The Caltech-UCSD Birds-200-2011 Dataset

Catherine Wah[1], Steve Branson[1], Peter Welinder[2], Pietro Perona[2], Serge Belongie[1]

[1] University of California, San Diego
La Jolla CA
{sbranson,cwah,sjb}@cs.ucsd.edu

[2] California Institute of Technology
Pasadena, CA
{welinder,perona}@caltech.edu

Abstract

CUB-200-2011 is an extended version of CUB-200 [7], a challenging dataset of 200 bird species. The extended version roughly doubles the number of images per category and adds new part localization annotations. All images are annotated with bounding boxes, part locations, and attribute labels. Images and annotations were filtered by multiple users of Mechanical Turk. We introduce benchmarks and baseline experiments for multi-class categorization and part localization.

1. Introduction

Bird species classification is a difficult problem that pushes the limits of the visual abilities for both humans and computers. Although different bird species share the same basic set of parts, different bird species can vary dramatically in shape and appearance (*e.g.*, consider pelicans vs. sparrows). At the same time, other pairs of bird species are nearly visually indistinguishable, even for expert bird watchers (*e.g.*, many sparrow species are visually similar). Intraclass variance is high due to variation in lighting and background and extreme variation in pose (*e.g.*, flying birds, swimming birds, and perched birds that are partially occluded by branches).

It is our hope that Birds-200 will facilitate research in subordinate categorization by providing a comprehensive set of benchmarks and annotation types for one particular domain (birds). We would like to cultivate a level of research depth that has thus far been reserved for a few select categories such as pedestrians and faces. Focusing on birds will help keep research more tractable from a logistical and computational perspective. At the same time, we believe that many of the lessons learned (in terms of annotation procedures, localization models, feature representations, and learning algorithms) will generalize to other domains such as different types of animals, plants, or objects.

2. Dataset Specification and Collection

Bird Species: The dataset contains 11,788 images of 200 bird species. Each species is associated with a Wikipedia article and organized by scientific classification (*order, family, genus, species*). The list of species names was obtained using an online field guide[1]. Images were harvested using Flickr image search and then filtered by showing each image to multiple users of Mechanical Turk [6]. Each image is annotated with bounding box, part location, and attribute labels. See Fig 1 for example images and Fig 6 for more detailed dataset statistics.

Bounding Boxes: Bounding boxes were obtained using the interface in Fig. 4.

Attributes: A vocabulary of 28 attribute groupings (see Fig 2(b)) and 312 binary attributes (*e.g.*, the attribute group *belly color* contains 15 different color choices) was selected based on an online tool for bird species identification[2]. All attributes are visual in nature, with most pertaining to a color, pattern, or shape of a particular part. Attribute annotations were obtained for each image using the interface in Fig. 5.

Part Locations: A total of 15 parts (see Fig 2(a)) were annotated by pixel location and visibility in each image using the GUI shown in Fig 3(a). The "ground truth" part locations were obtained as the median over locations for 5 different Mechanical Turk users per image.

3. Applications

Birds-200 has a number of unique properties that we believe are of interest to the research community:

Subordinate category recognition: Methods that are widely popular on datasets such as Caltech-101 [4] (*e.g.*, lossy representations based on histogramming and bag-of-words) are often less successful on subordinate categories, due to higher visual similarity of categories. Research in

[1] http://www.birdfieldguide.com
[2] http://www.whatbird.com

subordinate categorization may help encourage development of features or localization models that retain a greater level of discriminative power.

Multi-class object detection and part-based methods: Part-based methods have recently experienced renewed interest and success [3]. Unfortunately, availability of datasets with comprehensive part localization information is still fairly limited. Additionally, whereas datasets for image categorization often contain hundreds or thousands of categories [4, 1], popular datasets for object detection rarely contain more than 20 or so categories [2] (mostly due to computational challenges). Methods that employ shared part models offer great promise toward scaling object detection to a larger number of categories. Birds-200 contains a collection of 200 different bird species that are annotated using the same basic set of parts, thus making it uniquely suited toward research in shared part models.

Attribute-based methods: Attribute-based recognition is another form of model sharing that has recently become popular. Most existing datasets for attribute-based recognition (*e.g*. Animals With Attributes [5]) do not contain localization information. This is an obstacle to research in attributed-based recognition, because visual attributes are often naturally associated with a particular part or object (*e.g*. blue belly or cone-shaped beak).

Crowdsourcing and user studies: Annotations such as part locations and attributes open the door for new research opportunities, but are also subject to a larger degree of annotation error and user subjectivity as compared to object class labels. By releasing annotations from multiple MTurk users per training image, we hope to encourage research in crowdsourcing techniques for combining annotations from multiple users, and facilitate user studies evaluating the reliability and relative merit of different types of annotation.

