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

Abstract

We introduce tools and methodologies to collect high quality, large scale fine-grained computer vision datasets using citizen scientists – crowd annotators who are passionate and knowledgeable about specific domains such as birds or airplanes. We worked with citizen scientists and domain experts to collect NABirds, a new high quality dataset containing 48,562 images of North American birds with 555 categories, part annotations and bounding boxes. We find that citizen scientists are significantly more accurate than Mechanical Turkers at zero cost. We worked with bird experts to measure the quality of popular datasets like CUB-200-2011 and ImageNet and found class label error rates of at least 4%. Nevertheless, we found that learning algorithms are surprisingly robust to annotation errors and this level of training data corruption can lead to an acceptably small increase in test error if the training set has sufficient size. At the same time, we found that an expert-curated high quality test set like NABirds is necessary to accurately measure the performance of fine-grained computer vision systems. We used NABirds to train a publicly available bird recognition service deployed on the web site of the Cornell Lab of Ornithology.[1]

1. Introduction

Computer vision systems – catalyzed by the availability of new, larger scale datasets like ImageNet [6] – have recently obtained remarkable performance on object recognition [17, 32] and detection [10]. 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 performance improvement and expanding the set of solvable applications.

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, 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 nonprofessional 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 spammers. Additionally, citizen scientists can help define and organically grow the set of classes and its taxonomic

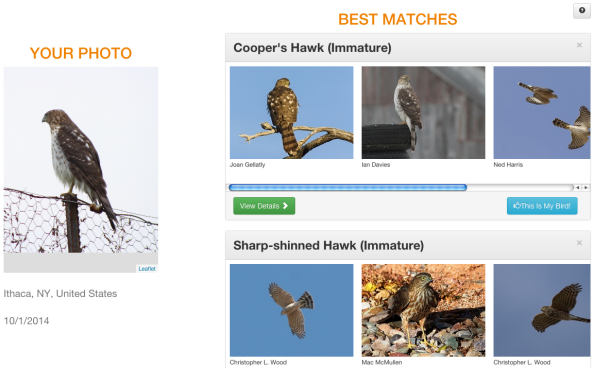


Figure 1: **Merlin Photo ID**: a publicly available tool for bird species classification built with the help of citizen scientists. The user uploaded a picture of a bird, and server-side computer vision algorithms identified it as an immature Cooper's Hawk.

[1]merlin.allaboutbirds.org

structure to match the interests of real users in a domain of interest. Whereas datasets like ImageNet [6] and CUB-200-2011 [35] have 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. 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.
2. NABirds: a large, high quality dataset of 555 categories curated by experts.
3. Merlin Photo ID: a publicly available 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 state-of-the-art computer vision algorithms for categorization.

A high-level summary of our findings is: a) 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. b) A curation-based interface for visualizing and manipulating the full dataset can further improve the speed and accuracy of citizen scientists and experts. c) Even when averaging answers from 10 MTurkers together, MTurkers have a more than 30% error-rate at 37-way bird classification. d) 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. e) MTurkers are as accurate and fast as citizen scientists at collecting part location annotations. f) MTurkers have faster throughput in collecting annotations than citizen scientists; however, using citizen scientists it is still realistic to annotate a dataset of around 100k images in a domain like birds in around 1 week. g) 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. h) 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. i) A consequence of findings (a), (d), and (h) 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.

2. Related Work

Crowdsourcing with Mechanical Turk: Amazon's Mechanical Turk (AMT) has been an invaluable tool that has allowed researchers to collect datasets of significantly larger size and scope than previously possible [31, 6, 20]. AMT makes it easy to outsource simple annotation tasks to a large pool of workers. Although these workers will usually be non-experts, for many tasks it has been shown that repeated labeling of examples by multiple non-expert workers can produce high quality labels [30, 37, 14]. Annotation of fine-grained categories is a possible counter-example, where the average annotator may have little to no prior knowledge to make a reasonable guess at fine-grained labels. For example, the average worker has little to no prior knowledge as to what type of bird a "Semipalmated Plover" is, and her ability to provide a useful guess is largely dependent on the efforts of the dataset collector to provide useful instructions or illustrative examples. Since our objective is to collect datasets of thousands of classes, generating high quality instructions for each category is difficult or infeasible.

Crowdsourcing with expertise estimation: A possible solution is to try to automatically identify the subset of work-

ers who have adequate expertise for fine-grained classification [36, 38, 28, 22]. Although such models are promising, it seems likely that the subset of Mechanical Turkers with expertise in a particular fine-grained domain is small enough to make such methods impractical or challenging.

