RESEARCH STATEMENT

Ryan Farrell

In recent years, the topic of object detection/recognition has rapidly gained in popularity and is now perhaps the most actively researched topic in computer vision. Object detection algorithms are becoming prevalent in consumer devices such as digital cameras (real-time face detection) and automobiles (pedestrian detection systems for collision avoidance are already available and will be a standard feature on new cars within a few years). Object recognition technology is quickly becoming widespread in smartphone apps; examples include Google Goggles, Amazon Flow and Leafsnap. I believe we are at a 'tipping point' towards the impending ubiquity of computer vision, specifically object recognition, in our everyday lives.

RESEARCH FOCUS

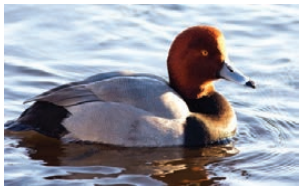
My research in object recognition focuses specifically on Fine-grained Visual Categorization (sometimes abbreviated FGVC). For many years, computer vision has focused on classifying an object in several basic-level categories such as person, car, frog, or piano. At the opposing end of the categorization spectrum (see Figure) is biometric identification - recognizing individuals within a population (e.g. face recognition or recognizing individual whales by unique fluke patterns). Between these two extremes lie what are called entry- and subordinate-level categories. Entry-level categories include penguin, owl, *etc.*; people generally use these more specific labels instead of simply saying "bird" (the basic-level category). Subordinate-level categories are highly specific. Continuing with the example of birds, categorizing at the subordinate-level would require differentiating two quite similar species (such as the Red-breasted and White-breasted Nuthatches). Fine-grained recognition addresses this situation where categories are distinguised by very subtle differences.

Figure 1: **The Categorization Spectrum.** Object recognition ranges from differentiating basic-level categories (on the left) to recognizing individuals in a population (on the right). The domain of fine-grained recognition, which includes entry-level and subordinate-level categories, lies between these two extremes.

Fine-grained categorization (classification) is a fundamental part of visual recognition. We can each recall a time where we came upon something unusual or of interest. Perhaps we saw an unfamiliar insect on a shrub or a patch of brightly colored flowers. Perhaps our attention was drawn toward a bird as it sang from a nearby tree or fencepost. While people can easily identify the proper basic-level categories (an insect, some flowers, and a bird in the above example), the domain specific knowledge needed for classification at a finer granularity (*e.g.*ebony jewelwing, hydrangea, and house wren) is generally rare. Much of the fascinating information that would satisfy our natural curiosity unfortunately remains out of reach. For example, how uncommon is the dragonfly that I'm watching? Do others see it regularly or is this sighting highly unusual? What does this bird eat? Is that snake crossing the path poisonous? Finding the correct fine-grained category needed to answer these questions with a simple textual search (e.g. googling "yellow bird with blue wings") is unlikely to succeed. I believe firmly, however, that computational techniques for fine-grained recognition will provide the needed bridge, connecting the physical world around us to the knowledge that the information world makes available.

I have pioneered techniques addressing the core challenge of fine-grained categorization - learning detailed models from few examples. Two of my recent papers have tackled this challenge by focusing on

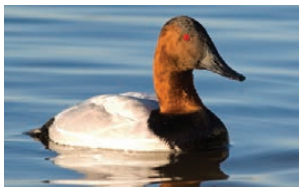
pose-normalization. Pose, or the relative configuration of an object's parts with respect to the camera, is a critical subproblem that must be solved to recognize objects. Much work has been done on pose estimation, both for rigid objects and for articulated objects such as humans. Being able to recognize an object in various poses is important but becomes critical in fine-grained domains as the visual characteristics which differentiate subordinate categories are often not only subtle but extremely localized. Consider, for example, the localized markings which differentiate each pair of species in Figure . The challenge is to design computational techniques which locate and measure these distinctive features despite image variations such as shadow, lighting, context and occlusion.