4. Benchmarks and Baseline Experiments

We introduce a set of benchmarks and baseline experiments for studying bird species categorization, detection, and part localization:

1. **Localized Species Categorization:** *Given the ground truth part locations, assign each image to one of 200 bird classes*. This benchmark is intended to facilitate studies of different localization models (*e.g*., to what extent does localization information improve classification accuracy?), and also provide greater accessibility to existing categorization algorithms. Using RGB color histograms and histograms of vector-quantized SIFT descriptors with a linear SVM, we obtained a classification accuracy of 17.3% (see Fig 7(d)).

2. **Part Localization:** *Given the full, uncropped bird images, predict the location and visibility of each bird part*. We measured the distance between predicted part locations and ground truth, normalized on a per-part basis by the standard deviation over part click locations for multiple MTurk users. The maximum error per part was bounded at 5 standard deviations. This was also the error associated with misclassification of part visibility. Using HOG-based part-detectors and a mixture of tree-structured pictorial structures, we obtained an average error of 2.47 standard deviations (by contrast, an average MTurk user should be off by 1 standard deviation). See Fig 8 for example part localization results and their associated loss.

3. **Species Categorization/Detection:** *Using only the full, uncropped bird images, assign each image to one of 200 bird classes.* For this benchmark, one can use the method of his/her choice (*e.g*., image categorization, object detection, segmentation, or part-based detection techniques); however, since the images are uncropped, we anticipate that the problem cannot be solved with high accuracy without obtaining some degree of localization. Detecting the most likely part configuration using a universal bird detector (as for benchmark 2) and then applying a localized species classifier (as for benchmark 1), we obtained a classification accuracy of 10.3% (see Fig 7(b)).

References

[1] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. CVPR, 2009. 2

[2] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. IJCV, 2010. 2

[3] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. CVPR, 2008. 2

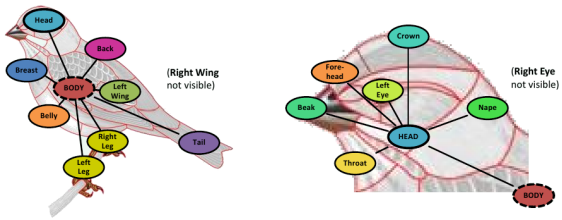
[4] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007. 1, 2

[5] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. CVPR, 2009. 2

[6] P. Welinder, S. Branson, S. Belongie, and P. Perona. The Multidimensional Wisdom of Crowds. NIPS, 2010. 1

[7] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 1

Figure 1. **CUB-200-2011 Example Images**



(a) Collected Parts

Part	Attributes	Part	Attributes	Part	Attributes
Beak	*HasBillShape, HasBillColor , HasBillLength*	**Back**	*HasBackColor, HasBackPattern*	**Breast**	*HasBreastPattern, HasBreastColor*
Belly	*HasBellyPattern, HasBellyColor*	**Fore-head**	*HasForehead Color*	**Bird** (all parts)	*HasSize, HasShape*
Throat	*HasThroatColor*	**Nape**	*HasNapeColor*	**Head**	*HasHeadPattern*
Crown	*HasCrownColor*	**Eye**	*HasEyeColor*	**Leg**	*HasLegColor*
Tail	*HasUpperTailColor, HasUnderTailColor, HasTailPattern, HasTailShape*	**Wing**	*HasWingPattern, Has WingColor, HasWingShape*	**Body**	*HasUnderpartsColor, HasUpperPartsColor, HasPrimaryColor*

(b) Attribute Part Associations

Figure 2. **Collected Parts and Attributes**. (a) The 15 part location labels collected for each image. (b) The 28 attribute-groupings that were collected for each image, and the associated part for localized attribute detectors.

(a) Part GUI

Figure 3. **MTurk GUI for collecting part location labels**, deployed on 11,788 images for 15 different parts and 5 workers per image.

