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 recognition of images are essential for most *human-centric* applications 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 the style of a shoe, 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 the people in a scene, but also to understand what kind (style and brand) of clothes they are wearing, whether they are wearing any accessories, and so on.

Despite several successes, such detailed recognition is beyond the current computer vision systems. This is a challenging task, and to make progress we have to make advances on several fronts. We need better *representations* of visual categories that can enable fine-grained reasoning about their properties, as well as *machine learning* methods that can leverage 'big-data' to learn such representations. In order to enable benchmarks for evaluating recognition tasks and to guide learning and inference in models that solve challenging problems, we need to develop better ways of *human-computer interaction*. My research touches upon several such themes in the intersection of computer vision, machine learning, and human-computer interaction including:

- **Machine learning for computer vision**, which includes classifiers that offer better tradeoffs between representation power and efficiency, leading to orders of magnitude savings in memory and training time on large-scale data [11]. Other examples are efficient classification techniques that are *exponentially* faster for commonly used classifiers in computer vision [12], techniques to speed-up object detection in images using part-based models [11], methods for efficiently combining high-level and low-level information for semantic segmentation [18], and learning and inference in structured models [6, 7].

- **Rich representations of visual categories** based on novel parts called 'poselets' that are currently the state of the art for person detection, segmentation, and estimation of fine-grained properties such as pose, action, gender, clothing style, and other attributes [1, 2, 3, 14]. Recently, we have developed methods to discover semantic parts and attributes of visual categories from noisy annotations collected via 'crowdsourcing' [9, 17], and built novel *describable* attribute based representations of textures that advance the state of the art in material recognition [4].

- **Advancing computer vision with 'humans in the loop'**, including tools [8] for building large-scale benchmarks for image understanding [2, 5, 14], intuitive interfaces for collecting annotations via 'crowdsourcing' [9, 16], interactive methods for fine-grained recognition [19], and active annotation methods that minimize user effort [15].

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 next generation applications for security, surveillance, robotics, computer graphics, and human-computer interaction. I present an overview of some of my past, ongoing, and future projects that stem from interacting with humans and large datasets for building better computer vision systems.

Machine learning for computer vision

Often, problems in computer vision can be tackled using statistical methods that can learn from 'big data'. However, as 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 and scales in an image to find the object. Traditionally, this bottleneck had limited these classifiers to boosted decision trees or linear SVMs because of their efficiency of classification, even though these 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 that are commonly used in computer vision, including the 'spatial pyramid match', 'pyramid match', and various χ^2 kernels, very efficient – classification time and memory complexity is same as that of a linear SVM [12, 13]. This result had a significant impact (cited over 450 times) in the computer vision community and has led to several state of the art systems. These include improvements over a state of the art pedestrian detector on INRIA dataset by allowing non-linear classifiers for detection and speedups of over *two orders of magnitude* for the 'spatial pyramid matching' method on the Caltech-101/256 datasets [12]. Besides these, 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, use our efficient classification algorithm to evaluate their classifiers. These methods combine several features that capture shape, color, and texture cues, with an additive kernel SVM.

Although testing speed and accuracy are the most important factors for many applications, in practice *training time* can sometimes dictate the choice of classifiers. Our analysis has led to new algorithms for *efficiently training* commonly used non-linear classifiers in computer vision [10, 11]. For example, 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. We also showed that these classifiers are identical to 'generalized additive models' (GAMs) thereby shedding some light on their representation power and enabling efficient learning of GAMs.

A particular emphasis of my research is the computational efficiency of models during learning and inference. Object detection is one such example where efficiency is important. We proposed a method to improve the accuracy and speed of object localization in images by using an efficient part-based voting method [11]. This led to the best results on ETHZ shape dataset and person detection on the PASCAL VOC dataset. Another example is a method to efficiently combine top-down information from object detectors and bottom-up information from low-level image features for semantic segmentation [18]. Our approach is based on a formulation of 'normalized cut' but allows 'biases' to influence the cut. Our analysis showed that solutions to these problems can be computed efficiently in linear time allowing interactive applications.

Beyond binary labels, computer vision problems require prediction of richly structured outputs such as segmentations, parses of human pose, and 3-D geometry estimates. Learning and inference in such cases is a challenge. Recently, we have developed methods that can leverage efficient maximum a-posteriori (MAP) solvers such as 'graph cuts' for learning and inference in a Bayesian framework. This allows better reasoning about uncertainty in the model, as well as learning using task-specific loss functions [6, 7]. Some of my recent work has also looked at analyzing captions using new variants of structured topic models for discovering attributes useful for fine-grained discrimination [9]. Better analysis of language can provide a means of using vast amounts of data from the web as supervision.

Rich representations of visual categories

Although current vision systems are successful in detecting faces, recognizing hand-written text, and estimating the 3D structure of a scene from large collections of photos, they are still far from the human ability of detailed image understanding. To this end, we have developed novel part and attribute based representations of visual categories that allow extraction of rich information from images.