Redhead (Aythya americana)
© Mark Johnson

Piping Plover (Charadrius melodus)
© Ryan Schain

Downy Woodpecker (Picoides pubescens)
© jfuhr



Canvasback (Aythya valisineria)
© Mark Johnson

Snowy Plover (Charadrius nivosus)
© Dan Irizarry

Hairy Woodpecker (Picoides villosus)
© Jim McCree

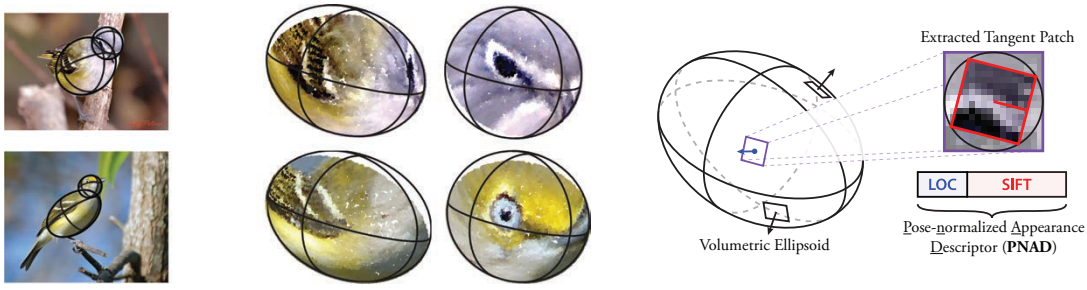
Figure 2: **Distinguishing characteristics are highly localized.** From left to right: eye/bill coloration and head/bill shape; bill/leg coloration and head/neck markings; bill length and black markings on outer tail feathers. The white patch on the Downy Woodpecker's back is not discriminative; it is simply not visible on the Hairy Woodpecker as it is obscured by its wings). *Note that these images were carefully selected, specifically to minimize pose difference.*

CONTRIBUTIONS

I worked on pose-normalization toward the end of my PhD, collaborating both with my advisor Larry Davis and with Trevor Darrell, whose group I had visited the preceding summer. Our efforts culminated in a paper [7] which was accepted at ICCV[1] 2011 as an oral (only 3.6% acceptance rate). Drawing inspiration from Biederman [1], we represented a bird using volumetric primitives, specifically modeling a bird's head and body with prolate ellipsoids. This volumetric representation was key to achieving pose-normalization. We leveraged a recently-developed framework called Poselets [2, 3] which utilizes an annotated training set to learn detection templates by clustering training examples with similar configuration. Each template is representative of a given partial pose or configuration. On a person, for example, it could capture a certain arm/torso or shoulder/head configuration. The power of the poselet approach is that when the detection templates are scanned across a test image looking for high-scoring regions, a high activation score provides evidence, not only for detecting the bird's presence (which poselets were designed for), but moreover, provides evidence about the location, orientation and scale (7 parameters) of the ellipsoidal part(s). As per-template predictions are combined, the parameterized ellipsoids provide a mapping from image pixels into a pose-normalized space

[1] International Conference on Computer Vision, the largest international computer vision conference. Occurs in alternating years with ECCV, the European Conference on Computer Vision.

(on the surface of the ellipsoids, see Figure). Extracting localized appearance features from within this pose-normalized space and training random forest classifiers on these features enables subordinate categorization of a query bird image independent of the bird's pose or the camera angle!



(a) Detection predicts ellipsoid parameters

(b) Map image pixels to pose-normalized space

(c) Extract pose-normalized features for classification

Figure 3: **Pose-normalization using Volumetric Primitives.** The poselet-based detection templates allow prediction of the volumetric part parameters, see (a). Once localized, the volumetric parts allow us to map image pixels into a pose-normalized space, see (b), from which we can extract pose-normalized appearance features for classification, see (c).

