Future Directions of Visual Common Sense & Recognition

A Summation of Workshop Findings

November 12–13, 2015

Larry Davis, University of Maryland
Devi Parikh, Virginia Tech
Fei-Fei Li, Stanford University

Prepared by Brian Hider and Jeremy Zeigler
Virginia Tech Applied Research Corporation

Workshop funded by the Basic Research Office, Office of the Assistant Secretary of Defense for Research & Engineering. This report does not necessarily reflect the policies or positions of the US Department of Defense

VT-ARC

Innovation is the key to the future, but basic research is the key to future innovation.

–Jerome Isaac Friedman,
Nobel Prize Recipient (1990)

Preface

Over the past century, science and technology has brought remarkable new capabilities to all sectors of the economy; from telecommunications, energy, and electronics to medicine, transportation and defense. Technologies that were fantasy decades ago, such as the internet and mobile devices, now inform the way we live, work, and interact with our environment. Key to this technological progress is the capacity of the global basic research community to create new knowledge and to develop new insights in science, technology, and engineering. Understanding the trajectories of this fundamental research, within the context of global challenges, empowers stakeholders to identify and seize potential opportunities.

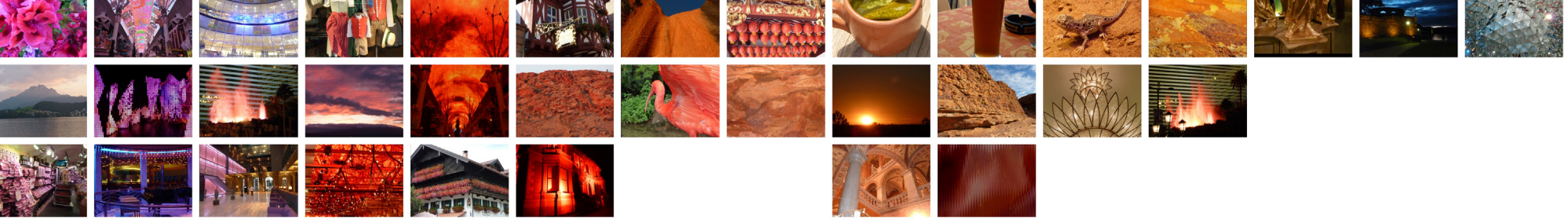
The Future Directions Workshop series, sponsored by the Basic Research Office of the Office of the Assistant Secretary of Defense for Research and Engineering, seeks to examine emerging research and engineering areas that are most likely to transform future technology capabilities.

These workshops gather distinguished academic and industry researchers from the world's top research institutions to engage in an interactive dialogue about the promises and challenges of these emerging basic research areas and how they could impact future capabilities. Chaired by leaders in the field, these workshops encourage unfettered considerations of the prospects of fundamental science areas from the most talented minds in the research community.

Reports from the Future Direction Workshop series capture these discussions and therefore play a vital role in the discussion of basic research priorities. In each report, participants are challenged to address the following important questions:

- How might the research impact science and technology capabilities of the future?
- What is the possible trajectory of scientific achievement over the next 10–15 years?
- What are the most fundamental challenges to progress?

This report is the product of a workshop held November 12–13, 2015 in Arlington VA on the future of Computer Vision research. It is intended as a resource to the S&T community including the broader federal funding community, federal laboratories, domestic industrial base, and academia.



Executive Summary

In November 2015, nineteen distinguished academic researchers met in Arlington Virginia to discuss future opportunities, challenges, and trajectories of the Computer Vision field. Participants included experts in computer vision, artificial intelligence, robotics, and cognitive psychology. The workshop provided a valuable forum for these experts to interact with their colleagues and engage in debate and discussion about the future research trends in computer vision and common sense reasoning. This report is the product of those discussions. Its purpose is to review the past, consider the present, and to frame the future of science and technology in Computer Vision.

The participants reviewed the critical advances over the past 10 years that have greatly improved performance for difficult computer vision tasks like object recognition, 3D reconstruction from multiple images, and tracking of rigid and articulated structures through video.

The consensus was that advancement in computer vision is the result of recent developments in:

- Design of practical algorithms that can train deep neural networks and large graphical models to learn visual recognition schemas from massive annotated data sets.
- Crowdsourcing systems that support collection of massive annotated sets.
- Advances in sensors, especially low cost, high definition video sensors and direct depth sensors (Microsoft Kinect).
- Widespread availability of open source software, such as Open CV.
- Advances in general computing. In particular, the availability of large networks of powerful GPUs and CPU's in conjunction with large, inexpensive data storage.

Because of these advances, the participants identified 6 research opportunities for computer vision that will address the next generation of challenges in building human-vision systems:

1. *Deep learning* theory to address basic questions of engineering vision systems, such as convergence.
2. Visual learning methods requiring far less manually annotated training data.
3. Human/robot systems with a deeper understanding of human intentionality that recover very fine grain descriptions of human activity from video.
4. Computational and biological models of goal-directed visual search for controlling where a robot looks and what it looks for, contingent on its goals.
5. Integration of Common Sense Reasoning to expand what is learned from images and videos from simply what things look like (for object and human activity recognition) to why people (or other agents) behave the way they do.
6. Visual learning theories based on recent advances in neuroscience such as isolating functional visual areas and exploring visual feedback mechanisms.

Progress on these research challenges is critical to the development of the next generation of perceptually enabled robotic systems for military and commercial applications. While today's vision systems have to be engineered to solve a limited set of pre-specified problems, the vision systems of the next two decades will exhibit much more flexibility to changes in goals and operating environments. They will work more naturally with people because they will possess a degree of social intelligence that current systems completely lack. Future vision systems should engender high levels of trust based on their ability to explain the basis for their observations.

The field of Computer Vision is experiencing remarkable progress that depends on sustained basic research to support these developments.
Workshop participants agreed that the United States is a leader in computer vision research but warned that constrained resources makes it difficult to attract and retain top research talent for crucial basic research.
The following sections details the discussion of the workshop participants on the state of the field and future research opportunities, challenges, and trajectories.

Dr. Aude Oliva (MIT) presenting her research on "The Emergence of Representations" at the Future Directions in Computer Vision workshop.

Introduction

Computer vision is concerned with the development of algorithms and systems that can transform images and videos of the natural world into geometric and linguistic descriptions. The field grew out of image processing, which emphasized image-to-image transformations like atmospheric correction, contrast enhancement, and video coding.

The workshop participants discussed the barriers to achieving the sophistication of a human vision system and agreed that there are three primary goals for a computer vision system:

1. **Recognition –** locating and naming objects in images and videos and describing their properties from their appearance.
2. **Reconstruction –** inferring the three dimensional geometry of objects and surfaces from images and videos.
3. **Explanation –** describing the contents of an image or a video in actionable language reflecting the inferred intent and goals of the agents observed.

Advances in hardware, software, and computational power have greatly advanced the capabilities of computer vision systems in these areas, but many challenges have yet to be addressed.

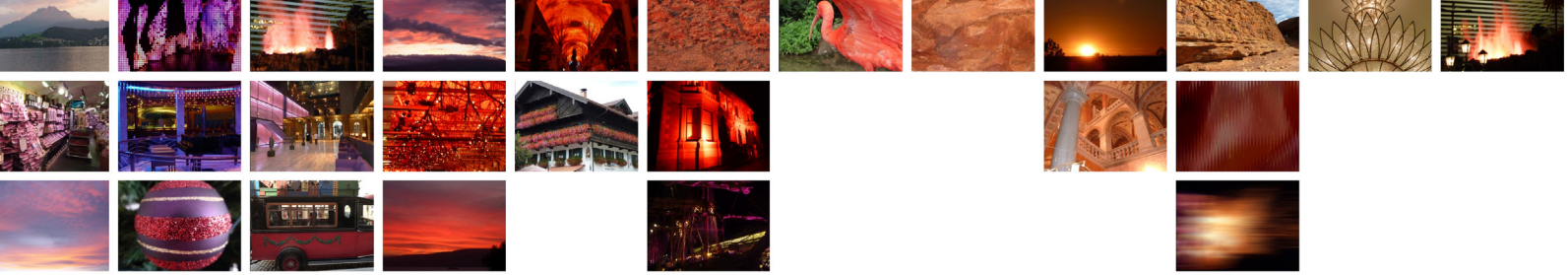
The Fundamental Challenges of Computer Vision

Workshop participants discussed the trajectory of the field to achieving these three goals and identified the fundamental challenges for recognition, reconstruction, and explanation:

For recognition, the computer vision system must both name and localize an object and its movements in an image or video. This is a primitive task for the human visual system, where even young children can flawlessly identify thousands of different object classes. Until recently, this has been far beyond the capabilities of computer vision systems.

