Birdsnap: Large-scale Fine-grained Visual Categorization of Birds

Thomas Berg[1], Jiongxin Liu[1], Seung Woo Lee[1], Michelle L. Alexander[1], David W. Jacobs[2], and Peter N. Belhumeur[1]

[1]Columbia University [2]University of Maryland

Abstract

We address the problem of large-scale fine-grained visual categorization, describing new methods we have used to produce an online field guide to 500 North American bird species. We focus on the challenges raised when such a system is asked to distinguish between highly similar species of birds. First, we introduce one-vs-most classifiers. *By eliminating highly similar species during training, these classifiers achieve more accurate and intuitive results than common one-vs-all classifiers. Second, we show how to estimate spatio-temporal class priors from observations that are sampled at irregular and biased locations. We show how these priors can be used to significantly improve performance. We then show state-of-the-art recognition performance on a new, large dataset that we make publicly available. These recognition methods are integrated into the online field guide, which is also publicly available.*

1. Introduction

Classification is one of the most fundamental problems of computer vision. It is generally assumed that objects are first detected at a basic level (*e.g.*, bird) and then further distinguished with finer granularity (*e.g.*, Tufted Titmouse). While most efforts have focused on basic level categorization, there has been exciting recent progress in fine-grained visual categorization (FGVC). Methods have been demonstrated in many domains, from shoes [5] to motorcycles [13], but biological categories–species and breeds–have been especially well-studied, with work tackling subcategory recognition of flowers [25], trees [14], dogs [1], butterflies [8], birds [4], and insects [17]. These biological domains, where taxonomy dictates a clear set of mutually exclusive subcategories, are wonderfully well-suited to the problem, and recognition systems in these domains are of practical use in ecology and agriculture [2, 17].

This work was supported by NSF awards 0968546 and 1116631, ONR award N00014-08-1-0638, and Gordon and Betty Moore Foundation grant 2987.

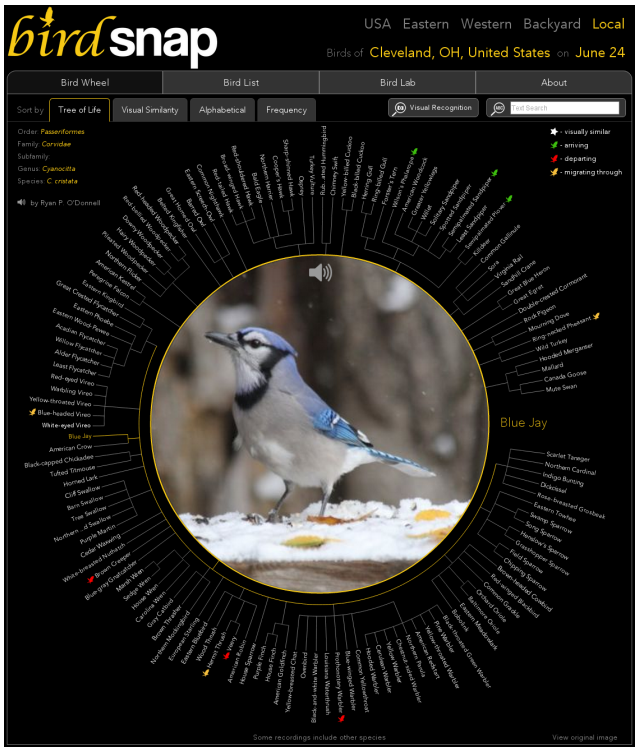

Figure 1. The Birdsnap web site, online at birdsnap.com.

Many of these applications require systems that scale to hundreds or even thousands of categories. A recent analysis [24] has shown that while state-of-the-art recognition methods perform well at basic-level recognition even on a 1000-category dataset such as that in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), these methods often confuse subcategories. This is intuitive; within the domain of a single basic-level category, visual similarity increases with the number of subcategories, often producing sets of subcategories that are nearly indistinguishable.

In this work, we approach the problem of large-scale fine-grained visual categorization by detailing methods needed to produce a digital field guide to 500 North American bird species. This online field guide, *Birdsnap*, avail-

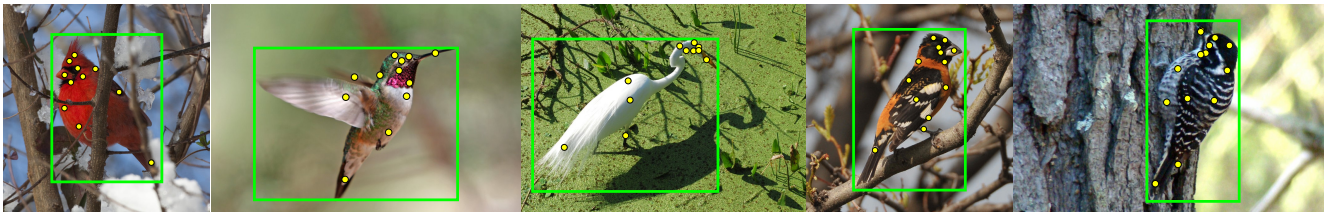

Figure 2. Sample images from the Birdsnap dataset, with bounding boxes and part annotations.

able at birdsnap.com, is a complete working system with a state-of-the-art visual recognition component that identifies birds in uploaded images. Figure 1 shows the home page. The 500 species (subcategories) have extensive visual overlap, with species within many genera, *e.g.*, terns (*Sterna*), scrub-jays (*Aphelocoma*), and some sparrows (*Melospiza*), exhibiting only slight visual differences. To address this, we introduce two ideas that mitigate complications arising from large numbers of highly similar subcategories.