While the above method was (and still is) very promising, one drawback is that it requires the training corpus to be carefully annotated with 7 degree-of-freedom ellipsoids. This limitation, coupled with the inherent difficulty in detecting birds of varied pose and appearance, prompted us to investigate alternative representations in two-dimensions. We developed an approach called Pose-pooling Kernels which we published at CVPR[2] 2012 [12] (24.1% acceptance rate). Over the last decade, the prevailing paradigm for object classification has been the bag-of-words model [11] which densely samples image features, aggregating or pooling them into a global histogram according to a predetermined vector quantization of the feature space. Spatial pyramid matching (SPM) [10] sought to capture some coarse spatial distribution of the features by imposing a multiresolution grid on the image and concatenating the bag-of-words representation of each region into a high-dimensional feature vector. While SPM has been effective for basic-level categories, it suffers in fine-grained domains due to the highly localized nature of distinguishing characteristics exacerbated by pose variability. Where SPM defines pooling regions based on a uniform grid at predefined scales (typically the whole image or bounding box, then quadrants, sixteenths, etc.), the pose-pooling kernel (see Figure) defines semantic pooling regions, aggregating features, for example, from the head area, the body area and so forth. To identify these semantic regions, we again leveraged the poselet framework, using 2D keypoints instead of 3D ellipsoids, to represent pose. Based on the keypoints that are present in each individual poselet, the set of poselets is clustered into groups with similar semantic content, thus defining semantic pooling regions. In practice, there may be two or three head pooling regions, spanning the space of diverse poses (e.g. left-facing, frontal and right-facing). The pose-pooling kernel harnesses these semantic pooling regions to define an approximately pose-normalized distance metric between images, again facilitating fine-grained recognition.

My contributions toward defining and advancing the topic of fine-grained recognition placed me amongst the pioneers in this emerging area. In the fall of 2010, I was a co-organizer of the First Workshop on Fine-Grained Visual Categorization (FGVC), held in conjunction with CVPR in June 2011. This workshop was the first meeting to address the recognition of fine-grained categories and featured invited talks and panel discussions by researchers from both computer science (from academia and industry) and psychology. Due to the immense success of this first workshop (which drew more than 100 participants), we proposed and are organizing a second FGVC workshop to be held in June 2013.

[2]Computer Vision and Pattern Recognition, the largest domestic conference on computer vision.

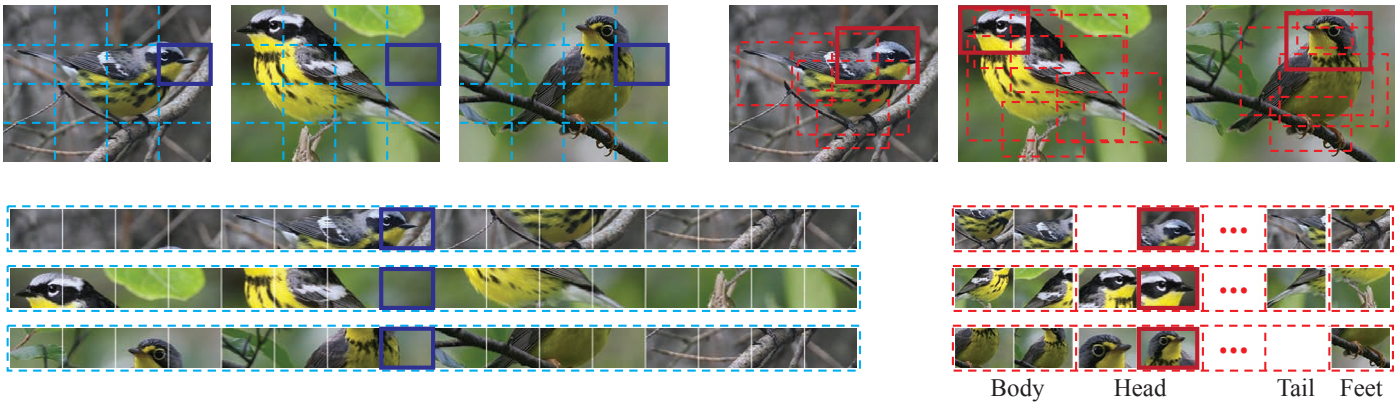


Figure 4: **Pose-pooling Kernel.** Spatial Pyramid Matching (SPM) [10] pools features by uniform spatial subdivision of the image (see the left half of the figure, and particularly note the differing feature vectors for the region framed in dark blue). The pose-pooling kernel [12] instead uses poselet-based detectors to define semantic pooling regions (see the rectangles on the right in dark red) which facilitate better comparison.

Datasets are key drivers of progress in this endeavor and I have also been leading on this front. As a part of the Visipedia project, the vision groups of Pietro Perona and Serge Belongie co-developed the Caltech-UCSD Birds (CUB-200) dataset comprising 200 bird categories, releasing it in 2010. This dataset has been adopted by many for the evaluation of fine-grained recognition approaches. When Pietro and Serge were presented with the opportunity to work with world-leading researchers at the Cornell Lab of Ornithology, they looked to me to assist in this collaboration. Being "bilingual" with expertise in both fine-grained vision and birds (the latter due to a decade of recreational birding), they felt that I was in a unique position to coordinate the collaboration with Cornell's experts.

