grown enormously. Medical images, mentioned in the 1991 report as a small application area, is now a large field in its own right with its own high quality conferences. There has been considerable progress on now established applications, such as web search and video processing, as well as an immense amount of new applications including cosmetic surgery, vision for the blind, forensic vision, analyzing plants.

To conclude, we argue that there has been considerable steady progress since 1991 mostly driven by improvements in hardware, mathematical and computational techniques, and experience. Nevertheless many of the concerns from 1991 remain. Indeed the growth in the number of researchers has arguably increased the fragmentation of the field and there remains lack of scholarship and little progress made on building on research done by others. Unlike most disciplines, computer vision still lacks a basic core of concepts and techniques. The name recognition of computer vision, and the importance and challenge of its problems, remains limited and largely unknown to the general public or politicians. There is still comparatively little interaction between the academic and industrial vision communities. Too much academic research is seen as being neither realistic enough to help develop practical real world systems nor insightful enough to yield new theories and techniques which could eventually lead to progress on

real world problems. The evaluation of computer vision systems on benchmarked datasets has been a big improvement, but these datasets do not compare yet to the complexity of the natural world.

D. A Vision for Computer Vision

We believe that the time has come to re-assess the state of computer vision, to see how it can build on its current successes to achieve its full potential as a mature academic discipline with close relations to industry. To set the stage for the workshop we propose ten key objectives for computer vision.

(I) **Better appreciation and understanding of Computer Vision** among the general public, funding agencies, industry and academia. This includes: (a) appreciation of the potential applications of Computer Vision – robotics is an obvious example, but many others are mentioned in this report, (b) appreciation of the difficulty of some vision problems – building a general purpose vision system is equivalent to understanding half the human cortex, (c) appreciation of what Computer Vision can achieve in the short term and in the long term (and avoid overpromising).

(II) **The establishment of Computer Vision as a coherent intellectual discipline** and clarifying its relationship to related disciplines. This requires Computer Vision practitioners in academia to have a unified core of concepts and techniques, similar to disciplines like Computer Science, Mathematics, Physics, and Statistics. This core should relate to real vision applications and include algorithms and evaluation procedures. Some foundational work should be encouraged particularly if it offers the possibility of yielding a unified conceptual framework, including links to related disciplines such as language processing and higher level cognitive processes such as reasoning. For example, probabilistic grammars and related machine learning approaches arguably have the potential to serve as a unifying conceptual framework for a range of disciplines including vision. In particular, machine learning techniques, coupled with the increasing availability of benchmarked data and the use of mechanical turk, have lead to many practical advances. This core framework should be embodied in books, reviews, web-resources, and other material which provides an efficient summary of the main techniques. In particular, online methods for disseminating this material should be developed. This material should include computer code and datasets of images. This core should encompass knowledge of related disciplines such as Signal Processing, Machine Learning, Natural Language processing, Reasoning, and Robotics. In general, computer vision should be seen as part of a bigger endeavor that includes these disciplines hence enabling multi-media projects (e.g., the combination of natural language and vision to address medical problems).

(III) **Exploiting the relationship of Computer Vision to studies of Biological Vision** systems as performed by Psychologists and Neuroscientists. The human visual system is a major part of the human brain, which is arguably one of the most complex physical systems we know of and understanding it is a major scientific challenge. The relationship between computer and biological vision has long been debated. On the one hand, the human visual system gives proof of concept for computer vision and has served as a source of inspiration for many vision researchers. It is argued that there should be a symbiosis between studying biological and computer vision systems since they must both perform similar tasks within the same visual environment. On the other hand, computer and biological systems function under very different

physical/biological constraints and currently have complementary strengths. Computer vision systems can outperform human systems on certain well-defined tasks in controlled environments, while human vision is much more robust and general purpose. From this perspective, we should seek to exploit their differences by, for example, building interactive vision systems where computer vision reliably solves the 'easy cases' and leaves the 'harder cases' to human experts. Either way, it seems that there is much to be gained by understanding the differences and similarities between computer and biological vision systems. But such understanding requires these two disciplines to develop a common language of theoretical concepts, the use of shared datasets, and the sharing of computer and experimental code.

(IV) **Establish a culture within Computer Vision of scholarship**, the sharing of code, the rigorous evaluation of theories, and a balance between short-term and long-term research. The current lack of scholarship not only results in the frequent reinvention of classic research but, perhaps more seriously, in good papers and grants being rejected by reviewers who lack the necessary expertise to evaluate them or a shared consensus about what constitutes good quality work. This is particularly true for conference papers. There is a culture which evaluates researchers based on the number of papers they produce without taking their quality into account. In addition, older work, beyond a ten year time-span, seems often forgotten and is frequently being re-invented. The conference cycle while adding dynamism often leads to a focus on short-term research, an emphasis on 'sound-bytes', and often small progress, improvements in performance on benchmarked datasets -- rather than long-term quality research. This disrupts the balance between short-term research -- picking the low-hanging fruit -- and long-term research which builds the tools to pick the rest. We suggest re-establishing journal publications with rigorous peer review as the 'gold standard' for referencing, for awarding prizes, for faculty appointments, and promotions.

(V) **Develop closer interaction between industry and academia** – with a few notable exceptions, there is little interaction between industrial vision and academic computer vision. This is unfortunate since one of the main goals of computer vision should be to develop techniques that can be applied to real world problems and used in industrial applications. It is a commonly heard criticism that when industry starts working on a vision problem then the problem is 'solved' and hence of little interest to the computer vision community. Computer vision researchers should understand better the tasks that industrial workers seek to solve and the impressive results they have achieved. In turn, industry should be able to rapidly access state of the art material on important and rapidly developing application areas such as video processing for surveillance. Some ways to achieve this could involve specialized sessions at conferences, including demos of real industrial vision systems, and summer schools sponsored by NSF (as already occur in Europe).