The first we call "one-vs-most" classification, a replacement for one-vs-all classification, which is popular in fine-grained recognition (*e.g.*, [4, 21]). One-vs-all classifiers can have particular difficulty with highly similar classes, as each one-vs-all classifier finds samples very similar to the positive class in the negative training set. We show that reducing this difficulty in the training set leads to better results.

Our second method is based on the observation that modern cameras embed more than image data in the images they capture. In particular, many cameras sold in recent years are phones, and embed the time and location of capture in the image files they produce. Biological categories in particular often have a well-studied geographic distribution, and it is wasteful not to use this information. For migratory animals, the distribution depends on time as well as location, and we will show how the estimation and use of a spatio-temporal prior dramatically improves classification accuracy.

Finally, a key requirement of a field guide is to instruct the user on how to distinguish visually similar species. We present a fully automatic method for providing this instruction, with better results than our previous method [3].

Details of the methods used to produce the Birdsnap field guide are laid out in Sections 3-6, after a discussion of the most closely related work in Section 2. For completeness, we summarize the main contributions of this paper below:

1. We release and give a complete description of a working online field guide to 500 of the most common North American bird species.
2. We propose "one-vs-most" classification, a method for improving the accuracy of multiclass recognition when subsets of the classes are nearly indistinguishable.
3. We introduce a spatio-temporal prior on bird species. We show how to estimate this prior from an irregularly-sampled dataset of 75 million sightings records, and show that use of the prior provides significant improvement in classification accuracy.
4. We present state-of-the-art bird species recognition results, with higher accuracy on a more difficult dataset than previous work.
5. We release the Birdsnap dataset for fine-grained visual classification, with 49,829 images spanning 500 species of North American birds, labeled by species, with the locations of 17 body parts, and additional attribute labels such as male, female, immature, etc.
6. We present a method for automatically illustrating the differences between similar classes.

2. Related Work

Much recent work in fine-grained visual categorization has focused on species identification, with work on leaves [14, 25], flowers [19, 25], butterflies [8, 29], insects [17], cats and dogs (*e.g.*, [16, 21]), and birds (*e.g.*, [4, 6, 7, 8, 12, 30, 32, 33]). In most of this work, features are extracted from discriminative parts of the object, and used in a set of one-vs-all classifiers. Our *one-vs-most* classifiers use the POOF features introduced in [4] due to their excellent reported results in bird classification.

The large amount of recent work on fine-grained recognition of birds has been spurred by the availability of the excellent CUB-200 dataset [28]. Unfortunately CUB-200 includes species from many parts of the world but does not provide coverage of all or most species for any one part of the world. Our dataset covers all the commonly sighted birds of the United States, allowing us to produce a useful regional guide, and is over twice the size of CUB-200.

The first modern, illustrated field guide to birds was Peterson's *A Field Guide to the Birds* [22], published in 1934, with many successors. Online or mobile app guides include translations of paper guide books [18] and digital-only guides [20], but do not offer automatic recognition. Compared to existing digital guides that perform automatic recognition, Leafsnap [14] and the Visipedia [6] iPad app, our guide covers more species and requires less user effort. The generation of the "instructive," part of Birdsnap (not the automatic recognition component) is based on [3], with improvements described in Section 8.

3. The Birdsnap Dataset

Our dataset contains 49,829 images of 500 of the most common species of North American birds. There are between 69 and 100 images per species, with most species