After a year of planning efforts and several iterations of tool design, we are nearing completion of a new bird dataset covering all of the species commonly observed in North America. The compelling elements of the new dataset surpass the increased scale (700+ categories). Instead of the traditional dataset collection approach of text-based web queries for image harvesting and a crowdsourcing service such as Amazon's Mechanical Turk (MTurk) for labeling and annotation, we designed an approach which leverages both the expertise and passion of the target community of bird photographers, experts and enthusiasts. More than 60 individual photographers (many semi-pro) have given us permission to use their entire photo collections. We have also created a tool called ImageShare through which hundreds of individuals have collectively contributed many thousands of photos. The next innovation was the use of bird identification experts to verify the categorization of each image, offering a dramatic improvement in accuracy over the prior standard of Mechanical Turk workers. Lastly, through our partnership with the Cornell Lab, we have engaged the broader birding community as volunteers who are eager to help us collect bounding box, keypoint and other annotations. To illustrate the significance of tapping this motivated user base, a single post on the Lab's Facebook page produced in mere minutes a 100-fold increase in the number of people helping to provide bounding box annotations, leading to nearly 50,000 annotations over the next 24 hours. As our annotators are driven by passion instead of pennies, they produce extremely high-quality results at zero cost.

The data we are collecting has two purposes: first, it will yield a computer vision dataset that advances the field of fine-grained recognition and second, it will be used as a part of Cornell's NSF-sponsored Merlin project which aims to create an interactive bird identification tool. Moreover, Merlin's technology, trained on the data we are collecting, will likely be deployed as an official smartphone-based bird identification app by the Cornell Lab of Ornithology. The technology and models developed using the data we are collecting will hopefully find their way into the hands of tens or hundreds of thousands of people.

RESEARCH AGENDA

Looking ahead, I am very passionate about fine-grained recognition and look forward to making many additional contributions in the coming years. In addition to continuing work on pose-normalization, I wish to focus on other aspects including taxonomy and one-shot learning. Models trained on subtrees of a hierarchy may perform better than those trained on the full tree. Certain properties, such as shape and proportion, begin to be rather similar in some subtrees, another cue that has yet to be properly exploited. One-shot learning challenges traditional supervised machine learning where a multi-category training set will have many exemplars from each category. In one-shot learning (or similarly, few-shot learning), a broad training set is available for many categories, but we wish to add additional categories each having just a single exemplar (or a few exemplars) and yet be able to recognize those novel categories with high accuracy. This is much closer to how a human child learns to differentiate objects or animals. For example, if a child grows up with a dog at home and then meets the neighbor's cat, the child will quickly determine which features differentiate cats from dogs. The child does not need to see hundreds (to billions in recent Google experiments) of cats or cat photos to discriminate between dogs and cats.

I am also interested in partnering with other researchers to tackle some of the other fundamental vision problems which are present in fine-grained domains. Advancing the state-of-the-art in object detection will facilitate better fine-grained results. Additionally, the fundamental problem of image segmentation endeavors to determine which pixels belong to a given object and which are background or belong to other objects. Being able to focus on the portions of an image which contain the object of interest and neglect portions that do not should yield cleaner object models. These are just two of the tasks that the new dataset will address. I hope that researchers in many areas of vision will utilize the dataset for their work on related problems including classification, detection, segmentation, pose estimation, part localization, and one-shot learning. I look forward to partnering with some of these researchers to work on these other fundamental problems.

While I greatly enjoy working on vision research with birds as the domain of focus, there are many other domains where fine-grained techniques can be applied. Recent work has addressed the recognition of dog breeds, tree and leaf species, and forthcoming work addresses the prediction of vehicle make and model. I look forward to working on such domains and to thinking about how to lift or generalize fine-grained approaches designed for a specific domain such that they can apply to recognition of many or even all visual categories. In looking at alternate fields, I have come to feel that collaborating with domain experts is indispensable. I look forward to pursuing interdisciplinary collaborations with biologists and others as I have previously with the Cornell Lab of Ornithology and with evolutionary biologists at the University of Maryland [8].