Games with a purpose: Games with a purpose target alternate crowds of workers that are incentivized by construction of annotation tasks that also provide some entertainment value. Notable examples include the ESP Game [33], reCAPTCHA [34], and BubbleBank [7]. A partial inspiration to our work was Quizz [13], a system to tap into new, larger pools of unpaid annotators using Google AdWords to help find and recruit workers with the applicable expertise.[2] A limitation of games with a purpose is that they require some artistry to design tools that can engage users.

Citizen science: The success of Wikipedia is another major inspiration to our work, where citizen scientists have collaborated to generate a large, high quality web-based encyclopedia. Studies have shown that citizen scientists are incentivized by altruism, sense of community, and reciprocity [18, 26, 39], and such incentives can lead to higher quality work than monetary incentives [11].

Datasets: Progress in object recognition has been accelerated by dataset construction. These advances are fueled both by the release and availability of each dataset but also by subsequent competitions on them. Key datasets/competitions in object recognition include Caltech-101 [9], Caltech-256 [12], Pascal VOC [8] and ImageNet/ILSVRC [6, 29].

Fine-grained object recognition is no exception to this trend. Various domains have already had datasets introduced including Birds (the CUB-200 [35] and recently announced Birdsnap [2] datasets), Flowers [25, 1], Dogs and Cats [15, 27, 21], Stoneflies [24], Butterflies [19] and Fish [4] along with man-made domains such as Airplanes [23], Cars [16], and Shoes[3].

3. Crowdsourcing with Citizen Scientists

The communities of enthusiasts for a taxon are an untapped work force and partner for vision researchers. The individuals comprising these communities tend to be very knowledgeable about the taxon. Even those that are novices make up for their lack of knowledge with passion and dedication. These characteristics make these communities a fundamentally different work force than the typical paid crowd workers. When building a large, fine-grained dataset they can be utilized to curate images with a level of accuracy that would be extremely costly with paid crowd workers, see Section 5. There is a mutual benefit as the taxon communities gain from having a direct influence on the construction of the dataset. They know their taxon, and their community, better than vision researchers, and so they can ensure that the resulting datasets are directed towards solving real world problems.

A connection must be established with these communities before they can be utilized. We worked with ornithologists at the Cornell Lab of Ornithology to build NABirds. The Lab of Ornithology provided a perfect conduit to tap into the large citizen scientist community surrounding birds. Our partners at the Lab of Ornithology described that the birding community, and perhaps many other taxon communities, can be segmented into several different groups, each with their own particular benefits. We built custom tools to take advantage of each of the segments.

3.1. Experts

Experts are the professionals of the community, and our partners at the Lab of Ornithology served this role. Figure 4 is an example of an expert management tool (Vibe[3]) and was designed to let expert users quickly and efficiently curate images and manipulate the taxonomy of a large dataset. Beyond simple image storage, tagging, and sharing, the benefit of this tool is that it lets the experts define the dataset taxonomy as they see fit, and allows for the dynamic changing of the taxonomy as the need arises. For NABirds, an interesting result of this flexibility is that bird species were further subdivided into "visual categories." A "visual category" marks a sex or age or plumage attribute of the species that results in a visually distinctive difference from other members within the same species, see Figure 2. This type of knowledge of visual variances at the species level would have been difficult to capture without the help of someone knowledgeable about the domain.

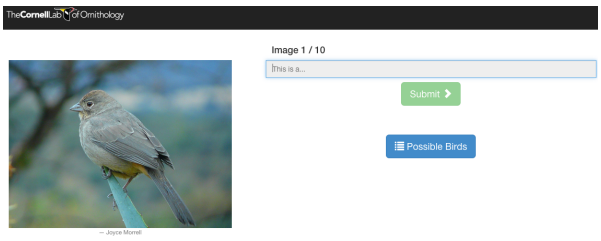
3.2. Citizen Scientist Experts

After the experts, these individuals of the community are the top tier, most skilled members. They have the confidence and experience to identify easily confused classes of the taxonomy. For the birding community these individuals

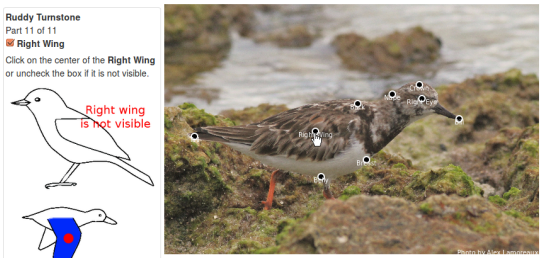
[2]The viability of this approach remains to be seen as our attempt to test it was foiled by a misunderstanding with the AdWords team.