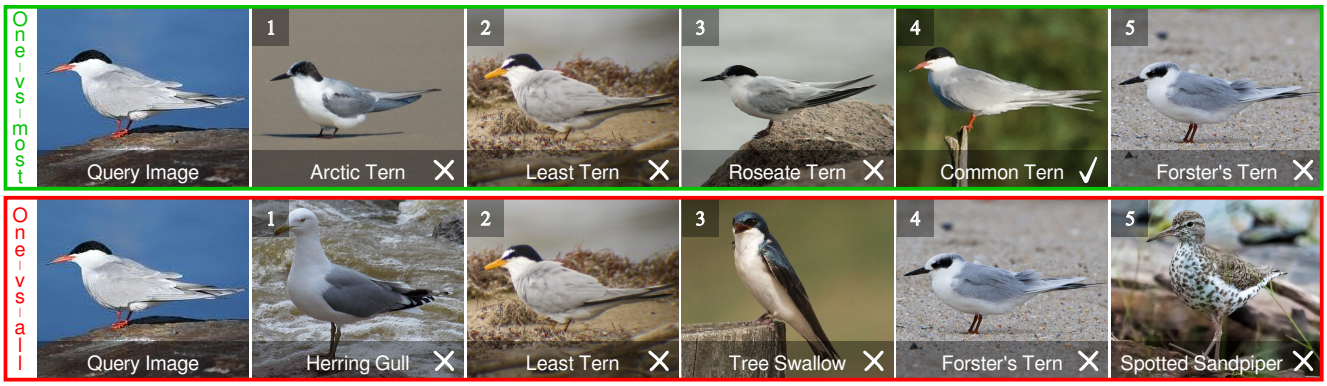

Figure 3. One-vs-most classifiers (top) improve both overall accuracy and the consistency and "reasonableness" of classification results. Here, they return the correct species at rank 4, with the top 5 results all terns (like the correct species). One-vs-all classifiers (bottom) omit the correct species from the top 5, and include a gull, a swallow, and a sandpiper. The supplementary material shows additional examples.

having 100. Each image is labeled with a bounding box and the location of 17 parts (see Figure 2). Some images are also labeled as male or female, immature or adult, and breeding or nonbreeding plumage.

The images were found by searching for each species' scientific name on Flickr. For species for which this did not yield enough images, we ran additional searches using the common names. The images were presented to labelers on Amazon Mechanical Turk, with illustrations of the species from a field guide, for confirmation of the species, and to flag images with no birds or multiple birds, or non-photographs. Labelers also marked the locations of the 17 parts. All labeling jobs were presented to multiple labelers, and images with inconsistent results were discarded.

Our dataset is similar in structure to CUB-200 [28], but has three important advantages. First, it contains two-and-a-half times the number of species and four times the number of images. Second, it covers all the most commonly sighted birds in one part of the world (the United States), which lets us build a tool that is useful in that region. Third, our dataset better reflects the appearance variation within many species. In particular, many bird species exhibit sexual dimorphism, with males and females having very different appearance. For example, in the red-winged blackbird, only the male has the distinctive red markings on the wing. CUB-200 contains only male red-winged blackbirds, while our dataset contains a mix of males and females.

4. One-vs-Most Classifiers

A fundamental problem in fine-grained visual categorization is how to handle subcategories that are nearly indistinguishable. In the bird world, an example of this problem is the terns, comprising ten species across six genera in our dataset, all of very similar appearance. If we train a discriminative one-vs-all classifier in the usual way for, say, the Common Tern, that classifier will be trained based on a positive set with images of just the common tern and a negative set that includes, in addition to non-terns, images of nine different species that look very much like the positive species. A classifier in this situation is very likely to latch on to accidental features that distinguish the Common Tern from other terns only in this particular training set and de-emphasize significant features that distinguish terns from non-terns.

To mitigate this issue, we omit from the negative training set all images of the k species most visually similar to the positive species (we use the similarity measure described in [3]). We call the resulting classifier a *one-vs-most* classifier. When the classifier omits similar terns from the negative training set, it is free to take advantage of features shared by terns (but different from other birds) as well as features that are unique to the common tern. Given a training set and a similarity measure, we choose the best value for k by evaluating performance on a held out set.

Note that one-vs-most classifiers can be implemented as a special case of cost-sensitive learning [9], by setting the cost of misclassification as the k most similar species to zero. However, while cost-sensitive learning usually sacrifices accuracy for lower cost, we will show in Section 6 that one-vs-most classifiers lead to both more reasonable (lower cost) errors and a reduction in overall error rate.

Birdsnap uses a set of one-vs-most SVMs based on POOFs, which are shown to be excellent features for bird species identification in [4]. Using one-vs-most classifiers brings a significant boost to accuracy. In addition, we find a qualitative benefit. Figure 3 shows the top 5 species returned for a query image of a Common Tern. The one-vs-all classifiers return two terns (very similar to the correct class), a gull (somewhat similar), and two "very wrong" species. The one-vs-most classifiers return 5 tern species, all very similar to (or equal to) the correct species. This pattern occurs for many queries; the one-vs-all classifiers, whether or not they find the correct species, often include species that are very different from the query image. Even when the rank-1 species is correct, this is a poor user experience. Results from the one-vs-most classifiers are more consistently