There are three psychologically motivated levels of object recognition. The first is category classification of an object—such as recognizing that an object is an instance of a car or a chair. The second is fine grained recognition and involves subcategory identification (for example, identifying the make, model and year of a vehicle or the breed of a dog). Finally, instance level recognition identifies a specific instance of an object like recognizing the face of an individual or a specific vehicle based on its idiosyncratic appearance.

Recognition also requires identification of attributes of objects and movements. Perhaps the most



important example of this is identifying the shape and material composition of an object or surface, because this determines how a robot might interact with the object (drive over it, grasp it, etc.). Attributes also play a central role in fine-grained and instance level recognition. Determining that a face is female constrains instance level recognition of that face.

To date, computer vision has focused a significant amount of attention on recognition of humans in images and videos. Identifying the locations and movements of body parts is critical to understanding what humans are doing and what their intentions might be. Humans are canonical examples of articulated objects. Methods developed for humans then can be applied to other articulated objects such as animals, vehicles, or machines.

Reconstruction is the recovery of the three dimensional structure of objects and surfaces from images and video. Humans use both stereo (binocular) vision and motion for this purpose, and these two modalities have been investigated intensively in computer vision. The *Structure from Motion* (SfM) problem involves reconstructing, from a freely moving camera or a collection of images, the relative locations of the images and a three dimensional model of the world portrayed by the images.

The most challenging goal for vision is to provide an explanation (description) in natural language of an image or a video because it involves integrating what is seen with large volumes of factual and common sense knowledge about the world. It requires new theories of knowledge acquisition, especially everyday common sense physical and spatial knowledge that we all take for granted. It also requires efficient inference models that can integrate this common sense knowledge with uncertain observations from vision. Tracking not just what a person is doing, but also to understand possible reasons for why is an example of common sense reasoning.

Relative Position of the United States in Computer Vision Science

The workshop participants discussed the position of the United States in the field of Computer vision. The consensus was that the United States has been the primary leader in computer vision since the field was established over 50 years ago. They agree that the most innovative research is still largely being conducted in the leading academic and industrial laboratories within the US. However, this lead has dissipated over the past decade because of three significant factors:

1. **China** has established a growing number of computer vision research groups at its leading universities and national centers. Many of these have large numbers (25 to over 100) of Ph. D. level computer vision and machine learning researchers working on both fundamental and applied problems.
2. **The European Community** has invested significantly in building and strengthening academic and industrial computer vision groups across the EU. Over the past decade a large number of multi-year, multi-million Euro projects to consortiums of industry and academia with an applications focus. These programs are principally commercial, but possess strong fundamental research components.
3. **The United States** has experienced a significant outflow of academic researchers from academia to industry in the computer vision field. For example, two leading academic researchers in Structure from Motion left academia for positions at Google, and three of the leading researchers in deep learning, (including two of the pioneers of that area) left senior academic positions for Facebook and Google. Recently, Uber established a research laboratory in Pittsburgh to collaborate with researchers from the Robotics Institute at Carnegie Mellon University, but then promptly hired 30 research scientists and engineers out of the Institute to work at their laboratory.

The United States cannot impact the funding decisions of China and the EU, but it can act to address the issue of outflow from academia to industry. The issue stems from two factors. First, and most importantly, progress in the field has been significant enough that industry now sees many opportunities to monetize computer vision technology. Companies like Google, Facebook, Amazon, Microsoft, and even startups like Uber and Tesla, have hired hundreds of computer vision researchers just over the past year or two.

Secondly, funding for academic research in computer vision—both facilities and personnel—is insufficient, especially for fundamental research.

Past, Present and Future Programs for Computer Vision

Historically, computer vision research has been funded by DOD defense programs for field applications like automatic target recognition, visual surveillance from both ground based and airborne platforms,

robot navigation, mission planning (using detailed and accurate 3D reconstructions obtained from computer vision), and forensics. For example, in the area of UAV video analysis, basic research twenty years ago investigated how the "soda-straw" quality of UAV video (which led to operator fatigue and reduced operator effectiveness) could be overcome by integrating many frames of UAV video into a wide area "mosaic" with moving objects displayed over the static background. This is now an operational capability.

Additionally, ONR, ARO and AFOSR have consistently supported computer vision research through small grants and MURI's that bring together several universities and disciplines to investigate broader problems over longer time periods. For more than two decades, DARPA supported the national computer vision community through its Image Understanding Program. This was by far the longest running research program in the history of DARPA. Over the past twenty years, DARPA funding for computer vision has been more mission oriented, based on more typical 2–4 year focused research projects. Examples of past programs include Visual Surveillance and Monitoring (ground based, fixed camera surveillance), VIRAT (especially UAV surveillance for recognition of people and vehicles), RADIUS (site modeling through 3D reconstruction of building and road networks from satellite imagery) and neoVision (understanding how models of human vision can lead to better computer vision systems). IARPA has also consistently funded computer vision over the past ten years. Recent programs include FINDER (geo-location of images and videos), ALADDIN (video classification and retrieval), and BEST/JANUS (still and video face recognition).

Programs like these are vital to the continued progress in the field and to ensure that the United States maintains and secures its position as the world leader in the field.

In particular, basic research programs that support smaller research institutions are absolutely critical to progress in Computer Vision research. The largest academic research centers, by virtue of their funding, have access to massive data sets required to train deep networks central to Computer Vision research. These centers include computer clusters that provide the computing cycles and memory needed for training algorithms. Unfortunately, the vast majority of researchers in academia do not share this access. **The establishment of a few computing centers would allow smaller research groups to conduct large-scale experimentation with critical access to large data sets.** Additionally, one of the most important future research directions for computer vision involves the integration of powerful computer vision methodologies with robotic systems that can cooperate with humans to perform tasks. This research will require that computer vision and robotics develop new methodologies for evaluating progress.

Currently in computer vision, progress is measured with respect to accuracy of recognition on very large publicly available data sets such as ImageNet. As performance on one data set saturates, researchers collect newer, larger, and more challenging ones with more categories and more complex images. But research in robotic systems cannot be evaluated against static datasets, no matter how large they are, because what a robot perceives depends on the movements and actions it performs. So, a researcher at one institution cannot exactly replicate an experiment performed at another, making overall progress on challenging problems difficult to assess.

The participants discussed solutions to the infrastructure problem:
One potential solution would be to establish a few national robotic challenge centers to act as challenging test environments with state of the art robotic platforms.

Another option could be the establishment of standardized test environments that can be easily deployed in research labs across the country. For instance, many labs might be able to install a pre-specified IKEA kitchen along with a manipulator to perform research in manipulation of complex items such as kitchen tools, fluids, mixtures, and other food items. Such an environment would also support standardized test scenarios. Researchers could download their algorithms and apply them in common settings for fair comparison and objective assessment of the overall progress of the field. Basic research programs are needed to move the computer vision community towards research directions with the greatest potential to impact future science and technology. For example, while engineering advances in deep learning have led to dramatic improvements in object and action recognition, understanding when and why fundamental computer vision models converge is critical. To drive the field forward, we need to be able to answer how these fundamental models can be trained with significantly lower cost than today. This will require interdisciplinary research coordinating computer vision, statistics and applied mathematicians.