[3]vibe.visipedia.org

Figure 2: Two species of hawks from the NABirds dataset are separated into 6 categories based on their visual attributes.



(a) Quiz Annotation GUI



(b) Part Annotation GUI

Figure 3: **(a)** This interface was used to collect category labels on images. Users could either use the autocomplete box or scroll through a gallery of possible birds. **(b)** This interface was used to collect part annotations on the images. Users were asked to mark the visibility and location of 11 parts. See Section 3.2 and 3.3

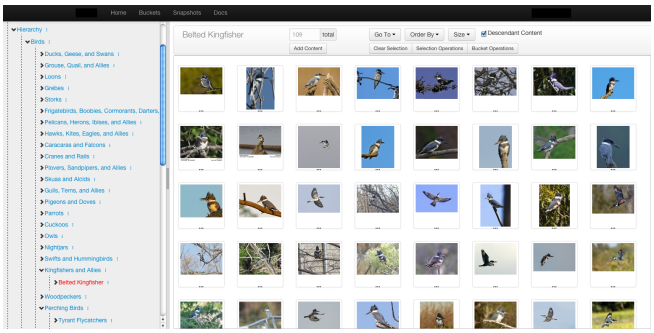


Figure 4: Expert interface for rapid and efficient curation of images, and easy modification of the taxonomy. The taxonomy is displayed on the left and is similar to a file system structure. See Section 3.1

were identified by their participation in eBird, a resource that allows birders to record and analyze their bird sightings.[4] Figure 3a shows a tool that allows these members to take bird quizzes. The tool presents the user with a series of images and requests the species labels. The user can supply the label using the autocomplete box, or, if they are not sure, they can browse through a gallery of possible answers. At the end of the quiz, their answers can be compared with other expert answers.

3.3. Citizen Scientist Turkers

This is a large, passionate segment of the community motivated to help their cause. This segment is not necessarily as skilled in difficult identification tasks, but they are capable of assisting in other ways. Figure 3b shows a part annotation task that we deployed to this segment. The task was to simply click on all parts of the bird. The size of this population should not be underestimated. Depending on how these communities are reached, this population could be larger than the audience reached in typical crowdsourcing platforms.

[4] ebird.org

4. NABirds

We used a combination of experts, citizen scientists, and MTurkers to build NABirds, a new bird dataset of 555 categories with a total of 48,562 images. Members from the birding community provided the images, the experts of the community curated the images, and a combination of CTurkers and MTurkers annotated 11 bird parts on every image along with bounding boxes. This dataset is free to use for the research community.

The taxonomy for this dataset contains 1011 nodes, and the categories cover the most common North American birds. These leaf categories were specifically chosen to allow for the creation of bird identification tools to help novice birders. Improvements on classification or detection accuracy by vision researchers will have a straightforward and meaningful impact on the birding community and their identification tools.

We used techniques from [5] to baseline performance on this dataset. Using Caffe and the fc6 layer features extracted from the entire image, we achieved an accuracy of 35.7%. Using the best performing technique from [5] with ground truth part locations, we achieved an accuracy of 75%.

5. Annotator Comparison

In this section we compare annotations performed by Amazon Mechanical Turk workers (MTurkers) with citizen scientists reached through the Lab of Ornithology's Facebook page. The goal of these experiments was to quantify the followings aspects of annotation tasks. 1) **Annotation Error**: The fraction of incorrect annotations., 2) **Annotation Time**: The average amount of human time required per annotation. 3) **Annotation Cost**: The average cost in dollars required per annotation. 4) **Annotation Throughput**: The average number of annotations obtainable per second, this scales with the total size of the pool of annotators.

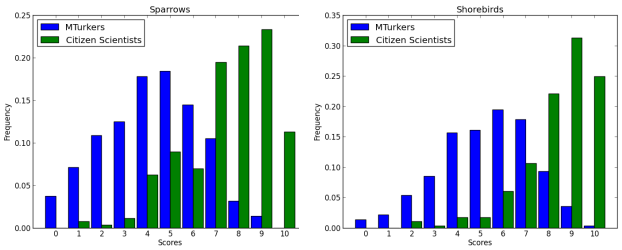
In order to compare the skill levels of different annotator groups directly, we chose a common user interface that we

considered to be appropriate for both citizen scientists and MTurkers. For category labeling tasks, we used the quiz tool that was discussed in section 3.2. Each question presented the user with an image of a bird and requested the species label. To make the task feasible for MTurkers, we allowed users to browse through galleries of each possible species and limited the space of possible answers to < 40 categories. Each quiz was focused on a particular group of birds, either sparrows or shorebirds. Random chance was 1 / 37 for the sparrows and 1 / 32 for the shorebirds. At the end of the quiz, users were given a score (the number of correct answers) and could view their results. Figure 3a shows our interface. We targeted the citizen scientist experts by posting the quizzes on the the eBird Facebook page.