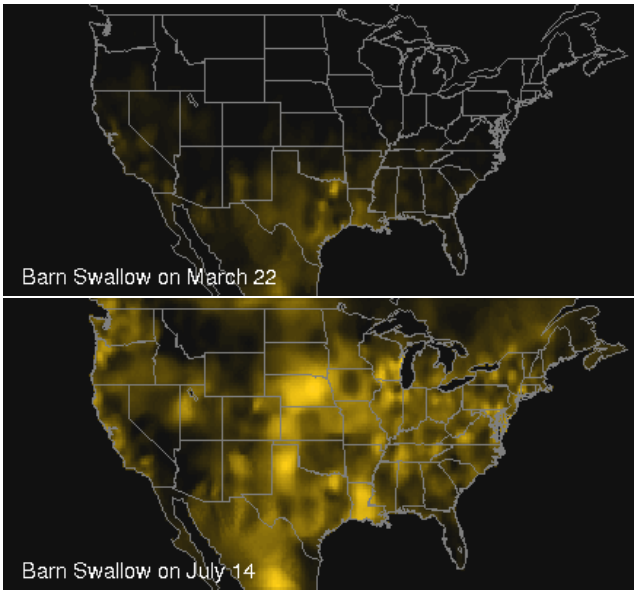

Figure 4. Fixed-time slices of our spatio-temporal prior show the Barn Swallow arriving from South America during its spring migration (above) and established in its summer grounds (below). Brighter regions indicate higher likelihood of a sighting.

similar to the query image. Experiments in Section 6 show the advantage of one-vs-most classifiers in both accuracy (Figures 5 and 7) and consistency (Figure 6).

5. A spatio-temporal prior for bird species

Prior knowledge can improve the performance of classification systems. A spatio-temporal prior is attractive for bird species identification, because the density of bird species varies considerably across the continent and throughout the year, due to migration. We see this in Figure 4, where slices of our spatio-temporal prior reveal the migration pattern of the Barn Swallow.

There is previous work using spatial priors to improve vision performance. For example, in pedestrian detection, knowledge of the ground plane and street layout can restrict a detector to regions of interest [10]. However, we are not aware of any work estimating spatio-temporal priors from large-scale observations to improve classification.

In order to combine a spatio-temporal prior with classifiers, we must convert the classifier output to a probability. As suggested by [31] we use the method of Platt [23] to produce probabilities from the output of the SVMs. This gives an estimate of $P(s|I)$ for each species s given image I, but these estimates may not be consistent with a single probability distribution. [31] note that simply normalizing the probabilities so that $\sum_s P(s|I) = 1$ works well in practice, and we follow this suggestion. To take advantage of the location x and date, t at which the photo was captured, we wish to find $P(s|I, x, t)$. Bayes' rule gives us

$$P(s|I, x, t) = P(I, x, t|s)P(s)/P(I, x, t). \quad (1)$$

We assume the image and the (location, date) pair are conditionally independent given the species, so this becomes

$$P(s|I, x, t) = P(I|s)P(x, t|s)P(s)/P(I, x, t). \quad (2)$$

Applying Bayes' rule to $P(I|s)$ and $P(x, t|s)$, we get

$$P(s|I, x, t) = \frac{P(s|I)P(I)}{P(s)} \frac{P(s|x, t)P(x, t)}{P(s)} P(s)/P(I, x, t)$$
$$\propto \frac{P(s|I)}{P(s)} P(s|x, t), \quad (3)$$

where we have dropped all factors that do not depend on s, as they will not affect the classification decision. $\frac{P(s|I)}{P(s)}$ is the calibrated classifier score ($P(s)$ appears in the denominator because in training the classifier we first equalize the number of images for each species). $P(s|x, t)$ is the spatio-temporal prior for the species.

5.1. Adaptive kernel density estimation of the spatio-temporal prior

In this section we construct an estimate for the prior probability that a bird observed at a given location and date belongs to a particular species. We use this prior to improve recognition performance of our classifiers (Section 5) and create visualizations that illustrate the varying distribution of a species throughout the year, or to provide a guide to the current species that one might observe at a particular place and time (Section 7).

Our prior is based on over 75 million records of North American bird sightings provided by eBird [26]. In addition, we make use of structural knowledge that some birds migrate annually, while others may remain year-round at a given location. We combine this information by first applying a variant of adaptive kernel density estimation to densely approximate the probability density of expected bird observations throughout the year in all parts of the US. We then post-process this density for each species to determine whether that species has been observed to migrate, and to determine the timing of migrations.

We wish to estimate the prior probability of a bird observation, $P(s|x, t)$, *i.e.* the probability that an observation made at time t and location x is of species s. As the density of a bird species displays much greater variation throughout the year than across different years [11], we let t denote a day and month, pooling data across years. Although we have a large volume of observational data available, direct estimation of the probability from this data is problematic, because of the uneven distribution of observations. Birding observations are concentrated near areas of high population density and/or at locations known to attract a wide variety of birds (for example, a high proportion of observations in New York City are reported from Central Park), and may occur disproportionately at certain times of year.