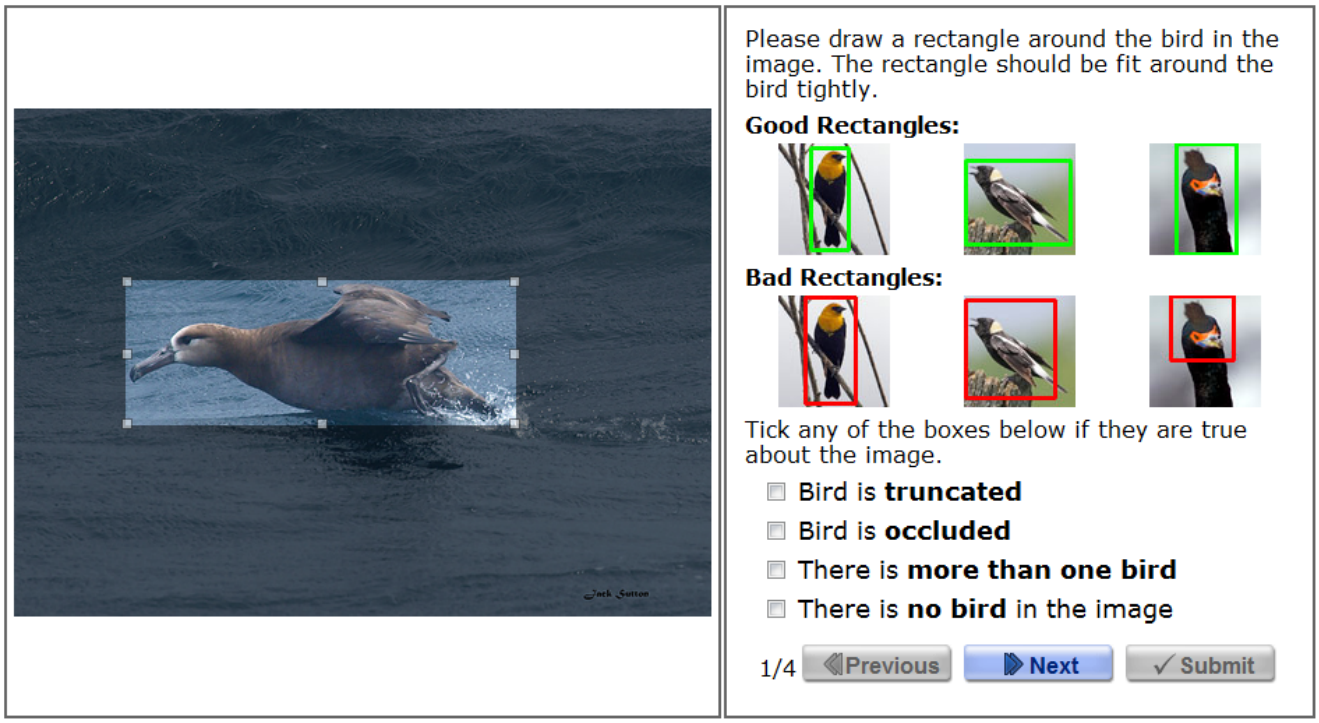


Figure 4. **MTurk GUI for collecting bounding box labels**, deployed on 11,788 images.