Figure 5 shows the distribution of scores achieved by the two different worker groups on the two different bird groups. Not surprisingly, citizen scientists had better performance on the classification task than MTurkers; however we were uncertain as to whether or not averaging a large number of MTukers could yield comparable performance. Figure 6a plots the time taken to achieve a certain error rate by combining multiple annotators for the same image using majority voting. From this figure we can see that citizen scientists not only have a lower median time per image (about 8 seconds vs 19 seconds), but that one citizen scientist expert label is more accurate than the average of 10 MTurker labels. We note that we are using a simple-as-possible (but commonly used) crowdsourcing method, and the performance of MTurkers could likely be improved by more sophisticated techniques such as CUBAM [36]. However, the magnitude of difference in the two groups and overall large error rate of MTurkers led us to believe that the problem could not be solved simply using better crowdsourcing models.

Figure 6c measures the raw throughput of the workers, highlighting the size of the MTurk worker pool. With citizen scientists, we noticed a spike of participation when the annotation task was first posted on Facebook, and then a quick tapering off of participation. Finally, Figure 6b measures the cost associated with the different levels of error–citizen scientists were unpaid.

We performed a similar analysis with part annotations. For this task we used the tool shown in Figure 3b. Workers from the two different groups were given an image and asked to specify the visibility and position of 11 different bird parts. We targeted the citizen scientist turkers with this task by posting the tool on the Lab of Ornithology's Facebook page. The interface for the tool was kept the same between the workers. Figures 7a, 7b, and 7c detail the results of this test. From Figure 7a we can see there is not a difference between the obtainable quality from the two worker groups, and that MTurkers tended to be faster at the task. Figure 7c again reveals that the raw throughput of Mturkers is larger than that of the citizen scientists. The primary benefit of using citizen scientists for this particular case is made clear by their zero cost in Figure 7b.



(a) Sparrow Quiz Scores (b) Shorebird Quiz Scores

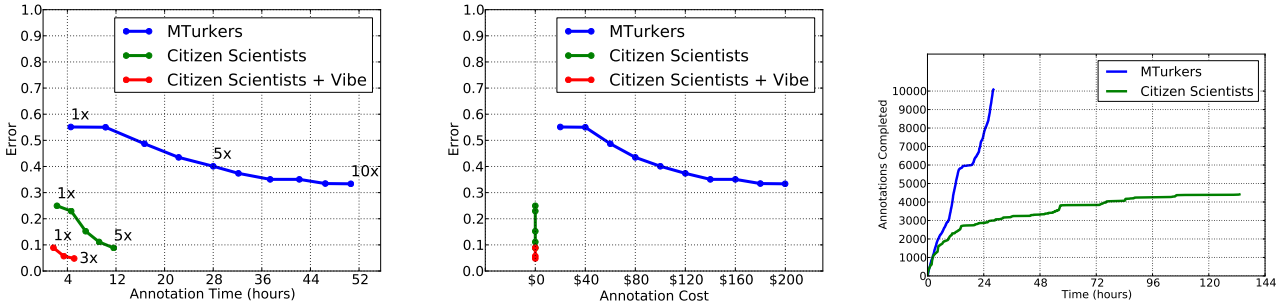
Figure 5: Histogram of quiz scores. Each quiz has 10 images, a perfect score is 10. **(a)** Score distributions for the sparrow quizzes. Random chance per image is 2.7%. **(b)** Score distributions for the shorebird quizzes. Random chance per image is 3.1%. See Section 5

Summary: From these results, we can see that there are clear distinctions between the two different worker pools. Citizen scientists are clearly more capable at labeling fine-grained categories than MTurkers. However, the raw throughput of MTurk means that you can finish annotating your dataset sooner than when using citizen scientists. If the annotation task does not require much domain knowledge (such as part annotation), then MTurkers can perform on par with citizen scientists. Gathering fine-grained category labels with MTurk should be done with care, as we have shown that naive averaging of labels does not converge to the correct label. Finally, the cost savings of using citizen scientists can be significant when the number of annotation tasks grows.