To deal with sparse data, we use adaptive kernel density estimation. First, we divide our problem into two parts. We

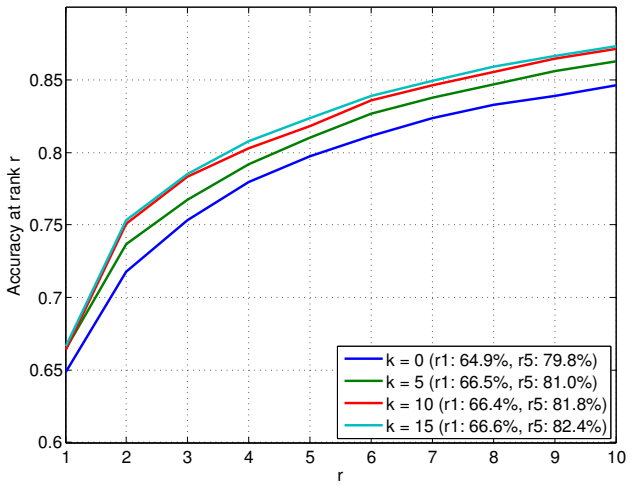

Figure 5. As we increase k, accuracy of the one-vs-most classifiers initially increases at all ranks. Results for additional values of k, shown in Table 1, are omitted for clarity.

estimate the density that any observation will occur at (x, t), and we also estimate the density of observations of species s at (x, t). $P(s|x, t)$ is then the ratio of these two densities.

We use a *balloon estimator* [27]:

$$\hat{f}(y) = \frac{1}{nh(y)^d} \sum_{i=1}^{n} K\left(\frac{y_i - y}{h(y)}\right). \quad (4)$$

Here, $\hat{f}(y)$ is the estimated density at $y = (x, t)$, n is the number of samples, d is the dimension of the space, $y_i = (x_i, t_i)$ is the ith sample, K is the kernel, in our case a Gaussian, and h is the bandwidth, which depends on the location and time, y, at which we are estimating the density. As noted by [27], the estimated density does not globally integrate to 1, but this is not a problem in our context, since we are taking the ratio of two estimates in which the same h is used for bandwidth. We set h, the standard deviation of the Gaussian, to half the distance to the 500th-nearest observation. We sum only over nearby observations, as distant observations contribute only small values to the sum. So we take

$$P(s|x, t) \approx \frac{\sum_{y_i \in N(y), s} K\left(\frac{y_i - y}{h_o(y)}\right)}{\sum_{y_i \in N(y)} K\left(\frac{y_i - y}{h_o(y)}\right)}. \quad (5)$$

The sum in the numerator is only over observations of species s. Note that h_o depends on all observations, not just those of species s. We take $N(y)$ to include all observations within a distance of $2h$ from y, guaranteeing that the estimate will be derived from a neighborhood containing at least 500 observations.

Even when we restrict sums to $N(y)$, this computation is potentially expensive. For this reason, we begin by discretizing all observations into spatio-temporal cubes with a spatial width of one-quarter degree of latitude/longitude and a temporal width of six days. This allows us to represent many observations with a single point, weighted by the number of observations. Distance calculations are done in units of these cubes, so a spatial distance between observations of a quarter degree is "equal" to a temporal distance of six days for purposes of kernel calculation.

k	rank 1	rank 3	rank 5	rank 10
0	0.649	0.753	0.798	0.846
1	0.658	0.755	0.799	0.851
3	0.660	0.762	0.807	0.863
5	0.665	0.768	0.810	0.863
7	0.666	0.779	0.816	0.869
10	0.664	0.783	0.819	0.872
15	0.666	0.785	0.824	0.873
20	0.661	0.786	0.823	0.877
30	0.657	0.792	0.836	0.879
40	0.659	0.790	0.830	0.885
50	0.648	0.787	0.830	0.882

Table 1. Accuracy of the one-vs-most classifiers increases at all ranks as k increases to 15. Beyond $k = 15$, high-rank accuracy continues to increase, but rank-1 accuracy decreases.

The problem of building spatio-temporal models of species distribution has been previously studied in the ecology literature. [11] contains a discussion of a number of prior methods, and proposes a new method in which spatially overlapping decision trees are combined to estimate the density of species observations. The input to the decision tree classifiers is a location and time, along with other meta-data about that location such as the elevation and type of land cover. Intuitively, one expects that this type of information can be useful, although [11] do not compare to a model that does not use this information. Unfortunately, while interesting, their system is rather complex, and they do not describe all parameters needed to replicate their results, nor do they make an implementation available for purposes of comparison.

6. Experiments on the Birdsnap Dataset

We hold out a test set of 2443 images–two to five per species–and train on the rest. Where images for a species include multiple images from a single Flickr account, we ensure those images are all in training or all in test, to avoid having test images of the same individual bird at the same time and place as any training image.

