Birdsnap: Large-Scale Fine-Grained Visual Categorization of Birds

Thomas Berg*, Jiongxin Liu*, Seung Woo Lee*, Michelle L. Alexander,* David W. Jacobs†, and Peter N. Belhumeur*

*Columbia University

†University of Maryland

Birdsnap

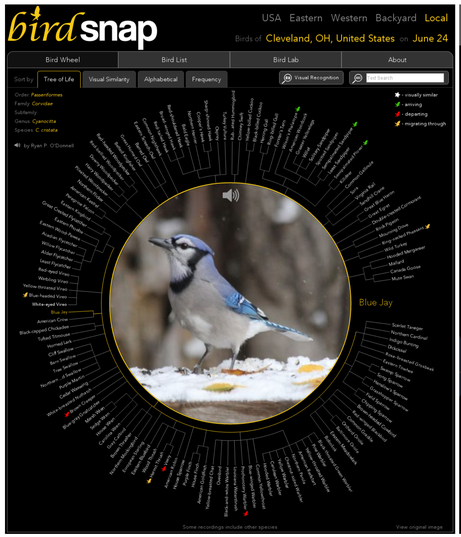
We have built a digital field guide to 500 common North American bird species. It identifies birds in photos. The recognition problem is difficult because:

There are a large number of classes, some of which are nearly indistinguishable.

We mitigate the problem with two techniques:

One-vs-Most Classifiers

A Spatio-Temporal Prior



http://birdsnap.com

In the Apple App Store

The Birdsnap Dataset

- 500 of the most common North American bird species, from Flickr
- 49,829 images
- Species labels confirmed on Mechanical Turk
- Bounding boxes and 17 part locations from Mechanical Turk
- Some images labeled with sex, age, and plumage
- Available at www.cs.columbia.edu/~tberg/

[1] T. Berg and P. N. Belhumeur, "Part-based One-vs-One Features for Fine-grained Categorization, Face Verification, and Attribute Estimation," CVPR 2013

[2] 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

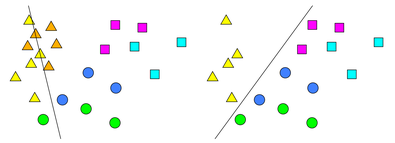
This work was supported by ONR award N00014-08-1-0638, Moore Foundation grant 2987, and NSF grants 0968546 and 1116631 .

One-vs-Most Classifiers

Some classes are nearly indistinguishable. Instead of *one-vs-all* classifiers, train *one-vs-most* classifiers, excluding the *k* classes most similar to the positive class from training.

Canada Goose

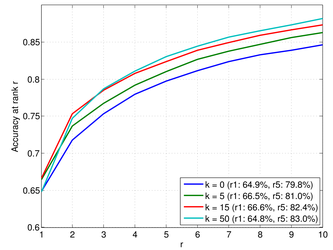
Cackling Goose

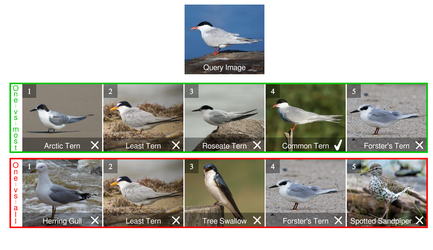


One-vs-all classifier

One-vs-most classifier

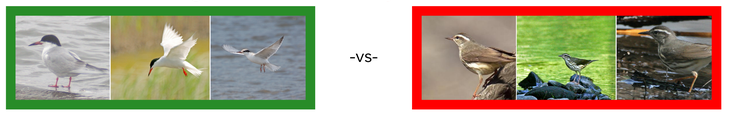
Results: Better accuracy and more reasonable mistakes.



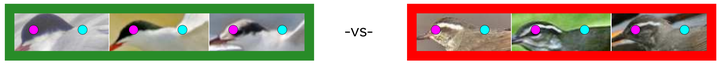


Background: Part-based One-vs-One Features [1]

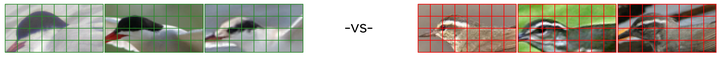
1. Choose two classes (e.g. *i* = common tern and *j* = Louisiana waterthrush)



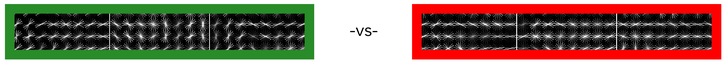
2. Choose a feature part and an alignment part (e.g. *f* = eye and *a* = back), align and crop



3. Divide cropped images into grids



4. Extract base features (e.g. *b* = gradient direction histograms (shown) or color histograms)



5. Train a linear SVM to separate the classes, then threshold the weights and retrain on just the discriminative region to get a POOF.

$$T^{i,j}_{f,a,b}$$

7. Repeat (e.g. with the Birdsnap dataset, can build millions of POOFs; we use 5000).

$$\binom{500}{2} \text{ class pairs} \cdot (13 \cdot 12) \text{ part pairs} \cdot 2 \text{ base features}$$
$$= 38,922,000 \text{ POOFs}$$

A Spatio-Temporal Prior

Given a bird image *I* captured at location *x* on date *t*, what is the probability the bird is of species *s*?

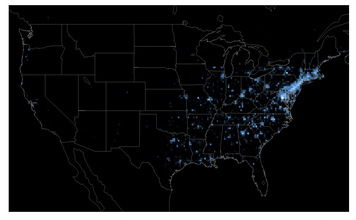
$$P(s|I,x,t) = \frac{P(I,x,t|s)P(s)}{P(I,x,t)}$$

Assuming conditional independence of *I* and (*x*, *t*) given *s*, we get

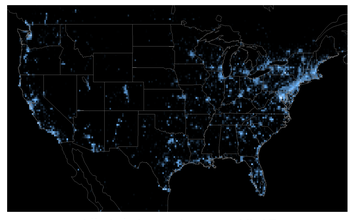
$$P(s|I,x,t) \propto \frac{P(s|I)}{P(s)} P(s|x,t)$$

From one-vs-most classifier — Spatio-temporal prior

Estimate spatio-temporal data from 75 million bird sighting records collected by eBird [2]. Problem: eBird data reflects bird distribution *and bird-watcher distribution*.



White-throated Sparrow sightings, January 12-17



All sightings, January 12-17

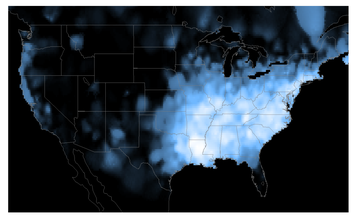
We use an adaptive kernel density estimate:

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

— Density estimate for species *s*

— Density estimate for all species

Where $y = (x,t)$, $N(y)$ is a neighborhood of y, and K is a Gaussian kernel of width h_o



$P(s|x,t)$ estimate

Recognition Performance

- Hold out 2443 images, 600k eBird checklists
- To generate a test sample:
  - Choose random *(s, x, t)* from held-out checklists
  - Choose random held-out image of *s*
- 10,000 test samples