The establishment of a few computing centers would allow smaller research groups to conduct large-scale experimentation with critical access to large data sets.

Recent Advances in Computer Vision

The workshop participants agree that general advances in computing, hardware, and software, have greatly contributed to the progress in the closely linked fields of machine learning and computer vision. This section reviews the discussion about these advances.

General Advances in Computing

The most successful computer vision systems typically operate on computing platforms consisting of hundreds to tens of thousands of computing cores and GPU's. This is possible because of the remarkable improvements in cost/performance of GPU's and storage, as well as improvements to CPU architectures like on-chip parallelism. Additionally, robotic platforms like UAVs and small ground-based robots with manipulators, can now be obtained by even small university research groups at reasonable cost. Sensors have also improved significantly over the past decade. RGBD cameras that measure both color and depth like the Kinect are widespread and inexpensive. HD video is now available on any smart phone and cameras like GoPro can be attached to any mobile platform and collect HD video for long periods of time.

Changes in software and dataset availability have also contributed significantly to progress in computer vision. First, the availability of large open source libraries for building vision applications, as well as, the general trend for researchers to make their software available for others to test and build, has greatly reduced the time needed to prototype new vision algorithms and systems. Second, human-centered computing platforms, such as Amazon's Mechanical Turk (AMT), have allowed vision researchers to obtain massive data sets of accurately annotated image and video data for training new computer vision algorithms. A good example is ImageNet, an image dataset akin to the WordNet hierarchy where each meaningful concept is described by multiple words or word phrases, called a "synonym set" or "synset". There are more than 100,000 synsets in WordNet, the majority of which are nouns (80,000+). ImageNet is still growing and will eventually provide on average 1000 images to illustrate each synset. Images of individual concepts are quality-controlled and human-annotated. It is anticipated that ImageNet will eventually offer tens of millions of cleanly sorted images that cover most of the concepts in the WordNet hierarchy.

Another good example is the MS COCO dataset with 300,000 images densely annotated with object regions, captions, questions, answers, and visual phrases. Workshop participants shared their experiences with both the ImageNet and MS COCO datasets.

Advances in Computer Vision and Machine Learning

Participants agreed that the ability to train (and subsequently employ for inference or testing) complex models like deep neural networks, graphical models such as conditional random fields, and structural models like probabilistic grammatical and logical models, have led to significant improvements in solving core vision problems. This includes recognition, tracking, and human activity recognition.

Consider the problem of *object detection*. The ImageNet Large Scale Visual Recognition Challenge (ILSVCR) is a community sponsored competition that challenges participants to detect 1,000 distinct object categories using more than 1,000,000 training images. In 2010, the best performing system in the *image classification* competition had an error rate of 28%. In 2012, a system using deep learning led to an astounding reduction in that error rate to 16%. Just two years later, a group at Google reported an error rate using deep learning of less than 7%, and in 2015 a group from Microsoft Asia reported a 50% reduction in that error rate to 3.5%. So, **in just 6 years, classification accuracy on what was regarded as an extremely challenging data set reduced by a factor of 7.** What accounts for this change? While the general computing architecture of deep networks can be traced back over forty years, their recent success is due to three factors:

1. The availability of large scale computing systems with thousands of cores and ever more powerful GPU's and big data sets involving millions of annotated images for training very deep networks. In just one generation, the Tesla family of GPU's from Nvidia improved from a peak performance of 3.95 single precision Teraflops to 5.04 Teraflops. Additionally, modern CPU's with onboard parallelism can rival GPU's on some deep network computations.

2. The number of critical engineering insights in network design which have led to faster convergence for training deep networks and faster inference. For example, on a popular image recognition benchmark, the best performing deep network in 2014 took 84 hours of GPU time to train. By 2015, a better performing model was able to train on the same dataset in only 1/8 the time. More importantly, the time to process an image with the trained network decreased from 47 seconds per image to only 1/3 second per image with no decrease in classification accuracy.
3. The development of powerful software platforms for architectural design and exploration of design spaces of deep networks. For example, the CAFFE toolbox developed at UC Berkeley, is used by thousands of researchers around the world to construct deep networks and add new functionality to the library. This further accelerates the research cycle as the science and engineering of deep learning progresses.

Today, networks with millions of parameters are routinely trained by researchers all over the world to solve vision problems. These deep networks have more than doubled the accuracy of computer vision systems on central problems such as object recognition (Figure 1) and semantic scene segmentation, in which every pixel in an image or video is labeled with the semantic category of the object class to which it belongs.

Just a decade ago, a fundamental research challenge facing the vision community related to the Structure from Motion (SfM) problem. Today, highly accurate 3D models of static scenes are regularly obtained from both airborne and ground based camera platforms (Figure 2) and used to support applications in autonomous driving, urban planning, or military mission planning. These improvements can be attributed to advances in the design of scalable and robust algorithms for solving underlying problems in geometric statistics (*bundle adjustment* problem).

Microsoft now provides a service called "photo-tourism" in which people upload their travel photographs of monuments and famous sites, and these images are then combined through SfM into a complete three dimensional model of the site (Figure 3).

Most importantly, recent research on SfM has also produced algorithms for building three dimensional models of humans in actions that will have profound implications for robotic systems, as well as, surveillance systems.

A decade ago the problem of detecting and tracking people and analyzing their activities through the movements of their body parts was largely unsolved. Workshop participants agreed that significant progress has been made on this problem with the development of new machine learning methods that can automatically learn a robust decomposition of an articulated object into parts that are then optimized for detection in images and videos. Most notably, *deformable part models* automatically learn the structure of articulated

Figure 1 – Computer vision systems can now recognize and delineate typical objects from images.

(courtesy of Prof. Jitendra Malik, Department of Electrical Engineering and Computer Science, University of California, Berkeley)

Figure 2 – Google Street View collects images as a vehicle drives along the road.

Figure 3 – Computer vision system can reconstruct 3D models of monuments using SfM.

(courtesy of Prof. Jan-Michel Frahm, Department of Computer Science, University of North Carolina)

objects, such as people, and have very efficient inference algorithms that can be used for detection and tracking. A related approach, referred to as "poselets", learns models of a small sets of body parts and joints and can detect and track people in video with great accuracy (Figure 4). Advances in tracking, largely based on new and powerful machine learning methodologies and deep learning, now allow computer vision systems to track people and vehicles even in the midst of moderately crowded scenes over long periods of time.

These advances in computer vision have had a significant economic impact in the U.S and overseas. Autonomous vehicle navigation can trace its roots back to the DARPA Strategic Computing Program of the 1980's. Martin Marietta in Denver Colorado, the prime contractor on the Autonomous Land Vehicle project in Strategic Computing, demonstrated a very rudimentary form of road-following and limited cross country navigation based on computer vision. **Today, Google cars have driven autonomously for more than 1.5 million miles and the Israeli company, Mobileye (started by an MIT computer vision graduate) is the world's major supplier of autonomous driving systems as well as the world's first billion dollar computer vision company.** Tesla, BMW, and Daimler, all have large computer vision groups developing autonomous systems for vehicles. Basic research in object recognition is being integrated in product search engines at Amazon, Walmart and a host of startups around the nation. Crunchbase reported over 150 computer vision startups in January 2016.

The participants agree that the pace of innovation in the field is remarkable and the scope of applications of computer vision technology is staggering. Some of the applications discussed include human behavioral modeling, image search, marketing, medical applications, computer security, UAV navigation and stabilization, and much more. Participants agree that without basic research, these recent advances and applications would not have been possible. And now, basic research is benefiting from the very advances it helped originate.