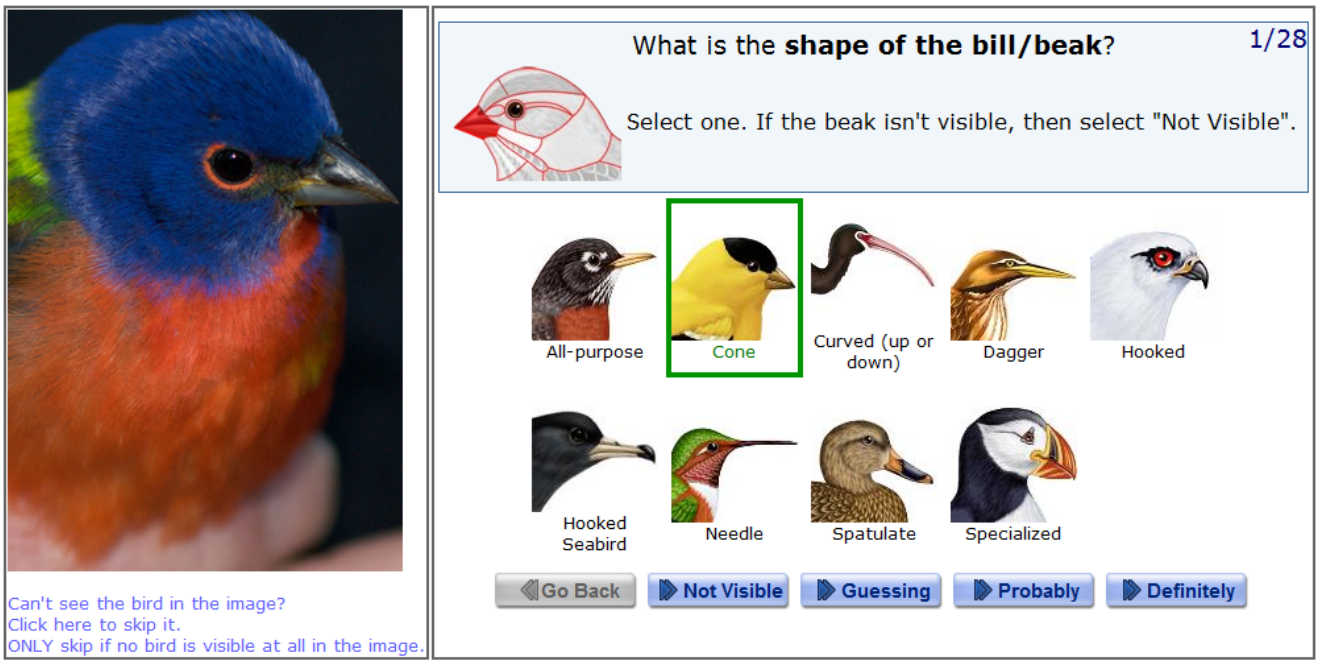
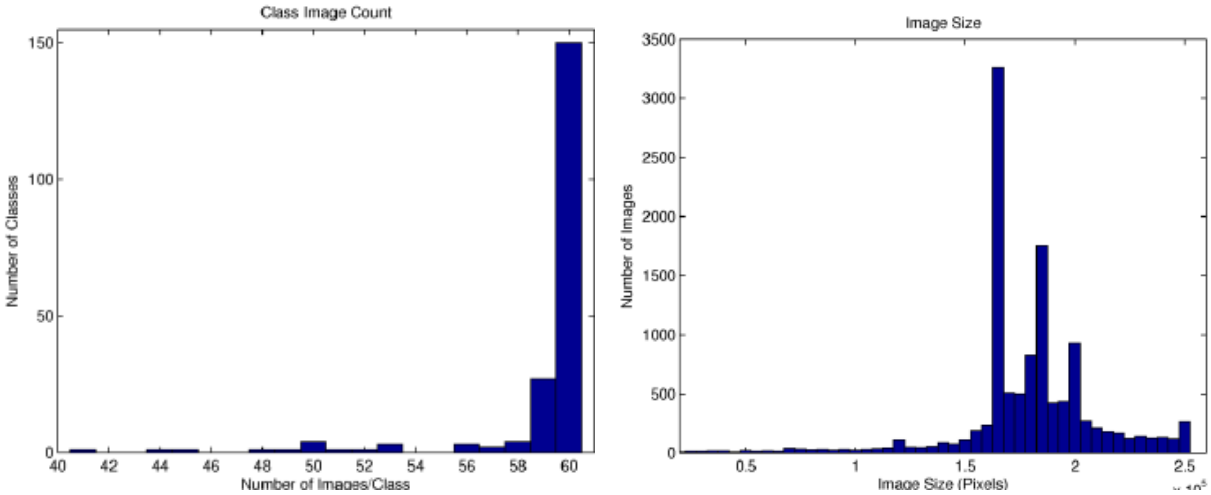


Figure 5. **MTurk GUI for collecting attribute labels**, deployed on 11,788 images for 28 different questions and 312 binary attributes.