6. Measuring the Quality of Existing Datasets

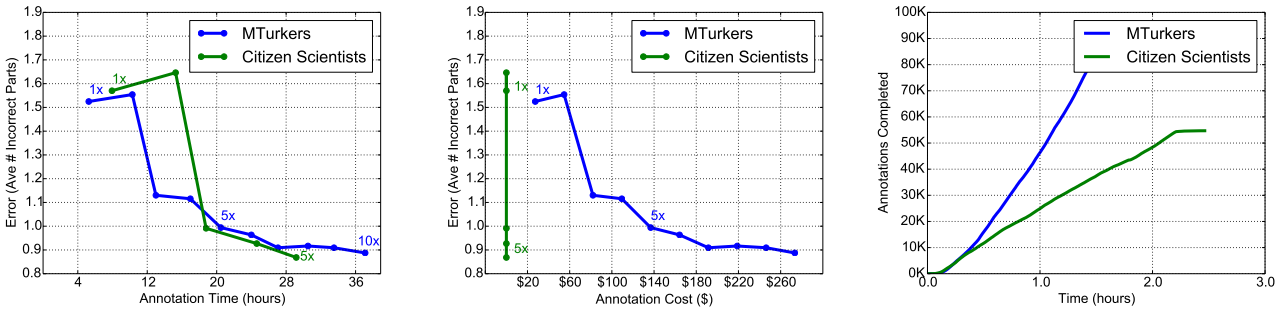
CUB-200-2011 [35] and ImageNet [6] are two popular datasets with fine-grained categories. Both of these datasets were collected by downloading images from web searches and curating them with Amazon Mechanical Turk. Given the results in the previous section, we were interested in analyzing the errors present in these datasets. With the help of experts from the Cornell Lab of Ornithology, we examined these datasets, specifically the bird categories, for false positives.

CUB-200-2011: The CUB-200-2011 dataset has 200 classes, each with roughly 60 images. Experts went through the entire dataset and identified a total of 494 errors, about 4.4% of the entire dataset. There was a total of 252 images that did not belong in the dataset because their category was not represented, and a total of 242 images that needed to be moved to existing categories. Beyond this 4.4% percent error, an additional potential concern comes from dataset bias



(a) Annotation Time (b) Annotation Cost (c) Throughput

Figure 6: **Category Labeling Tasks:** workers used the quiz interface (see Figure 3a) to label the species of birds in images. **(a)** Citizen scientists are more accurate and faster for each image than MTurkers. If the citizen scientists use an expert interface (Vibe), then they are even faster and more accurate. **(b)** Citizen scientists are not compensated monetarily, they donate their time to the task. **(c)** The total throughput of MTurk is still greater, meaning you can finish annotating your dataset sooner, however this comes at a monetary cost. See Section 5



(a) Annotation Time (b) Annotation Cost (c) Throughput

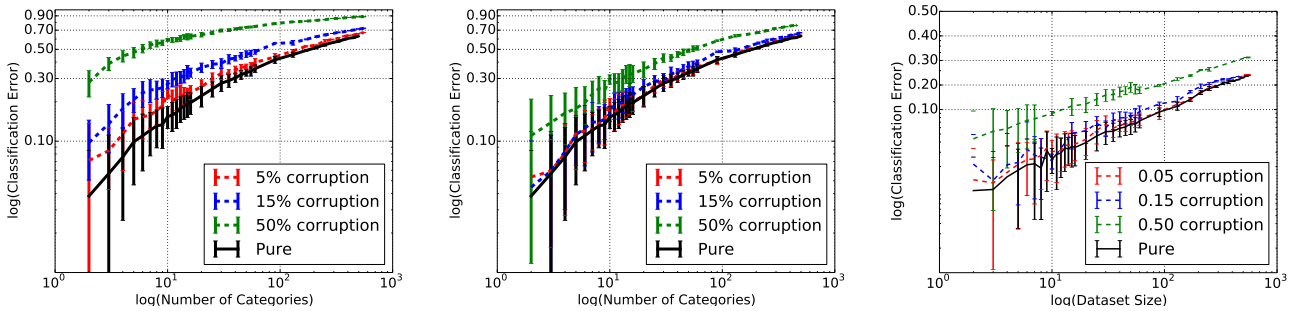
Figure 7: **Parts annotation tasks:** workers used the interface in Figure 3b to label the visibility and location of 11 parts. **(a)**: For this task, as opposed to the category labeling task, citizen scientists and MTurkers perform comparable on individual images. **(b)**: Citizen scientists donate their time, and are not compensated monetarily. **(c)**: The raw throughput of MTurk is greater than that of the citizen scientists, meaning you can finish your total annotation tasks sooner, but this comes at a cost. See Section 5

issues. CUB was collected by performing a Flickr image search for each species and using MTurkers to filter results. A consequence is that the most difficult images tended to be excluded from the dataset altogether. By having experts annotate the raw Flickr search results, we found that on average 11.6% of correct images of each species were incorrectly filtered out of the dataset. See Section 7.2 for additional analysis.