(VI) **Develop a taxonomy of Computer Vision problems, and short and long term challenges.** This taxonomy should address both 'big picture' issues -- such as action recognition, the representation of scenes, the roles of low-,mid- and high-level vision in image understanding among others – as well as specific issues like particular classes of object detection (e.g. car, pedestrian) and image segmentation. Too often vision taxonomies seem to try to subdivide computer vision into modular subparts leading to greater specialization on less and less at the expense of the bigger picture. This over-specialization is often followed by calls for 'integration of modules', which

is difficult since these modules are often designed without a unified conceptual framework or code base. Hence the taxonomy should recognize that the ultimate goal of computer vision is complete image understanding while acknowledging that there are many important real world vision problems which can be solved by more restricted systems. Benchmarked datasets should be developed to stimulate and evaluate methods for addressing these problems.

(VII) **Benchmarked datasets which address the scaling problem**. The fundamental challenge of computer vision is the enormous richness of images and the complexity of the visual environment. How can Computer Vision systems deal with this complexity? For example, how can we scale up video processing systems to deal with the enormous number of security and surveillance applications? Estimates of the number of visual objects range between 20,000 and 200,000 and objects are often partially occluded, can be illuminated in a large variety of different ways. This poses an enormous challenge to computer vision when it seeks to go beyond the limitations of performing restricted range of tasks in restricted environments. The use of benchmarked datasets has been of major benefit to computer vision and, in particular, has shown the feasibility of learning-based approaches. But to lead to useful real world applications, and avoid the risks of being 'toy worlds', these datasets should be large enough to be representative of the complexity of the visual environment and its high dimensionality. Understanding the structures of images – their patterns and redundancies – is critical to provide a basis for both Computer and Biological Vision. Most other disciplines have a clear understanding of their elementary components – for example Physicists study systems composed of quarks, atoms, and molecules – but vision researchers still have only limited understanding of the structure of images. Establishing large well-designed datasets is critical to this endeavor.

(VIII) **Sensing, computer and technology issues.** It is clear from studying the 1991 report that much progress since then was due to the enormous advances in related technology. This will remain true in the future. Many practical vision problems can greatly benefit from the design of novel sensors – e.g., laser sensors, different frequencies – and better understanding of the physics of imaging. Similarly, computer vision algorithms can greatly benefit from the introduction of novel types of computing – e.g., GPUs. Indeed much recent improvement in computer vision has been made possible by the ever increasingly availability of cheap processing power and memory.

(IX) **Develop new funding mechanisms for Computer Vision** which are appropriate to its current state of development. Computer Vision lacks the type of long term funding mechanisms provided to medical research by the National Institute of Health. National Science Foundation funding is rarely renewable and although it is important to keep funding 'transformative' and 'innovative' research, novel grant mechanisms could support the more methodical long term research which is often required to make progress on really difficult problems by enabling researchers to build on the results of their predecessors (e.g., Dickmann's research in Germany on automated cars where a small but coherent team obtained groundbreaking results over an extended period of 10-15 years). In general, funding should have a balance between high-risk potentially transformative research and low-risk solid and thorough research.

(X) **Make the field attractive to talented people**. Treat researchers as explorers and allow flexibility of approaches, while encouraging the development of a community which shares scholarship, datasets, computer code and other resources. Provide postdoctoral and faculty leave fellowships to foster training in new technologies and communication between research groups.

E. The NSF workshop on the Frontiers of Computer Vision

We have received funding from NSF to hold a three day workshop at MIT in late August/early September of 2011 to discuss the issues raised in this report. We will address issues such as what are the major open problems in computer vision? How can they best be addressed? What technical problems must be overcome in order to solve them? How to construct datasets and 'grand challenges' representative of real world problems? How to improve academic relationships with industry? How to enable computer vision to interact best with related disciplines, such as machine learning, cognitive processing, and the study of biological vision systems? How to exploit the growing amounts of visual data now available by learning or other techniques? How can computer vision establish a core set of techniques where academic research can lead directly to industrial applications?

The success of a meeting of this type depends crucially on attracting participants who are world leaders in computer vision and related disciplines, and represent different perspectives, topics, and universities. We propose an advisory board – see below --which will help provide guidance, help in the selection of participants, and provide some 'seed' input to the interactive webpage. All participants will be required to submit a two-page viewpoint paper addressing the topics of this meeting and will be strongly encouraged to join the discussions on the interactive webpage. We will particularly encourage the participation of experts from disciplines which bridge to computer vision – e.g., cognitive science, machine learning, computer graphics, robotic, perceptual science, neurophysiology.

David Forsyth	UIUC	Computer Vision.
Bill Freeman	MIT	Computer Vision, Computational Photography.
Martial Hebert	CMU	Computer Vision, Robotics.
Anil Jain	Michigan State	Computer Vision, Industrial Applications.
Daniel Kersten	UMN	Perceptual Science, Cognitive Neuroscience.
Daphne Koller	Stanford	Machine Learning, Robotics, Computer Vision.
Yann LeCun	NYU	Machine Learning, Computer Vision.
Jitendra Malik	Berkeley	Computer Vision.
Rich Szeliski	Microsoft	Computer Vision, Industry.
Antonio Torralba	MIT	Computer Vision