We learn 5000 random POOFs [4] from the training images using the labeled part locations, then extract the POOFs for one-vs-most training using detected part locations. We use the part detector of [15], which includes a random component, so we run it three times on each training image to augment the training set. This gives 250-285 training (image, parts) pairs per class, from which we use the 200 most accurate detections, reasoning that if the part detection fails badly, classification cannot succeed. Each one-vs-most classifier is a linear SVM trained on these 200 positive samples and 100 samples (randomly chosen from

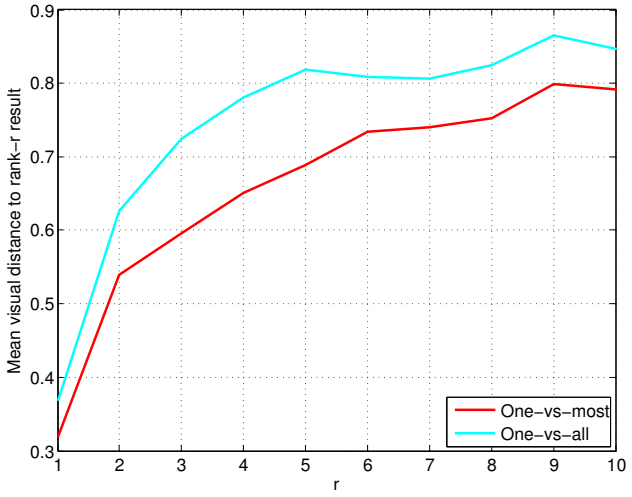

Figure 6. Mean visual distance between query species and returned species. One-vs-most classifiers return species that are more similar to the query species.

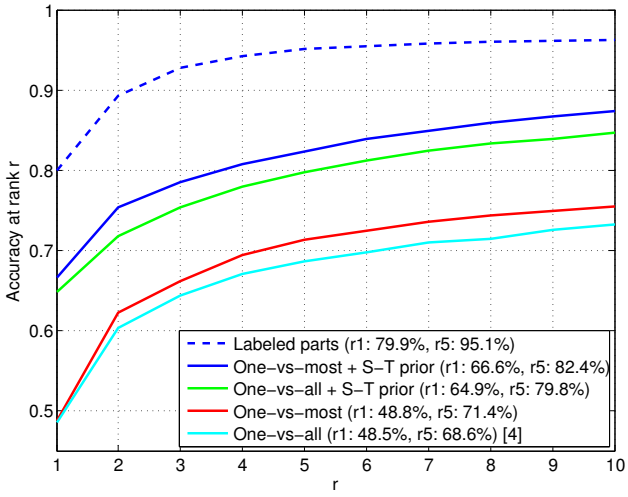

Figure 7. The one-vs-most classifiers and spatio-temporal prior each contributes significantly to overall performance. The dashed line, using labeled part locations, shows hypothetical performance with human-level part localization.

the 200) for each negative class. The extra positive samples improve the balance of the training set.

Many birds form flocks, and photographs often contain multiple birds–not always of the same species. To resolve this ambiguity and reduce response time, we ask users to click the rough location of the head and tail, giving us an approximate bounding box. This limits the search space considered by the part detector. In experiments, we generate these click locations by randomly perturbing the true location of the eye and tail in x and y by up to an eighth of the side length of the bounding box .

As with the images, we hold out a random subset of the bird sightings for testing. The North American portion of the eBird dataset includes 6,249,584 *checklists*–lists of the birds seen by an observer on a particular outing–with a total of 76,833,202 individual bird sightings. We hold out a randomly selected ten percent of the checklists for testing, and estimate the spatio-temporal prior from the remainder.

Each submission to the identification system consists of an (image, location, date) triple. We construct a test set by first choosing a random 10,000 sightings from the held-out eBird data, yielding a set of 10,000 (species, location, date) samples. For each sample, we randomly choose an image of that species from the held-out image set. This produces a test set of 10,000 (image, location, date) triples.

First, we seek the optimal value of k for the one-vs-most classifiers, *i.e*. how many species should be left out of the negative training sets. Figure 5 and Table 1 show accuracy within the top r guesses for several values of k. We see that while rank-1 accuracy peaks at $5 \leq k \leq 15$, rank-5 accuracy increases through $k = 30$, and rank-10 through at least $k = 40$. This is expected: at higher ranks, it is less useful to distinguish between highly similar species. For Birdsnap, we choose $k = 15$, which produces a nice boost at rank 5 without sacrificing accuracy at rank 1.

Figure 6 demonstrates the effect seen qualitively in Figure 3: that the top few species returned by the one-vs-most classifiers are more consistently similar to the query species than those returned by one-vs-all classifiers. We use the visual distance measure of [3], normalized so that the average distance between species is one, and find the mean over the test set of the distance from the species of the query image to the species returned at rank r. As suggested by Figure 3 and confirmed by Figure 6, the species returned by our one-vs-most classifiers are more visually similar to the query species than those returned by one-vs-all classifiers.

