As humans, we have a remarkable ability to perceive the world around us in minute detail purely from the light that is reflected off it – we can estimate material and metric properties of objects, localize people in images, describe what they are doing, and even identify them! Automatic methods for such detailed analysis of images are essential for most human-centric applications such as image search, organization of personal photo collections; and large scale analysis of the content of media collections for market research, advertisement, and social studies. For example, in order to shop for shoes in an on-line catalogue, a system should be able to understand a shoe style, the length of its heels, or the shininess of its material. In order to support visual demographics analysis for advertisement, a system should be able to not only identify people in a scene, but also to understand what kind (style and brand) of clothes they are wearing, whether they have any accessories, and so on.

Despite several successes, such detailed recognition is beyond the current state of the art computer vision systems. This is a challenging task and to make progress we have to make advances in many fronts; human-computer interaction to enable construction of datasets and benchmarks to evaluate various recognition tasks; computational models that can enable such rich recognition; study of natural language as a means to extract the semantics of “human understanding” that is exposed to us via people’s descriptions of their visual experiences; and machine learning algorithms that can learn from large amounts of data in a computationally efficient manner. My research touches upon some of these aspects of the problem, including:

- **Supervised methods for detailed visual recognition.** Together with collaborators, we have developed a novel representation called “poselets” that use more detailed supervision during learning than traditional approaches and provide a basis for building high-level recognition systems. Our methods are currently the state of the art for person detection [1], segmentation [3], and estimating fine-grained properties such as pose, action [12], gender, clothing style and other attributes [2].

- **Machine learning for computer vision.** This includes learning algorithms that offer better tradeoffs between representation power and efficiency of the classifier, leading to orders of magnitude savings in memory and training time on large-scale data [9], efficient classification techniques that are exponentially faster for commonly used classifiers in computer vision [10], techniques to speed-up object detection in images using part-based models [9], and methods for efficiently combining information from high-level and low-level image analysis for semantic segmentation [14].

- **Collecting annotations via crowdsourcing**. These include software tools [5] to collect and curate annotations enabling richer models and benchmarks for image understanding [2] [4] [12], and intuitive interfaces for collecting annotations that minimize annotator bias [6] [13].

- **Models that combine language and vision** for analyzing videos and transcripts to automatically tag faces [8] and modeling human responses in image description tasks to discover discriminative properties of objects [6].

The explosion of imagery on the web due to cheap availability of sensors, computation, and storage, combined with the ability to collect large amounts of images labeled and verified by real people via crowdsourcing, is enabling a new direction of computer vision research for building richer models for detailed visual recognition. This presents new scientific challenges that have to be solved to allow the next generation applications for security, surveillance, robotics, and human-computer interaction. I present an overview of the two main themes of my research that stem from interacting with humans and large datasets.

\(^1\)Services like Amazon Mechanical Turk [https://www.mturk.com]
Machine learning for visual recognition

Computer vision provides a rich source of problems which can often be tackled best using statistical methods that can learn from data. However, as more and more training data becomes available, training and evaluating such systems starts to become a bottleneck. This is especially true for object detectors as they face the complexity of searching over various locations in an image that score well according to a classifier. Traditionally, that had limited the kinds of classifiers to boosted decision trees or linear SVMs because of their efficiency of classification, even though linear SVMs are not the most accurate for many image classification tasks.

My work on efficiently evaluating additive kernel SVMs has made a large class of kernels commonly used in computer vision, including the “spatial pyramid match”, “pyramid match” and various $\chi^2$ kernels, very efficient – classification time and memory complexity is same as that of a linear SVM \cite{10, 11}. This result had a significant impact (cited over 295 times) in the computer vision community and has led to many state of the art systems (including some of our own) such as:

- Improvements over the state of the art pedestrian detector (of Dalal and Triggs) on INRIA pedestrian dataset by allowing non-linear classifiers for detection \cite{10}. Speedups of over two orders of magnitude for the state of the art image classification method based on spatial pyramid matching on Caltech-101/256 datasets \cite{10}, the most widely used dataset at the time of publication.

- Several top-performing methods in some of the most challenging benchmarks in computer vision and multimedia for image classification and detection, such as PASCAL VOC, ImageNet and TRECVID datasets, use our efficient classification algorithm to evaluate their classifiers. These methods combine several features corresponding to shape, color and texture cues with an additive kernel SVMs.

- Although testing speed and accuracy are important factors for many applications, training time can sometimes dictate the choice of classifiers. Our analysis has also led to the development of new algorithms for efficiently training commonly used non-linear classifiers in computer vision \cite{9, 7}. For e.g., an additive kernel SVM classifier that previously took several hours to train on a standard detection dataset can be trained in a few seconds using our method.

One theme of my research is building computationally efficient models for learning and inference for problems that arise in computer vision. For example in \cite{9}, we propose a method to improve the accuracy and speed of object localization in images by casting a voting based method (Hough transform) as a learning problem. This led to the best results on ETHZ shape dataset and person detection in the PASCAL VOC 2008 dataset.