ImageNet: There are 59 bird categories in ImageNet, each with about 1300 images in the training set. Table 1 shows the false positive counts for a subset of these categories. In addition to these numbers, it was our general impression that error rate of ImageNet is probably at least as high as CUB-200 within fine-grained categories; for example, the synset "ruffed grouse, partridge, Bonasa umbellus" had overlapping definition and image content with the synset "partridge" beyond what was quantified in our analysis.

Category	Training Images	False Positives
magpie	1300	11
kite	1294	260
dowitcher	1298	70
albatross, mollymark	1300	92
quail	1300	19
ptarmigan	1300	5
ruffed grouse, partridge, Bonasa umbellus	1300	69
prairie chicken, prairie grouse, prairie fowl	1300	52
partridge	1300	55

Table 1: False positives from ImageNet LSVRC dataset.



(a) Image level features, train+test corruption (b) Image level features, train corruption only (c) Localized features, train corruption only

Figure 8: Analysis of error degradation with corrupted training labels: **(a)** Both the training and testing sets are corrupted. There is a significant difference when compared to the clean data. **(b)** Only the training set is corrupted. The induced classification error is much less than the corruption level. **(c)** Only the training set is corrupted but more part localized features are utilized. The induced classification error is still much less than the corruption level. See Section 7.1

7. Effect of Annotation Quality & Quantity

In this section we analyze the effect of data quality and quantity on learned vision systems. Does the $4\%+$ error in CUB and ImageNet actually matter? We begin with simulated label corruption experiments to quantify reduction in classification accuracy for different levels of error in Section 7.1, then perform studies on real corrupted data using an expert-vetted version of CUB in Section 7.2.

7.1. Label Corruption Experiments

In this experiment, we attempted to measure the effect of dataset quality by corrupting the image labels of the NABirds dataset. We speculated that if an image of true class X is incorrectly labeled as class Y, the effect might be larger if class Y is included as a category in the dataset (*i.e.*, CUB and ImageNet include only a small subset of real bird species). We thus simulated class subsets by randomly picking $N \leq 555$ categories to comprise our sample dataset. Then, we randomly sampled M images from the N selected categories and corrupted these images by swapping their labels with another image randomly selected from all 555 categories of the original NABirds dataset. We used this corrupted dataset of N categories to build a classifier. Note that as the number of categories N within the dataset increases, the probability that a corrupted label is actually in the dataset increases. Figure 8 plots the results of this experiment for different configurations. We summarize our conclusions below:

5-10% Training error was tolerable: Figures 8b and 8c analyze the situation where only the training set is corrupted, and the ground truth testing set remains pure. We see that the increase in classification error due to 5% and even 15% corruption are remarkably low–much smaller than 5% and 15%. This result held regardless of the number of classes or computer vision algorithm. This suggests that the level of annotation error in CUB and ImageNet ($\approx 5\%$) might not be a big deal.

Obtaining a clean test set was important: On the other hand, one cannot accurately measure the performance of computer vision algorithms without a high quality test set, as demonstrated in Figure 8a, which measures performance when the test set is also corrupted. There is clearly a significant drop in performance with even 5% corruption. This highlights a potential problem with CUB and ImageNet, where train and test sets are equally corrupted.

Effect of computer vision algorithm: Figure 8b uses computer vision algorithms based on raw image-level CNN-fc6 features (obtaining an accuracy of 35% on 555 categories) while Figure 8c uses a more sophisticated method [5] based on pose normalization and features from multiple CNN layers (obtaining an accuracy of 74% on 555 categories). Label corruption caused similar additive increases in test error for both methods; however, this was a much higher percentage of the total test error for the higher performing method.