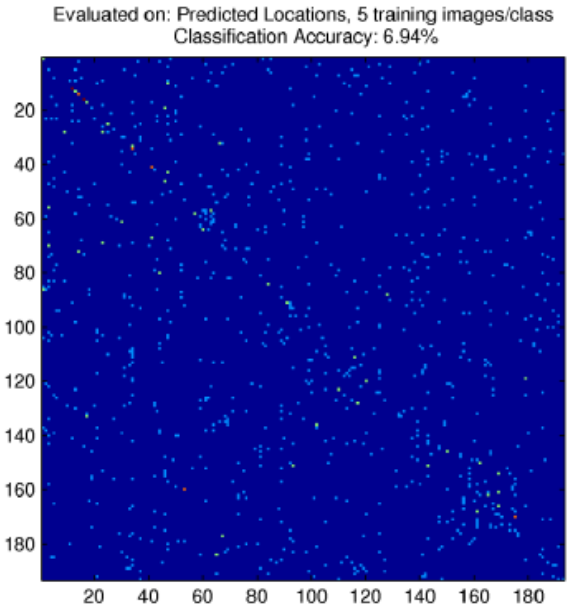
(a) Class Image Count

(b) Image Sizes

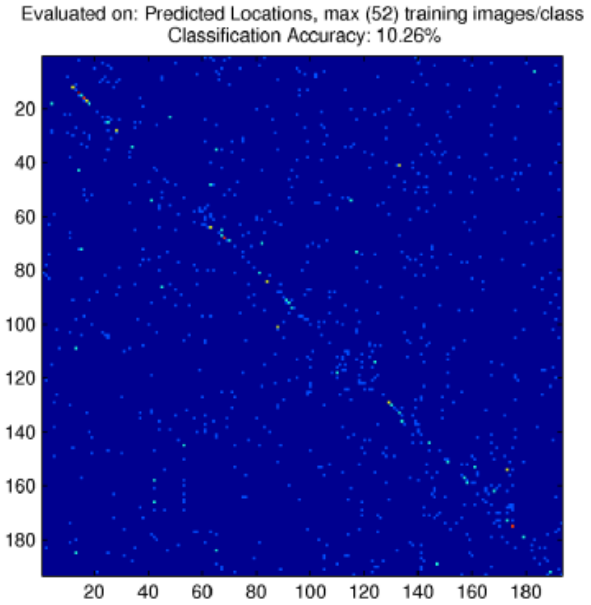
(c) Cropped/Uncropped Image Size Ratio

(d) Average Part Labeling Time

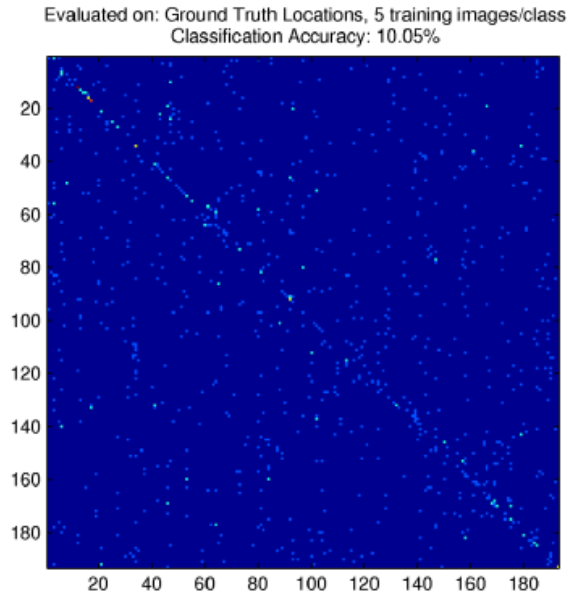
Figure 6. **Dataset Statistics**. (a) Distribution of the number of images per class (most classes have 60 images). (b) Distribution of the size of each image in pixels (most images are roughly 500X500). (c) Distribution of the ratio of the area of the bird's bounding box to the area of the entire image. (d) The average amount of time it took MTurkers to label each part.



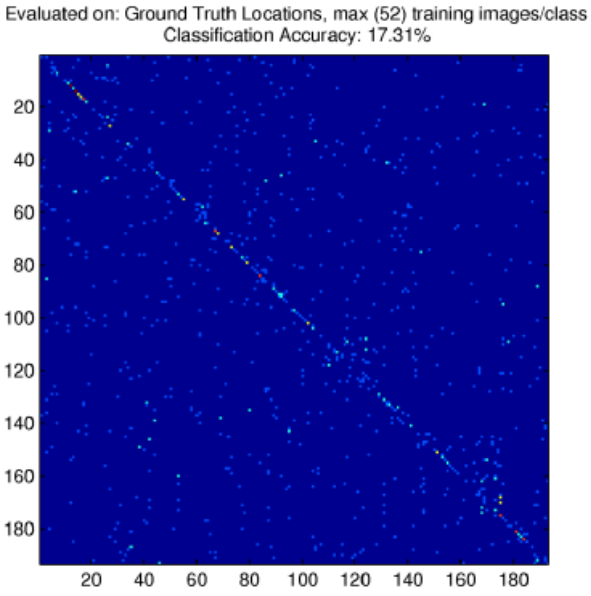
(a) Predicted Locations, 5 Images/Class



(b) Predicted Locations, 52 Images/Class



(c) Ground Truth Locations, 5 Images/Class



(d) Ground Truth Locations, 52 Images/Class

Figure 7. **Categorization Results** for 200-way bird species classification. The top 2 images show confusion matrices when using a universal bird detector to detect the most likely location of all parts and then evaluating a multiclass classifier. The bottom 2 images show confusion matrices when evaluating a multiclass classifier on the ground truth part locations. The 2 images on the left show results with 5 training images per class, and the images on the right show results with 52 training images per class.

Figure 8. **Example Part Detection Results**, with good detection results on the left and bad detection results on the right. A loss of 1.0 indicates that the predicted part locations are about as good as the average MTurk labeler.