Research Statement

Ross Girshick

1 Introduction

My primary research interests are in computer vision and machine learning. Broadly speaking, the goal of computer vision research is to develop algorithms that describe images. For example, given a photograph of a crowded Parisian cafe, an algorithm should be able to read the cafe's name from its awning, infer its geographic location, count the patrons, describe their poses and clothing, and so on. I want computers to *see* the world.

This grand goal is the motivation behind my work, which focuses on classical problems in computer vision, such as object recognition and human pose estimation. I am driven by a purely scientific desire to solve computer vision, as well as the desire to see its broader impact through practical applications. For example, I believe that advances in computer vision will greatly assist visually impaired users navigate unfamiliar environments. I find this application particularly inspiring.

I am drawn to approaches that can be derived from first principles and to algorithms that have strong theoretical guarantees. Ultimately, a computer vision system must also work well in practice. I have made releasing high-quality code a fundamental research objective. For example, I maintain open source releases of the object detection system that I developed during my Ph.D. This software has been downloaded more than 3,000 times in the last year and is in use by a wide range of research groups (including at MIT, CMU, UIUC, Oxford, and Google).

Computer vision cuts across many research areas. In my work, for example, I make extensive use of existing machine learning tools, in addition to developing new machine learning algorithms. One of my current projects [3] (in review) exemplifies these interactions. In this work, we show how techniques from computer vision (bottom-up image segmentation) and machine learning (convolutional neural networks) can be used in concert with "big visual data" (ImageNet) to yield dramatically better object detection results than previously reported. I look forward to rich collaborations with researchers working in different areas, in particular machine learning and high-performance computing. With hundreds of millions of photos being uploaded every day, these interactions are essential.

2 Research Contributions

Object detection. When you or I look at an image, we immediately extract an enormous amount of semantic and geometric information. For example, you might recognize *your* particular car, or more generally *some* car. The object detection task is to infer the latter type of information. That is, to localize objects belonging to a basic level category. Recognizing cars, generically, requires discriminating them from other non-car objects, while also generalizing over all car images. The challenge of this task lies in the delicate contention between generality and specificity.

I made fundamental contributions towards solving this problem during my Ph.D. by showing that a class of part-based models performs particularly well. Part-based modeling captures the intuitive notion that objects are composed of parts that can be described by their local appearance together with a geometric model governing their spatial layout. For example, people are composed of heads, torsos, arms, and legs with specific geometric constraints imposed by our skeletons. However, most object categories exhibit far too much within-class variation to be represented by a single part-based model. For example, a bicycle viewed from the side and a bicycle view from the front have distinct appearances and yet must be recognized as members of the same class.

In [2] we showed how to solve this problem by learning *mixtures* of part-based models in a discriminative latent support vector machine (SVM) framework. This idea can be further generalized to models defined by a type of context-free grammar. This representation naturally allows an object category to be modeled recursively in terms of parts, subparts, and so on. It also enables parts (or subparts) to have alternative modes of appearance. For example, the "head" part might appear as a front-facing head *or* a side-facing head. We explored this modeling paradigm by building object detection grammars in [5]. We were able to show state-of-the-art performance on the challenging task of detecting people, and even achieved results that are comparable to other approaches that use more supervision during training.

This line of work has had a significant impact on the computer vision community. Our part-based models, known as Deformable Part Models (or DPM), and variations on them proposed by other groups, were consistent winners of the 2008-2012 PASCAL Visual Object Classes Challenge—the premiere yearly challenge in computer vision for comparing object detection algorithms. We were awarded the PASCAL VOC "Lifetime Achievement" Prize for this work. With these results, and open source software [4], DPM is currently the "go to" object detector in use today.

Human pose estimation. Machine perception of the human form occupies a special place in computer vision research because it is fundamental to many applications. One of the most basic inferences to draw from an image of a person is a description of their pose. That is, to give a detailed account of where each of the person's joints is located (either in 2D or 3D). This task is simultaneously forgiving—the human skeleton enforces a strong prior on valid poses—and vexing—the appearance space due to articulated deformation and clothing (amongst other nuisance variables) is vast. In my research, I have made contributions to two versions of this problem: pose estimation in RGB color images and pose estimation in depth (RGB-D) images.

My internship project at Microsoft Research Cambridge lead to a new algorithm for 3D human pose estimation from a single depth image [7]. Our work improved on the previous Kinect pose estimation algorithm (which is in use by millions of Xbox Kinect customers) by directly regressing body joint locations from the visible depth pixels. Our approach uses regression forests to vote for the 3D position of each joint. We showed that learning a confidence value associated with each vote allows most of them to be ignored, which dramatically speeds up vote aggregation. Our approach is 4x faster than the previous Kinect pose estimation algorithm, while also being significantly more accurate, especially on joints obscured by self-occlusion.

In work currently in review, we tackle the 2D pose estimation problem in RGB color images. Our approach, which we call k-poselets, builds an ensemble of deformable part models, where each of the k parts in a single DPM is a *poselet*. A poselet is a part that, by construction, detects and localizes a subset of body joints striking a particular pose. Combining deformable part models and poselets improves both detection and pose estimation.

Efficient inference. Many computer vision algorithms involve an inference step in which an optimization problem is solved in order to infer an unobserved state (e.g., the location of a car and its parts in an image). Detecting objects with a deformable part model, for example, involves solving a discrete optimization problem with a dynamic programming algorithm.

In my master's thesis, and later in [1], we developed a fast inference algorithm for detection with deformable part models. The algorithm is based on the idea of a "cascade" in which a sequence of thresholds is used to quickly prune large portions of the state space. From a dynamic programming point of view, pruning results in sparse dynamic programming tables where most subproblems are left unsolved because they are unlikely to participate in a good solution.