Part-based representations are widely used in computer vision. However, most of these parts are either 'low-level' and lack semantic meaning or are not discriminative. For example, limbs modeled as a pair of lines are often confused with other structures in images. Having semantic parts allows us to answer questions such as 'people with hats on their head', but these parts also need to be discriminative to be able to reliably detect them in images. We have developed novel parts called 'poselets' [1] that achieve both these goals by relying on weak annotations in the form of sparse landmark locations on instances. Poselets are easy to detect patterns in images that have a semantic meaning – these correspond to faces or entire bodies, but are not just restricted to anatomical parts – a poselet can refer to a pattern that is the conjunction of half the head and shoulder (Fig. 1.1). These parts provide an underlying representation of images on which we have built state of the art systems for person detection [1], segmentation [3], pose estimation, action classification [14], and attribute recognition [2] (Fig. 1.2-1.4).

A small amount of extra supervision can make learning easy and interpretable. Our human attribute recognition system resembles deep convolutional neural networks (CNNs) that have recently demonstrated excellent results on image classification, but unlike with typical CNNs, in our model *every layer is supervised* [2] (Fig. 1.6) and can be trained using efficient linear SVM solvers. Moreover, such architectures can help answer questions like 'What features are most useful for discriminating hair styles?' or 'What parts are most useful for gender recognition?', revealing the bottlenecks in these recognition systems (Fig. 1.5). In the future, we aim to investigate the extent of the analogy between our models and deep CNNs.

Beyond objects, textures and patterns are also rich sources of visual information in images. Material recognition is an especially challenging task due to variations in shading and shape of the object, as well as the viewing direction of the observer. Recently, we have developed representations based on novel 'describable attributes' of textures that significantly outperform the state of the art in material recognition on a number of standard datasets [4]. Furthermore, these describable attributes provide a basis for visualizing and querying large datasets of patterns such as textiles and wallpapers (Fig. 1.7a, 1.7b), enabling new computer vision applications.

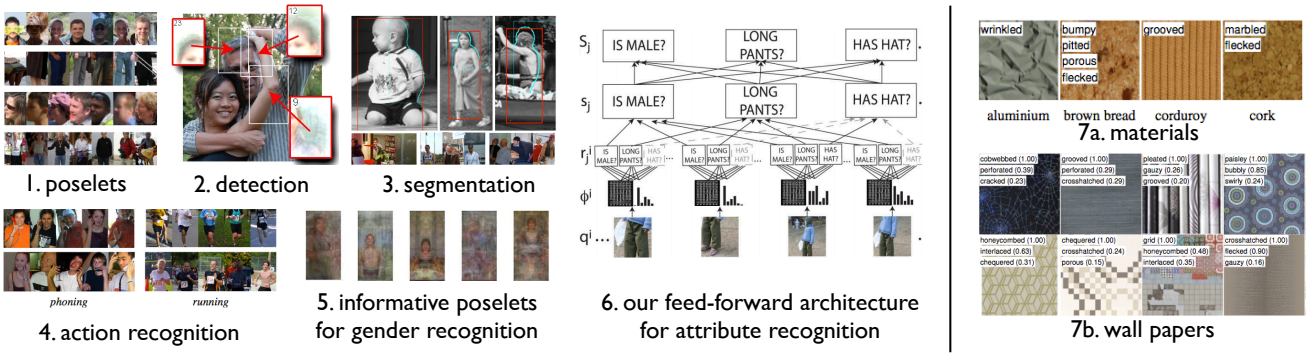

Figure 1: Our part-based representation (Fig. 1.1-1.6) using 'poselets' and their application to detection, segmentation, action, and attribute recognition. Fig. 1.7a, 1.7b overlay several materials and wallpapers with our describable attributes such as 'bumpy', 'wrinkled', 'groved', 'marbled', 'pitted', and 'swirly', that were automatically predicted from the images (best viewed digitally with zoom).

Advancing computer vision with 'humans in the loop'

In the past I have developed software tools [8] to collect *annotations* that have propelled research on rich structured methods for object detection such as poselets [1]. These tools have also enabled datasets for evaluating semantic contour detection [5], attribute recognition [2], and pose estimation [14]. More recently we have developed active learning methods to reduce the annotation effort by leveraging new methods for Bayesian active learning in structured models [15].

Collecting detailed annotations can be expensive and time consuming, which prohibits the construction of large datasets. Two of recent projects aim to build better *user-interfaces* for collecting annotations that reduce the need of expert knowledge and lead to easier annotation tasks. The first is a way of collecting annotations for part and layout discovery that avoids the need for expert specified part labels. It is based on collecting annotations via *correspondences* between pairs of instances within a category [16]. The second is a novel interface for collecting captions that coaxes the annotators to describe objects in detail by asking them to *describe the differences* between two instances [9]. Studying the correspondences and differences is a powerful way to understand the structure of visual categories, and our initial experiments show that such annotations can also be reliably collected via 'crowdsourcing'. In the future, we aim to develop models that can leverage these annotations directly for detailed recognition.