7.2. Error Analysis on Real CUB-200-2011 Labels

The results from the previous section were obtained on simulated label corruptions. We performed additional analysis on real annotation errors on CUB-200-2011. CUB-200-2011 was collected by performing Flickr image search queries for each species and filtering the results using votes from multiple MTurkers. We had experts provide ground truth labels for all Flickr search results on 40 randomly selected categories. In Figure 9, we compare different possible strategies for constructing a training set based on thresholding the number of MTurk votes. Each method resulted in a different training set size and level of precision and recall. For each training set, we measured the accuracy of a computer vision classifier on a common, expert-vetted test set. The classifier was based on CNN-fc6 features from bound-

ing box regions. Results are summarized below:

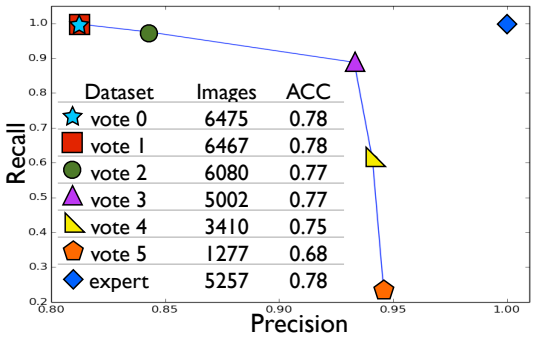


Figure 9: Different datasets can be built up when modifying the MTurker agreement requirement. Increasing the agreement requirement results in a dataset with low numbers of false positives and lower amounts of training data due to a high number of false negatives. A classifier trained on all the images performs as well or better than the datasets that attempt to clean up the data. See Section 7.2

The level of training error in CUB was tolerable: The results were consistent with the results predicted by the simulated label corruption experiments, where a 5-15% error rate in the training errors yielded only a very small (roughly 1%) increase in test error. This provides comfort that CUB-200-2011 and ImageNet are still useful despite label errors. We emphasize though that an error free test set is still necessary–this is still an advantage of NABirds over CUB and ImageNet.

Keeping all Flickr images without any MTurk curation does surprisingly well: This "free dataset" was as good as the expert dataset and slightly better than the MTurk curated datasets. The raw Flickr image search results had a reasonably high precision of 81%. Keeping all images resulted in more training images than the MTurk and expert filtered datasets. If we look at the voter agreement and the corresponding dataset training sizes, we see that having high MTurk agreement results in much smaller training set sizes and a correspondingly low recall.

Quantity can be more important than quality: This underlines the point that having a large training set is extremely important, and having strict requirements on annotation quality can come at an expense of reducing training set size. We randomly reduced the size of the training set within the 40 class dataset and measured performance of each resulting computer vision classifier. The results are shown in Table 2; we see that classification accuracy is more sensitive to training set size than it was to label corruption (see Figure 8b and 9).

Similar results when scaling to more classes: One caveat is that the above results were obtained on a 40 class subset, which was the limit of what was reasonable to ask experts to annotate all Flickr image search results. It is possible that annotation quality becomes more important as the number of classes in the dataset grows. To test this, we had experts go through all 200 classes in CUB-200-2011, annotating all images that were included in the dataset (see Section 6). We obtained a similar result as on the 40-class subset, where the expert filtered dataset performed at about the same level as the original CUB-200-2011 trainset that contains 4-5% error. These results are consistent with simulated label corruption experiments in Figure 8b.

Scale Size	1	1/2	1/4	1/8	1/16	1/32	1/64
ACC	.77	.73	.676	.612	.517	.43	.353

Table 2: Classification accuracy with reduced training set size. See Section 7.2

8. Conclusion

We introduced tools for crowdsourcing computer vision annotations using citizen scientists. In collecting a new expert-curated dataset of 48,562 images over 555 categories, we found that citizen scientists provide significantly higher quality labels than Mechanical Turk workers, and found that Turkers have alarmingly poor performance annotating fine-grained classes. This has resulted in error rates of over 4% in fine-grained categories in popular datasets like CUB-200-2011 and ImageNet. Despite this, we found that learning algorithms based on CNN features and part localization were surprisingly robust to mislabeled training examples as long as the error rate is not too high, and we would like to emphasize that ImageNet and CUB-200-2011 are still very useful and relevant datasets for research in computer vision.

Our results so far have focused on experiences in a single domain (birds) and has resulted in a new publicly available tool for bird species identification. We are currently working on expanding to other types of categories such as shoes and Lepidoptera. Given that over 500 citizen scientists helped provide high quality annotations in just a single domain, working with citizen scientists has potential to generate datasets of unprecedented size and quality while encouraging the landscape of computer vision research to shape around the interests of end users.