In addition to efficient models, I have looked at ways to model user response in complex image annotation tasks to extract useful information. Issues of language and pragmatics can often lead to unintended biases – for example, part annotation for visual categories such as “sofas” can be noisy since every annotator has a slightly different interpretation of what a “handrest” is. In these settings, models that treat the “true” annotation as a latent variable can be more robust. Some of my research aims to build models that can use such noisy supervision during learning \cite{13}. Another recent project aims to analyze human responses in description tasks to discover directions of visual variability in images \cite{6}. Future directions of research involve tighter integration of human input and learning in order to accomplish various computer vision tasks.
Detailed recognition of visual categories

Although current computer vision systems have been very successful in detecting faces, optical character recognition, and estimating the 3D structure of a scene from large collections of photos, they are still far from the human ability to perform detailed image understanding. Two challenges in building such systems are: (1) design of models that can learn both from rich annotations and large amounts of unlabeled data; and (2) collection of such annotations in a cost and time effective manner. I will discuss briefly our attempts at solving these problems.

Studying the role of supervision in computer vision

Typical computer vision systems involve a search over a large space spanned by the choice of features, learning methods, models, etc., so a “black-box” approach for learning can be unproductive. Rich supervision can enable feedback and an informed search over these choices by allowing intermediate tasks that can be tested and improved in isolation. These tasks can also provide a basis for building mid-level representations between pixels and high-level semantics to enable a variety of high-level recognition tasks.

Our work on “poselets” [1] tries to achieve this for object recognition by using rich annotations, somewhat contradictory to the favored approach of using large amounts of unsupervised or semi-supervised data. Poselets are easy to detect patterns in images that have a semantic meaning (Figure 1), unlike corners or parallel lines. These correspond to faces or entire bodies, but are not just restricted to anatomical parts. For e.g., a poselet can also refer to a pattern that is the conjunction of the head and shoulder. Additional supervision of the form of locations of various joints of the human body enables us to discover these parts. It provides an underlying representation of images on which we have build state of the art systems for person detection [1], segmentation [3], pose estimation, action classification [12], and attribute recognition [2] (Figures 2-6).

More supervision can also make learning easy and interpretable. Our state of the art human attribute recognition system uses a feed-forward neural-network like architecture where every layer is supervised [2] (Figure 6). Such architectures can help answer questions like “what features are most useful for discriminating hair styles?”, or “what parts are most revealing for gender recognition?” (Figure 7), thus providing directions for focussing future research.

Supervision also arises naturally in an interactive system. Some of my research has explored ways to efficiently combine top-down information provided by users, with bottom-up information from low-level image features for semantic segmentation [14].

Future work involves building models for analyzing diverse visual categories such as architectural scenes and texture, and ways of combining segmentation and detection approaches for better visual recognition.
Crowdsourcing for computer vision

I have developed software tools [5] to collect annotations at a large scale that has propelled research on supervised methods for object detection such as poselets [1]. It has also enabled datasets for evaluating semantic contour detection [4], attribute recognition [2], action classification and pose estimation [12].

Two of my ongoing research projects aim to build better user-interfaces to obtain supervision to bootstrap hard vision problems. The first is a way of collecting annotations to enable part and layout discovery that avoids deciding on the semantics of the parts ahead of time. It is based on collecting annotations via pairwise correspondence between instances of a category [13]. The second is a novel interface that forces annotators to describe objects in more detail than they normally would. Exploiting the same idea of paired comparison, the interface consists of randomly pairing each image with another within the category and asking people to describe the differences, which can be analyzed to discover a lexicon of parts, modifiers and their relations [6]. Future work involves extending these ideas to discover attributes for a taxonomy of object categories in a coarse to fine manner and analyzing descriptions of images on the internet to discover visual attributes.

Conclusions and outreach

The task of detailed and fine-grained visual recognition fascinates me and successful methods for doing so can tremendously improve human interaction with machines for navigating large amounts of visual data. I am interested in both the computer vision and machine learning problems that arise from this, as well as the aspects that touch upon areas such as psychophysics, cognitive science, language processing, data visualization and HCI.

As a community we have just started to look at this problem and very few benchmarks exist. At a recently held six-week workshop at the CLSP center at Johns Hopkins university [2] that I helped organize (along with researchers from Oxford, JHU, École Centrale and UC Berkeley), we have started to study the role of supervision for building better and interpretable models for detailed object detection. To encourage research in this area, we have collected a large dataset of several thousands of airplanes, completely annotated with attributes, segmentations and part bounding boxes, which can provide a test bed for such research. We have also started to do a similar analysis for texture to build models that describe texture in detail. We plan to release this dataset and a set of benchmarks to the community soon, as well as extend our annotations to other categories.

I collaborate with researchers from around the world (Oxford, École Centrale, UC Berkeley, Stony Brook, UCSD and TTI Chicago) and have worked with many others. I have organized a tutorial at ECCV 2012 (on additive kernels and explicit embeddings for large scale vision), released code for much of my research and contributed to building datasets and benchmarks for evaluating computer vision algorithms. I have also maintained my ties to the Indian community where I come from, by attending yearly workshops and/or vision conferences held there. I am putting together a team to organize tutorials at ICVGIP 2012 being held in Bombay, India this December.

Bibliography


A complete list of my publications can be found on my website: [http://ttic.uchicago.edu/~smaji](http://ttic.uchicago.edu/~smaji) Citation counts are according to Google Scholar as of November 3, 2012.