Beyond the use of annotations to guide learning, humans can play a key role during inference. This is exciting as it provides an opportunity for humans and machines to 'collaborate' for solving difficult problems. We are currently developing 'human in the loop' methods for classification that can leverage similarity comparisons provided by the user to improve recognition accuracy [19]. There are a number of applications in areas of search and interactive design that can benefit from such methods.

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 the vision and learning problems that arise from this, as well as aspects that touch upon areas such as psychophysics, cognitive science, language processing, data visualization, and HCI. This is an exciting direction of research with implications for building better tools for creating visualizations, computer graphics, modeling for art and architecture, as well as advancing AI in general.

As a community we have just started to look at this problem, and very few benchmarks exist. At a six-week workshop held last year at the CLSP center at Johns Hopkins university[1] that I co-organized, we started to study the role of supervision for building better and interpretable models for detailed object recognition. To encourage research we have collected a large dataset of several thousands of airplanes, completely annotated with *attributes*, *segmentations*, and *part bounding boxes*. We have contributed this dataset to the fine-grained recognition challenge being held at ICCV 2013. Such supervision provides a means of understanding and designing hierarchical models of recognition.

I currently collaborate with researchers from around the world such as Oxford, École Centrale, UC Berkeley, MIT, UIUC, Stony Brook, UCSD, Caltech, IIIT Hyderabad, and TTI Chicago. I have organized tutorials at leading computer vision conferences, released code for much of my research, and contributed to building datasets and benchmarks for evaluating computer vision algorithms. At TTI Chicago I regularly mentor graduate student interns. I maintain ties to the Indian community where I come from by attending yearly workshops and conferences held there, and by organizing tutorials such as one at ICVGIP 2012 that was held at Bombay, India.

[1] http://www.clsp.jhu.edu/workshops/archive/ws-12/groups/tduosn/

Bibliography

[1] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting People using Mutually Consistent Poselet Activations. In *European Conference on Computer Vision (ECCV)*, 2010.

[2] L. Bourdev, S. Maji, and J. Malik. Describing People: Poselet-Based Approach to Attribute Classification. In *International Conference on Computer Vision (ICCV)*, 2011.

[3] T. Brox, L. Bourdev, S. Maji, and J. Malik. Object Segmentation by Alignment of Poselet Activations to Image Contours. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[4] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing Textures in the Wild. *CoRR*, arXiv:1311.3618, 2013.

[5] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik. Semantic Contours from Inverse Detectors. In *International Conference on Computer Vision (ICCV)*, 2011.

[6] T. Hazan, S. Maji, and T. Jaakkola. On Sampling from the Gibbs Distribution with Random Maximum A-Posteriori Perturbations. In *Neural Information Processing Systems (NIPS)*, 2013.

[7] T. Hazan, S. Maji, J. Keshet, and T. Jaakkola. Learning Efficient Random Maximum A-Posteriori Predictors with Non-Decomposable Loss Functions. In *Neural Information Processing Systems (NIPS)*, 2013.

[8] S. Maji. Large Scale Image Annotations on Amazon Mechanical Turk. Technical Report UCB/EECS-2011-79, EECS Department, University of California, Berkeley, Jul 2011.

[9] S. Maji. Discovering a Lexicon of Parts and Attributes. In *Second International Workshop on Parts and Attributes, ECCV*, 2012.

[10] S. Maji. Linearized Smooth Additive Classifiers. In *Workshop on Web-scale Vision and Social Media, ECCV*, 2012.

[11] S. Maji and A. Berg. Max-Margin Additive Classifiers for Detection. In *International Conference on Computer Vision (ICCV)*, 2009.

[12] S. Maji, A. Berg, and J. Malik. Classification using Intersection Kernel Support Vector Machines is Efficient. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[13] S. Maji, A. Berg, and J. Malik. Efficient Classification for Additive Kernel SVMs. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, Vol. 35, No. 1, January 2013.

[14] S. Maji, L. Bourdev, and J. Malik. Action Recognition from a Distributed Representation of Pose and Appearance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[15] S. Maji, T. Hazan, and T. Jaakkola. Efficient Boundary Annotation using Random Maximum A-Posteriori Perturbations. *Submitted to AISTATS*, 2014.

[16] S. Maji and G. Shakhnarovich. Part Annotations via Pairwise Correspondence. In *4th Workshop on Human Computation, AAAI*, 2012.

[17] S. Maji and G. Shakhnarovich. Part Discovery from Partial Correspondence. In *Computer Vision and Pattern Recognition (CVPR)*, 2013.

[18] S. Maji, N. Vishnoi, and J. Malik. Biased Normalized Cuts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[19] C. Wah, G. V. Horn, S. Branson, S. Maji, P. Perona, and S. Belongie. Similarity Comparisons for Interactive Fine-Grained Categorization. *Submitted to CVPR*, 2014.

A complete list of my publications can be found on my website: http://ttic.uchicago.edu/~smaji