Figure 7 shows the contributions of the one-vs-most classifiers and the spatio-temporal prior over the standard one-vs-all classifiers (equivalent to one-vs-most with $k = 0$) without the prior. Note that this baseline–POOF-based one-vs-all classifiers–is the method of [4], which reports state of the art results on CUB-200. We see that at rank 5, the prior increases accuracy from 68.6% to 79.8%. This translates to a reduction in error rate of 35.6%, *i.e*. 35.6% of the errors of the baseline system are corrected by use of the spatio-temporal prior. Use of the one-vs-most classifiers brings rank-5 accuracy to 82.4%, an additional 12.9% reduction in error rate. Figure 7 also shows our system's accuracy if we use the manually labeled part location at training and test time. With manually labeled parts we achieve 79.9% accuracy at rank 1 and 95.1% at rank 5. The large boost from using manually labeled parts suggest there is still plenty of room for improvement in part detection.

7. Visualizing species frequency and migration

The density estimation method described in the previous section smooths our observation data and fills in the prior in

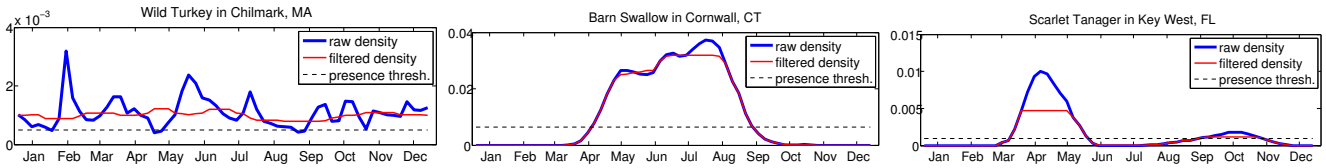

Figure 8. Species density over time in a fixed location. The "raw density" is the estimate from Section 5.1. Applying a median filter and adaptive threshold lets us recognize the Wild Turkey as present year round, despite the low frequency.

locations with few observations. Still, some noise remains. We can use structural knowledge of bird migrations to reduce this noise. For example, if we can determine that a bird has migrated away from a location in the winter, a few scattered observations can be treated as noise, and thresholded to zero. There is particular value in determining when a species is not present at a location, because we can use this knowledge to limit the species shown to a user browsing local birds. Also, we provide users with information about the timing of migration, which is of general interest.

Figure 8 shows the densities of three species. While most estimated densities are smooth over time, some rarely reported species, such as the Wild Turkey, have noisy densities. To smooth the noise without moving the edges, where the bird transitions between presence and absence, we apply a median filter. We then apply an adaptive threshold of 20% of the peak density to determine presence and absence.

At each location, a species can exhibit one of the following patterns of presence and absence:

1. in some locations, never present,
2. in some locations, present year-round, *e.g.*, the Wild Turkey in Chilmark, MA,
3. in the summer or winter grounds, present during one interval, *e.g.*, the Barn Swallow in Cornwall, CT, or
4. on the migration route, present during two intervals, *e.g.*, the Scarlet Tanager in Key West, FL.

(The examples are shown in Figure 8.) The 20% threshold is chosen empirically to make most species follow these patterns. To give users a sense of the bird activity around them, we give them the option of only showing birds that are currently in their area. Birds that follow the third pattern (indicated by two transition points during the year) and are close to transition are marked as "arriving" or "departing," while birds following the fourth pattern are marked as "migrating through."

8. Illustrating field marks

A traditional field guide is not a black box that identifies birds. Rather, through text and illustrations, it describes the distinguishing features, or *field marks*, of each species. This allows the user to justify the identification decision, and, once the field marks have been learned, to make future identifications without reference to the guide.

To achieve this in our online field guide, we illustrate, for any pair of similar species (s_i, s_j), features that effectively discriminate between them. To find such features, we consider a set of POOFs [4] as candidates. A POOF is a scalar-valued function trained to discriminate between two species based on features extracted from a particular region. We take the set of all POOFs trained on (s_i, s_j) and rank them by classification accuracy on a held-out set using a simple threshold classifier. Then we illustrate each of the top-ranked POOFs with a pair images, one of s_i and one of s_j, overlayed with ellipses that approximate the region used by the POOF, following the method of [3]. Each image pair illustrates a field mark.

The region used by each POOF is roughly set by the choice of two parts to an ellipse covering those two parts. Ellipses for different POOFs can have significant overlap, for example the POOF based on the beak and the crown often overlaps with that based on the beak and the forehead. To present a list of *distinct* field marks, we filter the ranked list of POOFs based on the Tanimoto similarity of the two ellipses, which is the ratio of the ellipses' intersection to their union. We define a *Tanimoto score* between two POOFs that discriminate between species s_i and s_j as the mean Tanimoto similarity between the ellipses drawn by the two POOFs, taken over the held-out images of s_i and s_j. We exclude any POOF whose Tanimoto score with a higher-ranked, non-excluded POOF is above a threshold. We find that a threshold of 0.05 gives a clear distinction between POOFs in the final list. Birdsnap displays the image pairs for the top three POOFs in the filtered list, with ellipses.