Ideally one could choose thresholds that are "admissible" in the sense that they never prune an optimal solution. However, provably admissible thresholds are typically slow in practice because they are too conservative. As an alternative, we introduced the notion of Probably Approximately Admissible (PAA) thresholds. An algorithm computes PAA thresholds if the probability that their pruning error rate is greater than ϵ is bounded by δ. Intuitively, the PAA property guarantees that the thresholds have a low probability of frequently pruning optimal solutions. In [1], we give a simple data-driven approach that selects PAA thresholds. In practice, our algorithm accelerates DPM detection by more than an order of magnitude.

I have also developed methods that speed up inference by compressing models with sparse coding. In [8, 9] we introduced the notion of a *sparselet*, which arises from applying sparse coding to learned model parameters, rather than to observational data as is typically done.

Learning. Machine learning is ubiquitous in computer vision. My work on object detection has contributed two general techniques for learning structured models with latent variables.

In supervised learning, the predictions made by a model are assumed to be in the same space as the labels provided in the training data (e.g., binary predictions and training labels). However, sometimes the label space is unnatural for a model, especially for one with richly structured outputs. For example, in [5] we trained a person detector that outputs derivation trees using training samples labeled with bounding boxes. Because many possible derivation trees can explain a single bounding box, the training data is weakly labeled with respect to the model's predictions. To handle such learning problems, we introduced an empirical risk minimization framework called weak-label structural SVM that generalizes a number of well-known learning framework, including structural SVM and latent structural SVM. We bridge the gap between these spaces with a loss function that defines how compatible a structured output is with a training label.

In [6], we introduced a new learning algorithm called latent linear discriminant analysis or latent LDA. Latent LDA can be used to train deformable part models instead of latent SVM. The advantage of LLDA over LSVM is speed. Both approaches solve a sequence of convex optimization subproblems, however in the case of LLDA each has a closed-form solution that can be computed quickly. The subproblems that arise in LSVM, in contrast, require an expensive iterative numerical optimization algorithm.

3 Future Work

My future research plans are organized around three themes: learning rich representations from big visual data, unifying object detection with semantic segmentation, and leveraging 3D information to improve recognition.

The notion of learning richly structured representations is central in my work. The visual world is filled with regularities that arise, ultimately, from physics. While it may be possible to solve vision problems by ignoring these regularities, I believe that solutions which exploit them are likely to have more attractive computational properties (e.g., require less training data, run more quickly at test time). In ongoing research [3], we are demonstrating that a large convolutional neural network (CNN) trained on more than a million images can yield previously unattainable object detection performance. These networks directly encode translationally invariant feature extraction in their convolutional structure. In only three months of work we achieved the same relative improvement in performance that I managed over 4 years during my Ph.D.

These detection results are exciting in their own right, but they also establish a new direction of inquiry. Little is understood about how to design the next generation of CNNs. Developing new methods for visualizing and debugging these models will be essential for making progress. To the extent that we can already visualize these networks, it's clear that they learn a rich hierarchy of features starting from simple oriented edge and opponent color filters up to human and animal face detectors. I also believe that there are deep connections with existing compositional part-based models. Both approaches share many of the same building blocks including convolutional filtering and max-response pooling. One hypothesis is that by encoding more prior knowledge of the regularities in the natural world, we'll be able to further improve performance.

"Big visual data" also plays an undeniable role. Large training sets are integral to the success of our method, clearly demonstrating the need to use even more data. However, additional data creates new scaling challenges that will require algorithmic, machine learning, and systems research to resolve. We will also need to devise new strategies for efficiently aquiring image annotations. Current methods are too costly and can only take advantage of a minute fraction of the images uploaded across the Internet each day. Here, there is ample room for interdiciplenary collaboration that might, for example, lead to game-based annotation tools in which players (unwittingly) generate labeled data.

This work has also established a path towards unifying object detection with semantic segmentation. Object detection has traditionally been concerned with roughly localizing objects by placing bounding boxes around them. Semantic segmentation can be thought of as object detection with fine-grained object boundary delineation. The output should be a nearly pixel-accurate mask covering the visible extent of a detected object. These two tasks are typically treated separately and evaluated with subtlely different metrics. I believe that a unified approach to both problems will benefit each more than disjoint solutions will. The CNN detection framework we're developing is making progress towards this goal.

Recognition should also benefit from inferences about the 3D geometry of a scene. In [7], and in ongoing work, we have seen that depth information can make challenging recognition problems more tractable. While depth sensors are becoming more widely available, both RGB and RGB-D modalities will remain important if for no other reason than that we already have billions of depthless images stored online. I find the prospect of leveraging depth in RGB color images, recorded without any form of depth sensor, even more exciting. Naturally, this will require new algorithms for monocular 3D reconstruction. Even if the recovered geometry is coarse, I believe this approach will help computers see the world.

References

[1] P. Felzenszwalb, R. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[2] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2010.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *Tech Report, arxiv e-prints*, 1311.2524 cs.CV, November 2013.

[4] R. Girshick, P. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, release 5. `http://www.cs.berkeley.edu/~rbg/latent-v5/`.

[5] R. Girshick, P. Felzenszwalb, and D. McAllester. Object detection with grammar models. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[6] R. Girshick and J. Malik. Training deformable part models with decorrelated features. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[7] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[8] R. Girshick, H.O. Song, and T. Darrell. Discriminatively activated sparselets. In *International Conference on Machine Learning (ICML)*, 2013.

[9] H. Song, S. Zickler, T. Althoff, R. Girshick, M. Fritz, C. Geyer, P. Felzenszwalb, and T. Darrell. Sparselet models for efficient multiclass object detection. In *European Conference on Computer Vision (ECCV)*, 2012.