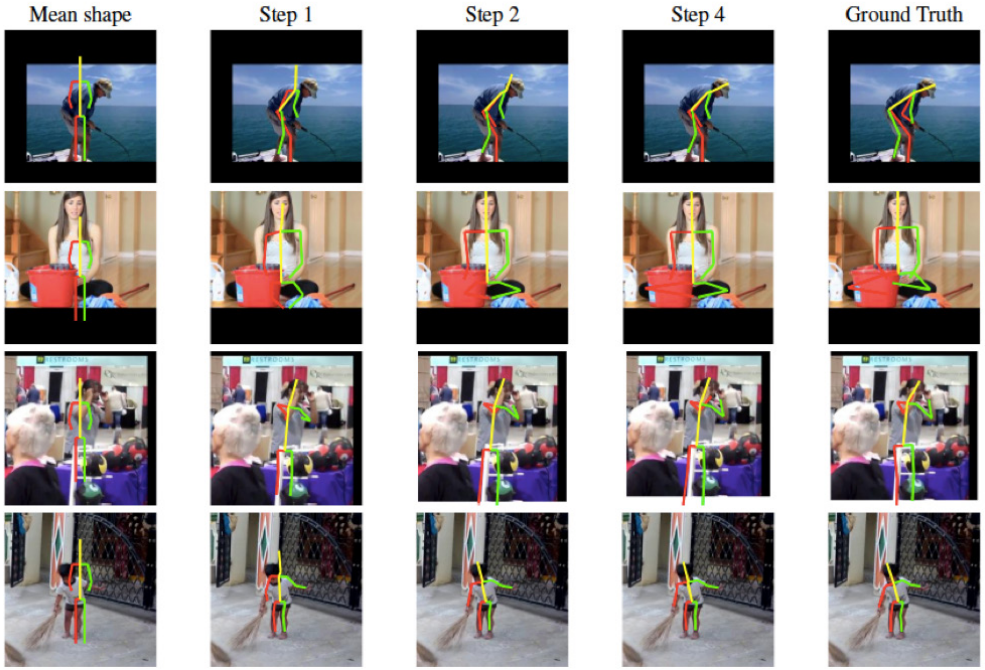


Figure 4 – Computer vision systems can now find body parts of humans in images using the poselet model invented at UC Berkeley.

(courtesy of Prof. Jitendra Malik, Department of Electrical Engineering and Computer Science, University of California, Berkeley)

Projected Opportunities, Trajectories and Goals

Where is computer vision heading in 10 and 20 years?

The workshop participants discussed some promising research trajectories and corresponding opportunities in computer vision where real potential is just starting to materialize. From the "beyond large annotated datasets" and towards more "pure visual learning models," to robots that may possess the same visual capabilities as a young child, there are many exciting trajectories and goals within computer vision worth pursuing. Specific challenges were also discussed. How do you model the complex relationships amongst parts, objects, and scenes? How do you create models that allow robots to generalize, reason, and build "knowledge" about the world around them? This section summarizes some of these conversations from the workshop. The context of these discussions frame general trends in the computer vision research of the future.

Visual Learning

Deep learning of large neural networks models (convergence training methodologies) and learning based on models of human development, will continue to be a central theme of computer vision over the next decade. Today's dominant computational paradigm in vision involves learning the parameters of a deep neural network from massive amounts of curated and manually labeled ground truth data. Researchers have made recent progress on understanding the convergence properties of these network models. Deep networks are highly nonlinear, and the algorithms that train them most successfully integrate a large number of engineering heuristics that appear critical to successful training.

During the next decade we expect major breakthroughs in making algorithms more powerful and efficient. Meanwhile, improvements in software systems for specifying and training these models will make training algorithms available to a much broader community of engineers and practitioners. Continued research in the design and development of both the theory and the practice of deep learning will be critical to success in this area.

Learning with Limited or No Supervision

As mentioned above, it is critical that practical training methods for deep networks and large graphical models be developed that require far less annotated training data than current models demand. There are a number of potentially powerful research themes that the community has begun to explore to address this challenge:

- **Developmental learning –** intelligent systems that actively explore (and manipulate) their environment in order to learn through a series of successes and failures in performing fundamental tasks such as grasping or navigating.
- **Lifelong learning –** learning systems that take advantage of previously learned visual concepts to acquire new ones with minimal supervision. The Never Ending Image Learning (NEIL) system at CMU is an example of a lifelong learning system that has acquired visual recognition models for hundreds of object classes without explicit supervision. It has discovered a large amount of visual common sense knowledge relating these visual classes.
- **Reinforcement learning –** inspired by behavioral psychology, focuses on how robots can learn to take appropriate actions in the world so as to maximize a model of long-term reward.
- **Implicit labeling –** Avoiding the expense of explicit annotation by specifying contextual information or other information that is highly predictive of explicit labels.

Research into efficient learning using these paradigms will lead to significant improvements in training large perceptual models for not only vision, but other modalities like hearing and touch, over the next decade.

Task Adaptation

Task adaptation is a little studied problem with great promise. In particular, the creation of tools and techniques to enable nonpractitioners to adapt visual systems developed on academic datasets to specific applications is just beginning to be explored. For example, construction and manufacturing companies have strong interest in using vision to improve safety, ergonomics, and efficiency of their workers. Currently, applying existing person detectors, pose estimators, trackers, attribute classifiers, etc. needs to be performed by a computer vision expert at much time and expense for each specific scenario and task. **Now that computer vision has reached a high level of competency on the core problems of detection and recognition, the time is right to design new supervision and training methods that enable non-specialists to benefit from progress in computer vision research.** Libraries like OpenCV represent a good first step in this direction, but more sophisticated software infrastructure that encapsulates more advanced computer vision methods may be required for some organizations.

"Now that computer vision has reached a high level of competency on the core problems of detection and recognition, the time is right to design new supervision and training methods that enable non-specialists to benefit from progress in computer vision research."

Figure 5 – Photo of civilians in Bosnia, seeking cover.

Courtesy of Photographer Chris Helgren; taken from http://www.chrishelgren.net/.

Common Sense Reasoning

Common sense reasoning involves the acquisition of visual common sense knowledge and using such common sense to answer questions about images and videos. While computer vision has finally advanced to the stage where it can delineate and identify many objects in an image, simply seeing what is in an image is just the first step in understanding the image in a useful way.

Consider Figure 5 (above) where people are walking in a crouched position along a wall. Computer vision algorithms can now identify the principal objects in the picture (four people, two carrying bags, a wall, a street), recover the poses of the people (qualitatively, they are crouching while walking), and even describe the contents of the image via natural language. But why are these people in these poses? It is certainly not because there is some physical surface above them that they need to avoid. Once it is revealed that the image was taken in Bosnia in 1983 (providing factual historical context) the entire story becomes clear. A war was ravaging that nation, characterized by random sniper killings in the streets, so the people are in crouching poses to reduce their visibility from possible snipers. What is the most likely location of the snipers? Clearly, to the right of the image. If the snipers were on the left then the wall would provide the civilians no protection and crouching would only have slowed their progress through this particular danger zone. We come to these conclusions through a combination of factual and common sense knowledge. Here, the common sense knowledge is about visibility. To reduce or eliminate visibility from some location you must position yourself (location and pose) so that some surface or object largely occludes you from that location.

The acquisition, representation, and utilization of visual common sense knowledge represents a set of critical opportunities in advancing computer vision past the stage where it simply identifies which objects occur in imagery. Over the next two decades workshop participants expect the field to develop computational models that are explanatory and can answer questions about images and videos such as: A. What is there? B. Who is there? C. What is the person doing? D. What environmental factors are influencing their activity?

Eventually, we want visual cognition to be able to answer more difficult questions such as who is doing what to whom, for what reason, and what is most likely to happen next? In fact, the area of data-driven visual prediction of human activity has emerged over just the past 18 months as a growing research theme in the academic community. Moreover, these questions need to be answered at the computational, psychophysical, and neural levels.

Acquiring visual common sense knowledge has only recently drawn the attention of computer vision researchers. Three promising classes of approaches are being pursued:

Figure 6 – NEIL learns visual models of object appearance, taxonomies, and visual common sense from the web.

