Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection

Grant Van Horn[1], Steve Branson[1], Ryan Farrell[2], Scott Haber[3], Jessie Barry[3], Panos Ipeirotis[4], Pietro Perona[1], Serge Belongie[5]
[1]Caltech. [2]BYU. [3]Cornell Lab of Ornithology. [4]NYU. [5]Cornell Tech.

Computer vision systems – catalyzed by the availability of new larger scale datasets like ImageNet – have recently obtained remarkable performance on object recognition and detection. Computer vision has entered an era of big data, where the ability to collect larger datasets–larger in terms of the number of classes, the number of images per class, and the level of annotation per image–appears to be paramount for continuing to improve performance and expand the set of applications solvable by computer vision.

Unfortunately, expanding datasets in this fashion introduces new challenges beyond just increasing the amount of human labor required. As we increase the number of classes of interest, classes become more fine-grained and difficult to distinguish for the average person (and the average annotator), more ambiguous, and less likely to obey an assumption of mutual exclusion. The annotation process becomes more challenging for human annotators, requiring an increasing amount of skill and knowledge. Dataset *quality* appears to be at direct odds with dataset *size*.

In this paper, we introduce tools and methodologies for constructing large, high quality computer vision datasets, based on tapping into an alternate pool of crowd annotators–citizen scientists. Citizen scientists are non-professional scientists or enthusiasts in a particular domain such as birds, insects, plants, airplanes, shoes, or architecture. Citizen scientists contribute annotations with the understanding that their expertise and passion in a domain of interest can help build tools that will be of service to a community of peers. Unlike workers on Mechanical Turk, citizen scientists are unpaid. Despite this, they produce higher quality annotations due to their greater expertise and the absence of disinterested spammers. Additionally, citizen scientists can help define and organically grow the set of classes and its taxonomic structure to match the interests of real users in a domain of interest. Whereas datasets like ImageNetand CUB-200-2011have been valuable in fostering the development of computer vision algorithms, the particular set of categories chosen is somewhat arbitrary and of limited use to real applications. The drawback of using citizen scientists instead of Mechanical Turkers is that the throughput of collecting annotations maybe lower, and computer vision researchers must take the time to figure out how to partner with different communities for each domain.

We collected a large dataset of 48,562 images over 555 categories of birds with part annotations and bounding boxes for each image, using a combination of citizen scientists, experts, and Mechanical Turkers. We used this dataset to build a publicly available application for bird species classification[1]. In this paper, we provide details and analysis of our experiences with the hope that they will be useful and informative for other researchers in computer vision working on collecting larger fine-grained image datasets. We address questions like: What is the relative skill level of different types of annotators (MTurkers, citizen scientists, and experts) for different types of annotations (fine-grained categories and parts)? What are the resulting implications in terms of annotation quality, annotation cost, human annotator time, and the time it takes a requester to finish a dataset? Which types of annotations are suitable for different pools of annotators? What types of annotation GUIs are best for each respective pools of annotators? How important is annotation quality for the accuracy of learned computer vision algorithms? How significant are the quality issues in existing datasets like CUB-200-2011 and ImageNet, and what impact has that had on computer vision performance?

We summarize our contributions below:

1. Methodologies to collect high quality, fine-grained computer vision datasets using a new type of crowd annotators: citizen scientists.

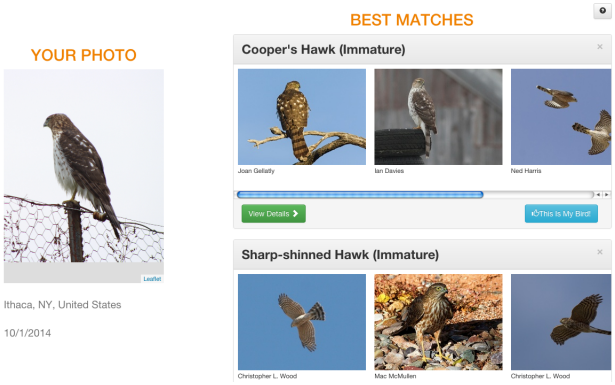

Figure 1: **Merlin Photo ID**: a publicly available tool for bird species classification that was built with the help of citizen scientists. The user uploaded a picture of a bird, and server-side computer vision algorithms identified it as an immature Cooper's hawk.

2. NABirds: a large, high quality dataset of 555 categories that was curated by experts.
3. Merlin Photo ID: a public tool for bird species classification.
4. Detailed analysis of annotation quality, time, cost, and throughput of MTurkers, citizen scientists, and experts for fine-grained category and part annotations.
5. Analysis of the annotation quality of the popular datasets CUB-200 and ImageNet.
6. Empirical analysis of the effect that annotation quality has when training computer vision algorithms for categorization.

A high-level summary of our findings is: 1) Citizen scientists have 2-4 times lower error rates than MTurkers at fine-grained bird annotation, while annotating images faster and at zero cost. Over 500 citizen scientists annotated images in our dataset–if we can expand beyond the domain of birds, the pool of possible citzen scientist annotators is massive. 2) A curation-based interface for visualizing and manipulating the full dataset can further improve the speed and accuracy of citizen scientists and experts. 3) Even when averaging answers from 10 MTurkers together, MTurkers have a more than 30% error-rate at 37-way bird classification. 4) The general high quality of Flickr search results (84% accurate when searching for a particular species) greatly mitigates the errors of MTurkers when collecting fine-grained datasets. 5) MTurkers are as accurate and fast as citizen scientists at collecting part location annotations. 6) MTurkers have faster throughput in collecting annotations than citizen scientists; however, using citizen scientists it is still realistic to collect a dataset of $\approx 100,000$ images in a domain like birds in around 1 week. 7) At least 4% of images in CUB-200-2011 and ImageNet have incorrect class labels, and numerous other issues including inconsistencies in the taxonomic structure, biases in terms of which images were selected, and the presence of duplicate images. 8) Despite these problems, these datasets are still effective for computer vision research; when training CNN-based computer vision algorithms with corrupted labels, the resulting increase in test error is surprisingly low and significantly less than the level of corruption. 9) A consequence of findings 3, 4, and 8 is that training computer vision algorithms on unfiltered Flickr search results (with no annotation) can often outperform algorithms trained when filtering by MTurker majority vote.

[1]merlin.allaboutbirds.org