We previously [3] proposed a similar method for displaying differences between classes, but with a different ranking function and without filtering the ranked list of POOFs. The new ranking function, classification accuracy, is simpler and more intuitively related to our goal (to find POOFs that successfully discriminate between the classes). Figure 9 shows illustrated images for the top three field marks distinguishing the Great Egret and the Snowy Egret by both methods, and particularly shows the need for the filtering step. Additional examples are included in the supplementary material.

References

[1] A. Angelova and S. Zhu. Efficient object detection and segmentation for fine-grained recognition. 2013.

[2] T. Arbuckle, S. Schroder, V. Steinhage, and D. Wittmann. Biodiversity informatics in action: identification and monitoring of bee species using ABIS. In *Int. Symp. Informatics for Environmental Protection*, 2001.

Figure 9. Field marks differentiating the Great Egret and the Snowy Egret. By filtering based on Tanimoto similarity, our method ensures we find three *different* features: beak color, the extension of the mouth beneath the eye, and the long, slender neck. In contrast, the top three features found by our previous method [3] all appear to relate to beak color.

[3] T. Berg and P. N. Belhumeur. How do you tell a blackbird from a crow? In *ICCV*, 2013.

[4] T. Berg and P. N. Belhumeur. POOF: Part-based One-vs-One Features for fine-grained categorization, face verification, and attribute estimation. In *CVPR*, 2013.

[5] T. L. Berg, A. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010.

[6] S. Branson, G. V. Horn, C. Wah, P. Perona, and S. Belongie. The ignorant led by the blind: A hybrid humanmachine vision system for fine-grained categorization. *IJCV*, 2014.

[7] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *CVPR*, 2013.

[8] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *CVPR*, 2012.

[9] C. Elkan. The foundations of cost-sensitive learning. In *Int. Joint Conf. on Artificial Intelligence*, 2001.

[10] M. Enzweiler and D. Gavrila. Monocular Pedestrian Detection: Survey and Experiments. *PAMI*, 31(12), 2009.

[11] D. Fink, W. Hochachka, B. Zuckerberg, D. Winkler, B. Shaby, M. A. Munson, G. Hooker, M. Riedewald, D. Sheldon, and S. Kelling. Spatiotemporal Exploratory Models for Broad-scale Survey Data. *Ecological Applications*, 20(8), 2010.

[12] E. Gavves, B. Fernando, C. G. M. Snoek, A. W. M. Smeulders, and T. Tuytelaars. Fine-grained categorization by alignments. In *ICCV*, 2013.

[13] A. B. Hillel and D. Weinshall. Subordinate class recognition using relational object models. *NIPS*, 19:73, 2007.

[14] N. Kumar, P. N. Belhumeur, A. Biswas, D. W. Jacobs, W. J. Kress, I. Lopez, and J. V. B. Soares. Leafsnap: A computer vision system for automatic plant species identification. In *ECCV*, 2012.

[15] J. Liu and P. N. Belhumeur. Bird part localization using exemplar-based models with enforced pose and subcategory consistency. In *ICCV*, 2013.

[16] J. Liu, A. Kanazawa, D. Jacobs, and P. N. Belhumeur. Dog breed classification using part localization. In *ECCV*, 2012.

[17] G. Martinez-Munoz, N. Larios, E. Mortensen, W. Zhang, A. Yamamuro, R. Paasch, N. Payet, D. Lytle, L. Shapiro, S. Todorovic, A. Moldenke, and T. Dietterich. Dictionary-free categorization of very similar objects via stacked evidence trees. In *CVPR*, 2009.

[18] mydigitalearth.com. Sibley eGuide to birds (mobile app).

[19] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Indian Conf. Computer Vision Graphics and Image Processing*, 2008.

[20] C. L. of Ornithology. Merlin bird ID (mobile app).

[21] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012.

[22] R. T. Peterson. *A Field Guide to the Birds*. Houghton Mifflin Company, 1934.

[23] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3), 1999.

[24] O. Russakovsky, J. Deng, Z. Huang, A. C. Berg, and L. Fei-Fei. Detecting avocados to zucchinis: What have we done, and where are we going? In *ICCV*, 2013.

[25] A. R. Sfar, N. Boujemaa, and D. Geman. Vantage feature frames for fine-grained categorization. In *CVPR*, 2013.

[26] B. L. Sullivan, C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10), 2009.

[27] G. Terrell and D. Scott. Variable Kernel Density Estimation. *The Annals of Statistics*, 20(3), 1992.

[28] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Inst. Tech., 2011.

[29] J. Wang, K. Markert, and M. Everingham. Learning models for object recognition from natural language descriptions. In *BMVC*, 2009.

[30] B. Yao, G. Bradski, and L. Fei-Fei. A codebook-free and annotation-free approach for fine-grained image categorization. In *CVPR*, 2012.

[31] B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *SIGKDD*, 2002.

[32] N. Zhang, R. Farrell, and T. Darrell. Pose pooling kernels for sub-category recognition. In *CVPR*, 2012.

[33] N. Zhang, R. Farrell, F. Iandola, and T. Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *ICCV*, 2013.