Courtesy of Prof. Abhinav Gupta, School of Computer Science, Robotics Institute, Carnegie Mellon University.

"The representation of visual common sense knowledge is a central research issue in designing computer vision systems that can explain what they see."

1. Automatic systems like NEIL that follow paths of concepts through the internet and build and expand knowledge from graphs over these concepts (Figure 6).
2. Crowdsourcing in which humans are asked to pose and/or answer questions about images that involve the application of common sense reasoning.
3. Generation of synthetic imagery—clip art—by people (another application of crowd sourcing) that capture common sense facts and relationships.

The representation of visual common sense knowledge is a central research issue in designing computer vision systems that can explain what they see. Significant progress has been made in AI over the past decade that should lead to vision systems with the ability to explain what they see in specific high-value domains over the next decade. There are two complementary approaches that the field is exploring.

First, recent advances in probabilistic logical reasoning, have fundamentally extended classical artificial intelligence approaches based on explicit and semantic representations of knowledge so that they can be utilized for visual reasoning. Second, embedding methods where words, phrases and sentences are mapped to coordinates in a high dimensional "embedding space," suggest that some visual common sense reasoning can be conducted using vector space representations of knowledge obtained from massive data sets. In the "embedded space," reasoning is more implicit. In large part it is absorbed into the knowledge representation.

The most successful word to vector space models are learned by neural networks. Trained on billions of words of text from the web, each word is associated with a point in a high dimensional vector space. The vectors representing the words exhibit some remarkable linguistic structure. For example, if one computes the expression vec(Madrid)- vec(Spain) + vec(France), where vec(word) is the location, that word is mapped to the neural network. The result is the expression is mapped to a vector whose closest word is Paris! Workshop attendees agree that both explicit and implicit approaches need to be investigated. The challenge is how best to integrate these reasoning systems to achieve common sense decisions.

Integration of Computer Vision with Robot Systems

Computer vision systems will soon move out of the internet and into the physical world through joint research projects with roboticists. There is a significant opportunity over the next decade to develop robot systems that can intelligently interact with people to help achieve specific goals. This is intimately related to visual common sense reasoning, because **a robot can only understand what a person's goals are by integrating what activities it observes a person doing with common sense knowledge about how actions reflect goals and constraints.** This is a form of fine grain activity recognition. For example, a vision system might observe a person running in a train station (a category level movement recognition), but only through common sense reasoning can it determine if the person's goal is to catch a departing train or to escape from danger. Once we understand how to represent and acquire visual common sense, robots possessing "social understanding"

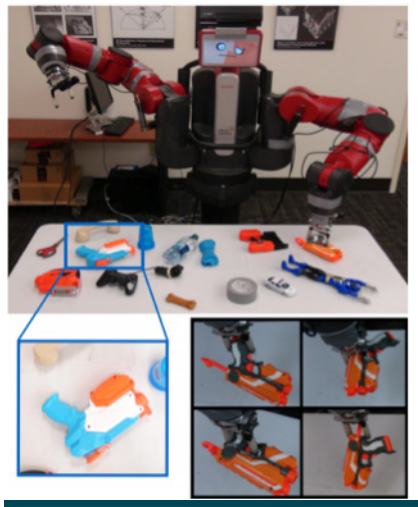


Figure 7 – This robot learned how to grasp simple objects after 50,000 random attempts over 700 hours using deep learning.

Courtesy of Prof. Abhinav Gupta, School of Computer Science, Robotics Institute, Carnegie Mellon University.

can be designed. Social understanding will allow them to successfully reason about how human tasks, goals, and utilities drive their actions. Robots with these visual skills could be used to dramatically improve the situational awareness of environments, scanning the world for people and vehicles whose movements, interpreted using visual common sense, might be reasonably determined as affecting the robot's actions.

There are also significant opportunities to improve and expand visual learning through a robot actively exploring its environment. First, a robot might be able to learn a recognition model for an object class with far less supervision than is currently required. In the future, a robot might be told the class identity of an object it sees, and it could then independently move while tracking that object to collect many more views of the object with no explicit manual labeling. How might a robot accomplish this? While computer vision systems can currently learn what an object class looks like by passive exposure to large sets of images from an object class, determining the so-called "affordances" of objects would most naturally be done through active interactions of the robot with the physical world. Affordances capture the possible uses of an object such as whether an object can be opened (a refrigerator, a door, a soda can), or not (a baseball or tree). Determining affordances of objects is what will eventually enable robots to flexibly reach goals across varying environments.

Q4: How many road vehicles in this image?
A4: Three.
R4: There are two cars and one truck.

Figure 8 – By adaptively tapping into Wikipedia, a VQA system not only answers the question, but explains the answer.

Prof. Anton can den Hengel, Department of Computer Science University of Adelaide, Australia

It is well-established that humans and animals operate under intuitive physical models rather than exact physics models. Humans can predict what happens when they push a mug filled with coffee off a table or when they try to catch a ball. The power of these intuitive physics models is not in modeling the physics of a situation exactly (which is often intractable), but in the fact that they are grounded in real world experience. Right now, in robotic vision labs around the country, robots are interacting with simple physical worlds on a 24/7 basis with no human supervision to learn intuitive physics models (Figure 6). Even with the mere application of random force and grasp attempts in tabletop worlds, these robots are learning not just what things look like, but how they react to forces (balls move with only small forces) and how they interact with one another. As these models and learning algorithms mature, they will form a foundation for robotic vision systems to acquire more complete intuitive physics models of how actions affect the world.

Effective human/robot interaction also requires advances in the integration of vision with language. This interaction should ideally involve robots that understand what a human intends from listening to what that person says and watching the world that they share. Object (noun)/attribute (adjective) recognition is the simplest example of such integration. When a human refers to a blue car in an image, the robot should be able to determine its location to resolve the reference. Of course language contains much more than just nouns and adjectives, so a human might refer to the blue car in front of the building entrance to distinguish it from another blue car in the scene. It is therefore critical for robots to be able to correctly explain to people what they see, and for people to be able to pose questions about what robots see.

Figure 9 – When asked what the man in the photo is doing, the VQA system answered "surfing".

Prof. Fei Fei Li, Computer Science Department, Stanford University

Researchers in computer vision and computational linguistics have begun to study these problems over the past few years. For example, models **like Visual Question Answering (VQA)** involve integrating vision, language, and common sense knowledge (Figure 8). Given an image, and a free-form natural language question about the image, VQA might attempt to answer questions like "What kind of store is this?", "How many people are waiting in the queue?", or "Is it safe to cross the street?" The machine's task is to automatically produce a concise, accurate, free-form, natural language answer ("bakery", "5", "Yes"). VQA represents not a single narrowly-defined problem (e.g., image classification) but rather a rich spectrum of semantic scene understanding problems (Figure 9).

A question may lie at a different point along the semantic scene understanding spectrum. Questions may directly map to existing well-studied computer-vision problems ("What is this room called?" is an example of indoor scene recognition) all the way to those that require an integrated approach of vision, language, and reasoning

over a knowledge base ("Does the pizza in the back row next to the bottle of Coke seem vegetarian?"). Unlike previous computer vision systems, VQA attempts to answer a question that is not determined until run time. For example, in traditional visual surveillance systems, the questions to be answered by an algorithm are predetermined and only the video changes. In contrast, VQA's question form is unknown, as is the set of operations required to answer it. This brings us much closer to general image understanding.

