High-Level, Part-Based Features for Fine-Grained Visual Categorization

Thomas Berg

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2017

ABSTRACT

High-Level, Part-Based Features for Fine-Grained Visual Categorization

Thomas Berg

Object recognition–"What is in this image?"–is one of the basic problems of computer vision. Most work in this area has been on finding basic-level object categories such as *plant*, *car*, and *bird*, but recently there has been an increasing amount of work in "fine-grained" visual categorization, in which the task is to recognize subcategories of a basic-level category, such as *blue jay* and *bluebird*.

Experimental psychology has found that while basic-level categories are distinguished by the presence or absence of parts (a bird has a beak but car does not), subcategories are more often distinguished by the characteristics of their parts (a starling has a narrow, yellow beak while a cardinal has a wide, red beak). In this thesis we tackle fine-grained visual categorization, guided by this observation. We develop alignment procedures that let us compare corresponding parts, build classifiers tailored to finding the interclass differences at each part, and then combine the per-part classifiers to build subcategory classifiers.

Using this approach, we outperform previous work in several fine-grained categorization settings: bird species identification, face recognition, and face attribute classification. In addition, the construction of subcategory classifiers from part classifiers allows us to automatically determine which parts are most relevant when distinguishing between any two subcategories. We can use this to generate illustrations of the differences between subcategories. To demonstrate this, we have built a digital field guide to North American birds which includes automatically generated images highlighting the key differences between visually similar species. This guide, "Birdsnap," also identifies bird species in users' up-

loaded photos using our subcategory classifiers. We have released Birdsnap as a web site and iPhone application.

Table of Contents

List of Figures

List of Tables

Acknowledgments

After spending a few years programming front-office systems in Tokyo, facing yet another project where the great challenge would be getting our clients to agree on the optimal tab-ordering of fields on the data entry screens, I began to wonder if learning a little computer science might put me in the way of more interesting work. So I convinced my girlfriend (now wife) to "visit" the States and headed back home to go to school. Thank you, Aya, for coming with me. I couldn't have got through this on my own.

With no training in computer science, I didn't interest the graduate schools that interested me, so I began taking classes at Harvard's (open admissions) extension school. Two instructors in particular, Jamie Frankel and David Albert, fed my interest and had enough faith to write the recommendation letters I needed to get in to Columbia. Thank you, Jamie and David, for pushing me along.

At Columbia, I planned on a quick master's degree, then back to work at a higher pay- and interest- grade. But in my first semester, I took some *very interesting classes*. Shree Nayar's and Peter Belhumeur's classes got me interested in computer vision and pattern recognition. Then a summer project with Peter Allen and Corey Goldfeder gave me a taste of research. Thank you, Shree and Peter and Peter and Corey, for showing me the good stuff. And especially thank you Peter, for letting me stick around when I decided I wanted to stay for a PhD.

I couldn't seem to get anything to work my first couple years, but the advice, friendship, and examples of my fellow students kept me going and even made it fun. Thank you Neeraj Kumar for your enthusiasm and patience, Matei Ciocarlie for your welcome, Ollie Cossairt for your quiet excellence and hidden hippie outlook, Austin Reiter for your matter-of-fact

skills and speed, Jiongxin Liu for your discipline, persistence, and singing voice, Hao Dang for your cheer.

With an inexhaustible well of ideas and intuition from Peter, and his unerring nose for implementation bugs, our work began to work. Thank you Peter, for teaching me how to think clearly, how to write, and how to argue with you.

After a few years, with my graduation "imminent", I left New York for California, remaining a Columbia student until I put the final touches on my thesis. These final touches have now taken over two years, during which Daniel Miau has leapt into the breach whenever a server went down or a disk failed. Thank you, Daniel, for being my eyes and hands on campus.

During all my time at Columbia, Anne Fleming has made the administrative bumps as smooth as they could possibly be. Daisy Nguyen has done the same for all my hardware troubles. Thank you Anne and Daisy.