Such collaborations help us as computational researchers to see real-world problems and search for solutions that can actually be applied. Another example of such a real-world problem was revealed to me recently by Dr. Peter Oboyski who manages the insect collection at the Essig Museum of Entomology. Dr. Oboyski shared with me that one of his colleagues is a moth expert and he maintains a moth trapping station in his yard during certain seasons. One evening, he noted an unfamiliar moth in his trap. As a moth expert, he rarely saw anything unfamiliar. It turns out that even a small population of this new moth's offspring would be heavily destructive to local crops. Accurate insect identification, both near farmland and at agricultural border crossings, has the potential for tremendous economic impact. This is just one of many compelling yet overlooked opportunities, chances where I will apply fine-grained vision techniques and hope to realize the great potential benefits to society.

Lastly, I maintain an interest in other fields of computer science and look forward to opportunities to partner with faculty members that work in these areas. Interestingly, my desire to do research was first kindled by an HCI project using Post-it notes on a digital smartboard surface for information architecture and layout (published at UIST 2001 [9]). I also have spent time working on sensor and camera networks, both using cameras to localize nodes in a sensor network [6] and estimating the topology and transition models within camera networks [5, 4]. I also have interest in applying my computer vision background into promising

fields such as big data and computational biology. The images and video that make up computer vision datasets are inherently "big data" with some datasets now topping a terabyte. The challenges inherent in distributed storage, efficient processing and analysis of such large-scale datasets are being addressed by vision researchers. In the field of computational biology, I feel that my expertise in localizing distinctive visual features has application to such as sequence alignment and analysis.

CONCLUSION

In conclusion, I have been a leader in the field of fine-grained visual categorization (FGVC), contributing to significant recent progress through my work on pose-normalization, the joint creation of a new fine-grained dataset made by engaging domain experts and enthusiasts, and the bringing together the fine-grained research community through the organizing of FGVC workshops. I will continue to be a leader in my field by pursuing the research agenda outlined above which includes future work on fine-grained recognition topics such as pose-normalization, leveraging taxonomy, image segmentation, and an increased focus on one-shot learning. I also am eager to forge collaborations both with faculty from other areas of computer science and interdisciplinary partnerships with researchers from other fields such as biology.

REFERENCES

[1] Irving Biederman. Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review*, 94(2):115–147, 1987. 2

[2] Lubomir Bourdev, Subhransu Maji, Thomas Brox, and Jitendra Malik. Detecting People Using Mutually Consistent Poselet Activations. In *ECCV*, 2010. 2

[3] Lubomir Bourdev and Jitendra Malik. Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations. In *ICCV*, 2009. 2

[4] Ryan Farrell and Larry S. Davis. Decentralized Discovery of Camera Network Topology. In *ICDSC*, 2008. 6

[5] Ryan Farrell, David Doermann, and Larry S. Davis. Learning Higher-order Transition Models in Medium-scale Camera Networks. In *OMNIVIS*, 2007. 6

[6] Ryan Farrell, Roberto Garcia, Dennis Lucarrelli, Andreas Terzis, and I-Jeng Wang. Localization in Multi-Modal Sensor Networks. In *ISSNIP*, 2007. 6

[7] Ryan Farrell, Om Oza, Ning Zhang, Vlad I. Morariu, Trevor Darrell, and Larry S. Davis. Birdlets: Subordinate Categorization using Volumetric Primitives and Pose-normalized Appearance. In *ICCV*, 2011. 2

[8] Aniruddha Kembhavi*, Ryan Farrell*, Yuancheng Luo, David Jacobs, Ramani Duraiswami, and Larry S. Davis. Tracking Down Under: Following the Satin Bowerbird. In *WACV*, 2008. 5

[9] Scott R. Klemmer, Mark W. Newman, Ryan Farrell, Mark Bilezikjian, and James A. Landay. The Designers' Outpost: a Tangible Interface for Collaborative Web Site Design. In *UIST*, 2001. 5

[10] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR*, 2006. 3, 4

[11] Josef Sivic and Andrew Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *ICCV*, 2003. 3

[12] Ning Zhang, Ryan Farrell, and Trevor Darrell. Pose Pooling Kernels for Sub-category Recognition. In *CVPR*, 2012. 3, 4