VQA typically requires processing both visual information (the image) and textual information (the question and answer). One approach to Vision-to-Language problems, such as VQA and image captioning, is based on a method pioneered in machine language translation. The direct approach develops an implicit vector space encoding of the input text using a Recurrent Neural Network (RNN) and passes it to another RNN for decoding. This differs from building a high-level representation of the text and re-rendering that high level understanding into a different language. The significance of this method is that it does not form an explicit representation of the meaning of the text and it performs well in practice. In contrast, other implicit approaches to VQA actually go to web sources like Wikipedia, to acquire the explicit knowledge needed to answer a question about an image. Such systems can also explain the basis for their answer. This is crucial for establishing trust between a human and a robot. This approach to Vision-to-Language problems will lead to robotic assistants for people that can be much more generally tasked than current systems. For example, a person might ask a surveillance robot to inform her if a person is seen leaving a building carrying anything related to entertainment. The robot could then interrogate the web to learn that televisions, radios, and musical instruments are possible reasons for alerting her.

Answering questions about images naturally requires task-dependent processing, or processing that requires different visual processes to achieve different goals.

Semantic image interpretation typically requires an extended process directed to specific objects and relations in a task-dependent manner. For example, what is person X looking at or touching, or is object Y stable? One potentially productive line of future research would be the combination of probabilistic inference with policy learning to generate an appropriate sequence of operations applied to the image. The first stage could construct, in a bottom up manner, an initial interpretation of the scene. The second could generate and apply an interpretation in a task dependent manner in which different processes can be synthesized in response to different queries. This is critical for robotic systems, because a robot cannot sense its entire environment at a resolution sufficient to answer any arbitrary question about it.

Progress on developing such task dependent processing models can be accelerated by advances in understanding human intelligence, in particular, the role of feedback signals in human and computer visual perception. Most visual recognition models, both computational and neurophysiological, include only feedforward paths despite the fact that a very large percentage of connections in the visual cortex are feedback connections (Figure 10). Research on probabilistic logical models (Markov Logic Networks) and other generative graphical models represent promising directions for integrating goals with feedforward processes in computational systems. But the role of feedback in both human and computer vision systems is not yet well understood and its unraveling will be critical to future vision systems. A better understanding of human visual learning would generally have a disruptive, yet positive, impact on the design of computational learning models.

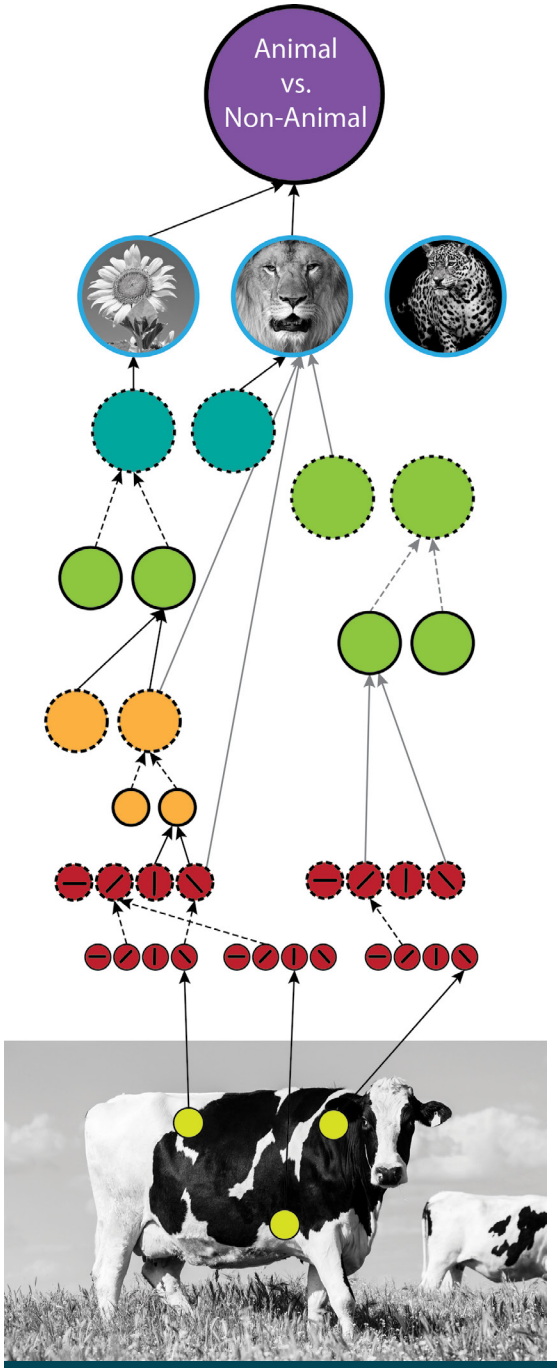
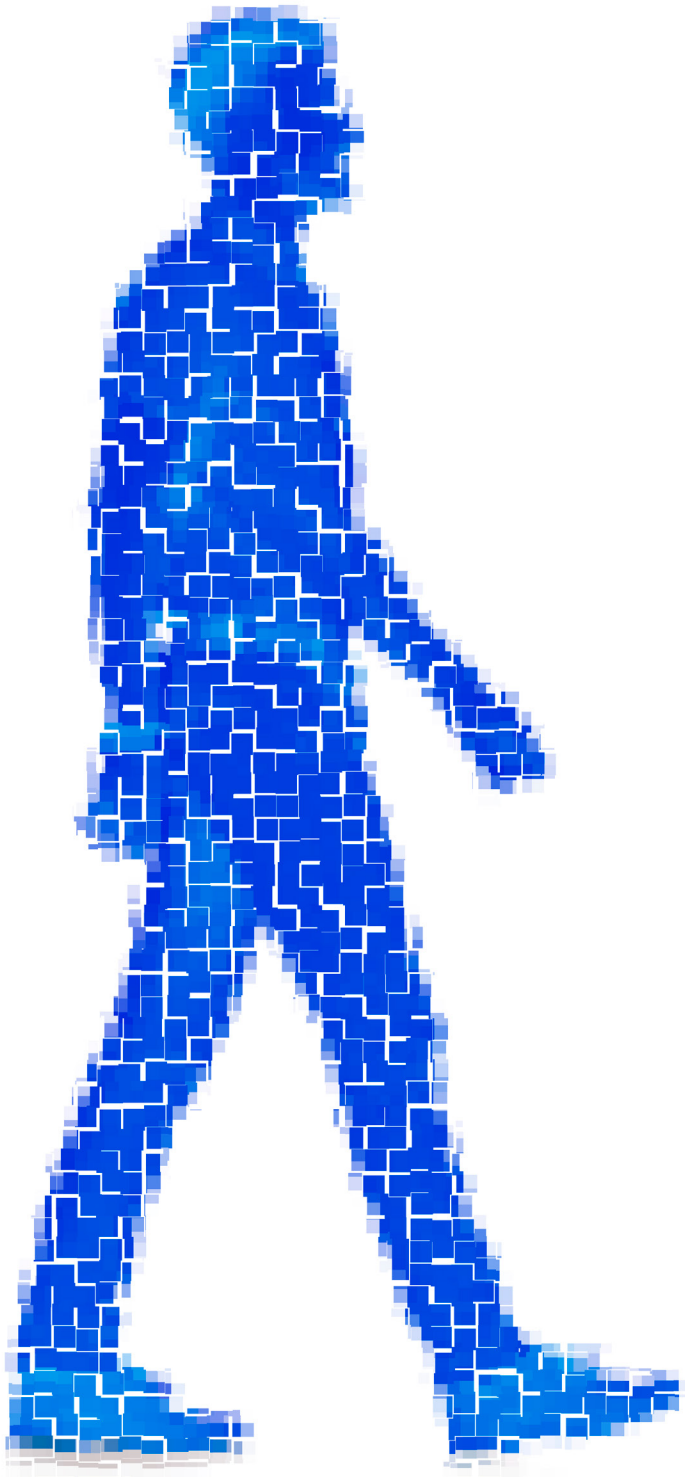


Figure 10 – A biologically inspired feedforward visual recognition model.



Conclusions

Progress in computer vision and machine learning, coupled with technological advances in computer and software systems, has resulted in dramatic advances in the ability of computer vision systems to locate, identify, and track objects in images and videos has vastly improved. The Computer Vision workshop gathered experts to discuss these advances, the trajectory of the field over the next few decades and review the challenges to achieve these goals.