This thesis is by all of us.

For Aya, as everything

Chapter 1

Introduction

Object detection and classification are among the most-studied problems in computer vision. Where are the objects in this image, and what are they? A lot of research effort has gone toward this problem, and a lot of progress has been made.

One measure of this progress is the PASCAL Visual Object Classes (VOC) challenges [Everingham *et al.*, 2014]. In the classification challenge, as administered each year from 2007 to 2012, we are presented with an image and asked to classify it according to the most prominent object it contains: an airplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa, TV / monitor, bird, cat, cow, dog, horse, sheep, or person. The training set consists of 11,540 images across these classes. As shown in Table 1.1, mean average precision of the top-performing method rose from 59.4% in 2007 to 82.2% in 2012 [Everingham *et al.*, 2005 2013]. Although the challenge officially ended in 2012, later work using the 2012 dataset has achieved mean average precision as high as 85.4%, or 94.3% when using additional training data [Everingham *et al.*, 2016].

The heir to the PASCAL VOC challenges is the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), run annually since 2010. In comparison with the PASCAL VOC, ILSVRC has one thousand categories instead of twenty and 1.2 million training images instead of twelve thousand. To account for the possibility of additional, unlabeled objects in the test images, classification is considered correct if any of five class labels pre-

Year	Winning mAP
2007	59.4%
2008	58.6%
2009	66.5%
2010	73.8%
2011	78.6%
2012	82.2%

Table 1.1: Winning mean average precision on the Pascal VOC Classification Challenge (Competition 1) [Everingham *et al.*, 2005 2013]

dicted by the algorithm match the test label. As shown in Table 1.2, this top-5 accuracy has improved each year, from 71.8% in 2010 to 97.0% in 2016 [Russakovsky *et al.*, 2015; Liu *et al.*, 2015; Liu *et al.*, 2016].

As object detection and classification improve, and further gains become harder to come by, it has become more important to explore the errors made by state-of-the-art methods and how these errors can inform further research. In an analysis of PASCAL VOC object detection results from two top methods, Hoiem *et al.* [Hoiem *et al.*, 2012] find the primary cause of false positives, to be "similar category" errors, in which a detector for one class fires on an instance of a similar class, for example, when the horse detector fires on a cow. These are the most common errors on average across all twenty classes, even though some classes (bottle, potted plant, and TV / monitor) have *no* similar classes and therefor *no* errors of this type at all. If we increase the number of categories, we expect this type of error to become more common, as each category will have more similar categories and the difference between similar categories will decrease. An analysis of the best results on the ILSVRC [Russakovsky *et al.*, 2013] confirms this, noting for example that the best method can distinguish dogs from non-dogs with 99% accuracy, but is much less reliable in distinguishing the 120 different dog breeds in the challenge from each other.

This problem of distinguishing very similar categories from each other is the problem

Year	Winning Top-5 Accuracy
2010	71.8%
2011	74.2%
2012	83.6%
2013	88.3%
2014	93.3%
2015	96.4%
2016	97.0%

Table 1.2: Winning top-5 accuracy on the ImageNET classification challenge. For 2015 and 2016 we show the best classification accuracy on the classification and localization challenge, as the separate classification challege was discontinued after 2014.

of *fine-grained visual categorization* (FGVC) and is the topic of this thesis. Examples of fine-grained categorization include recognizing breeds of dog, species of bird, or models of automobile. This thesis focuses on the use of *parts* of the object for fine-grained recognition, for two reasons.

First, we are guided by results from the study of human perception. Psychologists use the terms "superordinate level," "basic level," and "subordinate level" to describe levels in the taxonomic hierarchy with which we label the objects in our environment. While not always perfectly well-defined, in general the basic level is the level at which we most readily recognize and label objects. For example, barring a reason to be more or less specific, we are more likely to say we saw a "car" (basic level) than a "vehicle" (superordinate level) or a "Toyota Camry" (subordinate level). Similarly we say "bird" rather than "animal" or "starling," and "hammer" rather than "tool" or "ball-peen hammer." In these terms, fine-grained categorization is the problem of distinguishing subordinate-level categories of the same basic-level category from each other.