In summary, the participants consider the key research goals for the computer vision field to be:

- Achieving a more fundamental understanding of when and why complex learning models like deep learning converge and how they can be integrated with other forms of reasoning.
- Learning how to train visual recognition models with far less manually annotated training data for both passive (by watching online video sources, for example) and active training (using robots that continuously explore the world and learn how to see better through that experience).
- Providing computer vision systems with common sense knowledge—including intuitive physics knowledge and knowledge about human activities and interactions (social intelligence)—so that they can go beyond visual recognition to a deeper understanding of human intentions and goals.
- Creating a new generation of robotic systems that can intelligently interact with humans to solve problems; this will involve addressing problems in integrating language understanding with vision, developing more powerful models for top down control from task specifications and feedback for vision systems, and the design of new methodologies to assess progress in the field that go beyond static databases of images and videos.
- Developing new models of visual learning and control based on advances in brain and cognitive science.

Participants agree that these advances will have a profound influence on intelligent autonomous systems with both high social and military impact including:

- Surveillance systems that, rather than being pre-programmed to perform specific surveillance tasks, can acquire knowledge from online sources to perform novel surveillance tasks.
- Systems that monitor human behavior for safety and security that can not only recognize the activities that a person is engaged in but, based on common sense knowledge and context, understand whether these activities should result in intervention.
- Robots that can work with humans on a wide variety of tasks because they have a basic understanding of human intentions and goals, personalized to the individual in control of the robot.

Continued progress depends on sustained basic research efforts with more cross-disciplinary approaches and a strong commitment to improved infrastructure. These efforts establish a strong foundation that will propel computer vision closer to the capabilities of human vision.

Glossary

Deep learning – class of machine learning algorithms that use multiple layers of nonlinear processing units to learn task-specific features at different levels of scale and abstraction. They have been successfully applied to many fundamental problems in computer vision including face recognition and object detection in still images, and semantic labeling in still images and video. 3

Semantic labeling – given an image or a video and a set of semantic class labels (e.g., road, building, person), semantic labeling assigns a unique class label (or unknown) to each pixel in the image or video.

Structure from Motion – a computer vision ranging method—given a set of two dimensional images possibly coupled with local motion signals, construct the geometry of e three dimensional world covered by those images. 5

Deformable part model – a learned decomposition into parts and their geometric statistics of an object category from a set of training samples. 8

Object detection – Given an image and an object class, determine whether the image contains an instance of that object class, and if so, delineate each instance. 7

Image classification – given an image of a single object viewed against a possibly complex background, identify the category of the object. 7

Scene classification – given an image, identify the type of location (urban, beach, interior of a school) where the image was taken.

Bundle adjustment – refining a visual reconstruction to produce jointly optimal 3D structure and viewing parameter (camera pose and calibration) estimates. 8

Visual attribute – a property of an object or an action that can be recognized visually—for example, the color of a car or the intensity of a facial expression.

Image segmentation – decomposition or partitioning of an image into regions that, with high probability, correspond to objects and their parts.

Appendix I

Compressive Sensing Researchers

Pedro Domingos – *http://homes.cs.washington.edu/~pedrod/vita.pdf*
University of Washington, *pedromdd@gmail.com*
Department of Computer Science & Engineering
PhD (1997) Information & Computer Science, University of California Berkeley

A former ONR Young Investigator Award winner, Dr. Domingos' research interests include machine learning and data science. He is an SIGKDD Innovation and NSF Career award winner, Sloan Fellowship scholar, editorial member of the Machine Learning journal, and the co-founder of the International Machine Learning Society.

Kristen Grauman – *http://www.cs.utexas.edu/~grauman/grauman_cv.pdf*
University of Texas, *grauman@cs.utexas.edu*
Department of Computer Science
PhD (2006) Computer Science & Artificial Intelligence, MIT

A former ONR Young Investigator Award winner, Dr. Grauman is also the recipient of the Presidential Early Career Award for Scientists and Engineers (PECASE) and a Sloan Fellowship award. Her research interests include applications of information retrieval within the field of computer vision and machine learning, object and activity recognition, image search, and large-scale retrieval.

Brian Scholl – *http://www.yale.edu/perception/Brian/misc/bjs-CV.pdf*
Yale University, *brian.scholl@yale.edu*
Department of Psychology
PhD (1999) Cognitive Psychology, Rutgers University

Brian Scholl is a recognized distinguished faculty member at Yale University. His research interests include tackling difficult questions that address how the mind constructs conscious visual awareness, how seeing relates to thinking, and how visual systems represent objects.

Derek Hoiem – *http://dhoiem.cs.illinois.edu/*
University of Illinois, *dhoiem@uiuc.edu*
Department of Computer Science
PhD (2007) Robotics, Carnegie Mellon University

Dr. Hoiem has recently earned the Intel Early Career Faculty award winner, is an ACM Doctoral Dissertation Award honorable mention, and Sloan scholar. His research goal is to "model the physical and semantic structure of the world, so computers can better understand scenes from images."

Larry Zitnick – *http://larryzitnick.org/*
Facebook, *zitnick@fb.com*
Facebook AI Research
PhD (2003), Robotics, Carnegie Mellon University

Dr. Zitnick is currently pursuing applied research projects in computer vision at Facebook's AI Research division. His projects involve object recognition, language and vision, and methods for gathering commonsense knowledge. He has given numerous talks at MIT, Stanford, Berkeley, and DARPA on the topic of advancing the image recognition capability of machines.

Stefanie Tellex – *https://www.linkedin.com/in/stefanie-tellex-38468818*
Brown University, *stefie10@cs.brown.edu*
Department of Computer Science
PhD (2010) Media Arts & Science, MIT

Dr. Tellex's research goal is to eventually "construct robots that seamlessly use natural language to communicate with humans." She has published at SIGIR, HRI, RSS, IROS, and ICMI, winning Best Student Paper at SIGIR and ICMI. She was also named IEEE Spectrum's AI's "10 to Watch" and has won the Richard B. Solomon Faculty Research Award at Brown University.

Dieter Fox – *https://homes.cs.washington.edu/~fox/*
University of Washington, *fox@cs.washington.edu*
Department of Computer Science & Engineering
PhD (1998), Doctor of Philosophy, Universitat Bonn (Germany)

Dr. Fox's research interests are in robotics, AI, and state estimation. He is the head of the Robotics and State Estimation Lab (RSE-Lab) and currently serves as the academic PI of the Intel Science and Technology Center for Pervasive Computing (ISTC-PC). He is a current fellow of AAAI, IEEE, and is the editor of the IEEE Transactions on Robotics.

Hal Daumé III – *http://www.umiacs.umd.edu/~hal/*
University of Maryland, *hal@cs.umd.edu*
Department of Computer Science
PhD (2006), Computer Science, University of Southern California

Dr. Daume's primary research interests are in developing new learning algorithms for prototypical problems that arise in the context of language processing and artificial intelligence. He is a regular attendee of ACL, ICML, NIPS, and EMNLP. He is currently director of the CLIP lab at the University of Maryland.

Aude Oliva – *http://cvcl.mit.edu/audeoliva.html*
Massachusetts Institute of Technology, *oliva@csail.mit.edu*
Computer Science and Artificial Intelligence Lab
PhD (1995), Cognitive Science, Institut National Polytechnic of Grenoble (France).

Dr. Oliva's research interests are very cross-disciplinary. They span human perception/cognition, computer vision, and cognitive neuroscience, focusing on research questions at the intersection of these three aforementioned domains. Her work is regularly featured in the scientific and popular press, in museums of Art and Science, as well as in textbooks of Perception, Cognition, Computer Vision, and Design. She is the recipient of a National Science Foundation CAREER Award (2006) in Computational Neuroscience, an elected Fellow of the Association for Psychological Science (APA), and the recipient of the 2014 Guggenheim fellowship in Computer Science.

Pushmeet Kohli – *http://research.microsoft.com/en-us/um/people/pkohli/*
Microsoft Corporation, *pkohli@microsoft.com*

Dr. Kohli's reseach concerns the development of intelligent machines – to "teach" computers to (1) understand the behavior and intent of humans, and (2) to correctly interpret objects and scenes depicted in color/depth of images or videos. Kohli is currently working on several projects at Microsoft including "Deep Interpretable Models for Visual and Conversational Data," "Probabilistic Programming for Perception," and "Generative Models for Adaptive Crowdsourcing and Aggregation." He has been the recipient of several "Best Paper" awards and his work is regularly featured in such popular publications as WIRED, the BBC, and Scientific Computing. Recently, Dr. Kohli's work was featured in the Neural Information Processing Systems (NIPS) 2015 conference.

Abhinav Gupta – *http://www.cs.cmu.edu/~abhinavg/*
Carnegie Mellon University, *abhinavg@cs.cmu.edu*
The Robotics Institute & School of Computer Science
PhD (2009) Doctor of Philosophy, University of Maryland

Dr. Gupta seeks to answer three critical questions with his research: (1) How do we represent the visual world? (2) What is the link between language and vision? (3) How are actions and objects related to each other? His work on the Never Ending Image Learner (NEIL) was featured in Discover magazine, the BBC, WIRED, and Forbes. CNN included it in an article titled the "Top 10 Ideas of 2013."

Jitendra Malik – *http://www.eecs.berkeley.edu/Faculty/Homepages/malik.html*
University of California Berkeley, *malik@eecs.berkeley.edu*
University of California Berkeley, Electrical Engineering & Computer Sciences
PhD (1985), Stanford University, Computer Science

Dr. Malik is the Arthur J. Chick Professor of EECS at the University of California Berkeley - a department he chaired in 2004-2006. His research group has worked on many different topics in computer vision, computational modeling of human vision, computer graphics, and the analysis of biological images. Additionally, his research group is responsible for many well-known concepts and algorithms such as anisotropic diffusion, normalized cuts, high dynamic range imaging, and shape contexts. He has personally mentored more than 50 PhD students and postdoctoral fellows and is the recipient of numerous awards, including most recently, the PAMI-TC Distinguished Researcher in Computer Vision Award (2013) and the K.S. Fu Prize from the International Association of Pattern Recognition (2014).

Tomaso Poggio – *https://mcgovern.mit.edu/principal-investigators/tomaso-poggio*, *http://cbcl.mit.edu/publications/tomasopoggio.pdf*
Massachusetts Institute of Technology, *tp@ai.mit.edu*
Department of Brain and Cognitive Sciences
PhD (1970), Physics, University of Genoa

Dr. Poggio's research is in the development of computational models of brain function in order to understand intelligence and build intelligent machines that can mimic human performance. With David Marr, he introduced the seminal idea of levels of analysis in computational neuroscience. He introduced regularization as a mathematical framework to approach the ill-posed problems of vision and—more importantly—the key problem of learning from data. He has contributed to the early development of the theory of learning—in particular introducing the mathematics of radial basis functions (RBF), supervised learning in reproducing kernel Hilbert spaces (RKHSs) and stability. In the last decade, he has developed an influential quantitative model of visual recognition in the visual cortex, recently extended in a theory of sensory perception. He is one of the most cited computational scientists with contributions ranging from the biophysical and behavioral studies of the visual system to the computational analyses of vision and learning in humans and machines.

Song-Chun Zhu – *http://www.stat.ucla.edu/~sczhu/*
University of California Los Angeles, *sczhu@stat.ucla.edu*
Departments of Statistics and Computer Science
PhD (1996) Harvard University

Dr. Zhu is the Principal Investigator at UCLA's Center for Vision, Cognition, Learning, and Autonomy (VCLA). The objective of this lab is to pursue a unified framework for representation, learning, inference and reasoning, and to build intelligent computer systems for real world applications. For his work on computer vision, Dr. Zhu has been awarded or nominated for numerous prestigious awards, including the Marr prize (2003) and the Aggarwal prize (2006). In addition to these awards, Dr. Zhu is a Sloan and Harvard fellow. He has served as Vice-Chair for the IEEE Computer Society.

Alex Berg – *http://cs.unc.edu/people/alex-berg/*
University of North Carolina, *alex.c.berg@gmail.com*
Department of Computer Science
PhD (2005), Computer Science, University of California Berkeley

Dr. Berg's research concerns computational visual recognition. He has worked on general object recognition in images, action recognition in video, human pose identification in images, image parsing, face recognition, image search, and machine learning for computer vision. He co-organizes the ImageNet Large Scale Visual Recognition Challenge, and has co-organized a series of workshops on large scale recognition in computer vision. Dr. Berg is the recipient of the NSF CAREER award and the Marr prize (2013).

Government Observers

Yang Jie
National Science Foundation, *jyang@nsf.gov*

David Han
Office of the Assistant Secretary of Defense for Research and Engineering (Basic Science), *david.k.han.civ@mail.mil*

Jiwei Lu
Office of the Assistant Secretary of Defense for Research and Engineering (Basic Science), *jiwei.lu.civ@mail.mil*

Behzad Kamgarparsi
Office of Naval Research, *behzad.kamgarparsi@navy.mil*

David Montgomery
United States Air Force, *david.w.montgomery61.ctr@mail.mil*

Rapporteurs

Brian Hider, Project Manager
Virginia Tech Applied Research Corporation, *brian.hider@vt-arc.org*

Thomas Hussey, Senior Consultant
Virginia Tech Applied Research Corporation, *twhussey@flash.net*

Kate Klemic, Research Scientist
Virginia Tech Applied Research Corporation, *kate.klemic@vt-arc.org*

Workshop Chairs & Report Authors

Larry S. Davis – *https://www.umiacs.umd.edu/people/lsd*, *https://www.cs.umd.edu/people/lsdavis*
University of Maryland, *lsd@umiacs@umd.edu*
Department of Computer Science
PhD (1975) University of Maryland

Larry S. Davis is a Professor in the Institute for Advanced Computer Studies and the Department of Computer Science at the University of Maryland. His research focuses on object/action recognition and scene analysis, event and modeling recognition, image and video databases, tracking, human movement modeling, 3-D human motion capture, and camera networks. He was the Chair of the University of Maryland's Department of Computer Science from 1999-2012 and is a fellow for both the International Association for Pattern Recognition (2002) and the Institute of Electrical and Electronic Engineers (1998).

Devi Parikh – *https://filebox.ece.vt.edu/~parikh/*
Virginia Tech, *parikh@vt.edu*
Department of Electrical and Computer Engineering
PhD (2009, Electrical and Computer Engineering, Carnegie Mellon University

Devi Parikh leads the Computer Vision Lab at Virginia Tech and is also a member of the Virginia Center for Autonomous Systems (VaCAS) and the VT Discovery Analytics Center (DAC).

Her research interests include computer vision, pattern recognition, AI, and visual recognition problems in particular. Her recent work involves leveraging human-machine collaboration for building smarter machines, and exploring problems at the intersection of vision and language. She has also worked on other topics such as ensemble of classifiers, data fusion, inference in probabilistic models, 3D reassembly, barcode segmentation, computational photography, interactive computer vision, contextual reasoning, hierarchical representations of images, and human-debugging.

She is a recipient of an NSF CAREER award, a Sloan Research Fellowship, an Office of Naval Research (ONR) Young Investigator Program (YIP) award, an Army Research Office (ARO) Young Investigator Program (YIP) award, an Allen Distinguished Investigator Award in Artificial Intelligence from the Paul G. Allen Family Foundation, three Google Faculty Research Awards, an Outstanding New Assistant Professor award from the College of Engineering at Virginia Tech, and a Marr Best Paper Prize awarded at the International Conference on Computer Vision (ICCV).