Work in experimental pyschology suggests that the basic level is often defined by the presence or absence of parts [Tversky and Hemenway, 1984]. A car has four wheels, a

bird has a beak and feathers, and a hammer has a head and a handle. Subordinate-level categories, in contrast, are often distinguished by characteristics of their parts: a starling has an orange beak and spotted feathers. If humans perform fine-grained categorization by considering characteristics of the parts, it's appealing to have our algorithms look to the parts as well.

Second, although FGVC is relatively young as a named area of research in computer vision, it has much in common with a very well-studied *instance*-level classification problem: face recognition. Instance recognition is often considered a different problem from FGVC, at one end of a granularity spectrum with basic-level recognition at the other end and FGVC in the middle. But the difficulty in fine-grained and instance-level recognition is essentially the same: small inter-class differences often swamped by intra-class differences, so we consider face recognition (and instance recognition in general) as an example of fine-grained categorization, where the basic-level category is *face* and the subordinate-level categories are individuals. The best methods of face recognition all include finding parts of the face (eyes, nose, etc.) so that corresponding parts of faces can be compared with each other, so it's natural to investigate whether the use of parts is important for fine-grained recognition in other domains as well.

In this thesis, we develop a set of part-based features for fine-grained visual categorization and demonstrate their application to several problems.

After discussing relevant prior work in Chapter 2, in Chapter 3 we consider the problem of face verification, matching faces by identity. Using a set of 95 parts on the face, we design a method for alignment that brings the faces into correspondence while preserving interclass differences, and a method for learning a set of stacked classifiers that distinguish one face from another. In Chapter 4, we generalize this to learn a set of part-based features we call "POOFs" to distinguish subcategories in any domain, and demonstrate the generalized method's effectiveness at classification of human faces and birds. Subcategorization using POOFs closely follows the intuition from experimental psychology, as each feature is built to measure a characteristic of a particular part, and these part-specific features are

combined to build the subcategory classifiers.

In Chapter 5, we consider applications of part-based features beyond automatic recognition, in particular how we can use these features to develop an understanding of the visual domain defined by a basic-level category. Again taking birds as our example, we automatically determine which subcategories (species) are most similar to each other and annotate images to show the key differences that distinguish similar species from each other. Finally, in Chapter 6, we describe the application of these ideas to build Birdsnap, a publicly-available digital field guide to birds, implemented as a web site and an iPhone App. We use the methods of Chapter 5 to build the guide, illustrating the similarities and differences between species, and use the methods of Chapter 4 to perform automatic identification of the birds in users' uploaded photos. This section describes the challenges we encountered when applying our methods to build a real, useful system, and the modifications to our methods that these challenges necessitated.

Chapter 2

Prior Work

2.1 Face Recognition

The "finest" fine-grained categorization is instance recognition, where we must identify individual instances of a class, and the best-studied example of instance recognition is face recognition. So we look to prior work on faces. The full body of work on face recognition is too large to survey here, so we focus on two aspects relevant to our work, alignment and hierarchical classifiers.

2.1.1 Alignment

It is well established that alignment is critical for good performance in face recognition with uncontrolled images ([Gu and Kanade, 2008; Wang *et al.*, 2006; Wolf *et al.*, 2009]). One method often applied is [Huang *et al.*, 2007a]'s "funneling," which extends the congealing method of [Learned-Miller, 2006] to handle noisy, real-world images. These methods find transformations that minimize differences in images that are initially only roughly aligned. Another common technique is to apply a similarity or affine transformation to the images based on the locations of detected fiducial points such as the corners of the eyes and mouth. Due to both their effectiveness and the fact that pre-aligned images for the standard "Labeled Faces in the Wild" (LFW) face verification dataset are publicly avail-