9. Acknowledgments

We would like to thank Nathan Goldberg, Ben Barkley, Brendan Fogarty, Graham Montgomery, and Nathaniel Hernandez for assisting with the user experiments. We appreciate the feedback and general guidance from Miyoko Chu, Steve Kelling, Chris Wood and Alex Chang. This work was supported in part by a Google Focused Research Award, the Jacobs Technion-Cornell Joint Research Fund, and Office of Naval Research MURI N000141010933.

References

[1] A. Angelova and S. Zhu. Efficient Object Detection and Segmentation for Fine-Grained Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013. 3

[2] T. Berg, J. Liu, S. W. Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur. Birdsnap: Large-Scale Fine-Grained Visual Categorization of Birds. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2019–2026. IEEE, June 2014. 3

[3] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *Proceedings of the 11th European Conference on Computer Vision: Part I*, ECCV'10, pages 663–676, Berlin, Heidelberg, 2010. Springer-Verlag. 3

[4] B. J. Boom, J. He, S. Palazzo, P. X. Huang, C. Beyan, H.-M. Chou, F.-P. Lin, C. Spampinato, and R. B. Fisher. A research tool for long-term and continuous analysis of fish assemblage in coral-reefs using underwater camera footage. *Ecological Informatics*, 23:83–97, Sept. 2014. 3

[5] S. Branson, G. Van Horn, S. Belongie, and P. Perona. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv:1406.2952*, 2014. 4, 7

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 1, 2, 3, 5

[7] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 580–587. IEEE, 2013. 3

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 3

[9] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(4):594–611, 2006. 3

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv preprint arXiv:1311.2524*, 2013. 1

[11] U. Gneezy and A. Rustichini. Pay enough or don't pay at all. *Quarterly journal of economics*, pages 791–810, 2000. 3

[12] G. Griffin, A. Holub, and P. Perona. Caltech-256 Object Category Dataset. Technical Report CNS-TR-2007-001, California Institute of Technology, 2007. 3

[13] P. G. Ipeirotis and E. Gabrilovich. Quizz: targeted crowdsourcing with a billion (potential) users. pages 143–154, Apr. 2014. 3

[14] P. G. Ipeirotis, F. Provost, V. S. Sheng, and J. Wang. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery*, 28(2):402–441, Mar. 2013. 2

[15] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011. 3

[16] J. Krause, J. Deng, M. Stark, and L. Fei-Fei. Collecting a Large-Scale Dataset of Fine-Grained Cars. In *Second Workshop on Fine-Grained Visual Categorization (FGVC2)*, 2013. 3

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[18] S. Kuznetsov. Motivations of contributors to wikipedia. *ACM SIGCAS computers and society*, 36(2):1, 2006. 3

[19] S. Lazebnik, C. Schmid, and J. Ponce. Semi-local affine parts for object recognition. In *Proc. BMVC*, pages 98.1–98.10, 2004. doi:10.5244/C.18.98. 3

[20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014. 2

[21] J. Liu, A. Kanazawa, D. W. Jacobs, and P. N. Belhumeur. Dog Breed Classification Using Part Localization. In *ECCV*, 2012. 3

[22] C. Long, G. Hua, and A. Kapoor. Active Visual Recognition with Expertise Estimation in Crowdsourcing. In *2013 IEEE International Conference on Computer Vision*, pages 3000–3007. IEEE, Dec. 2013. 3

[23] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 3

[24] G. Martinez-Munoz, N. Larios, E. Mortensen, 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 *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 549–556. IEEE, June 2009. 3

[25] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008. 3

[26] O. Nov. What motivates wikipedians? *Communications of the ACM*, 50(11):60–64, 2007. 3

[27] O. M. Parkhi, A. Vedaldi, C. V. Jawahar, and A. Zisserman. The truth about cats and dogs. In *ICCV*, 2011. 3

[28] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *Proceedings of the 26th Annual international conference on machine learning*, pages 889–896. ACM, 2009. 3

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, 2014. 3

[30] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, page 614, New York, New York, USA, Aug. 2008. ACM Press. 2

[31] A. Sorokin and D. Forsyth. Utility data annotation with Amazon Mechanical Turk. In *2008 IEEE Computer Soci-*

ety Conference on Computer Vision and Pattern Recognition Workshops, pages 1–8. IEEE, June 2008. 2

[32] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1701–1708. IEEE, 2014. 1

[33] L. Von Ahn. Games with a purpose. *Computer*, 39(6):92–94, 2006. 3

[34] L. Von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum. recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008. 3

[35] 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 Institute of Technology, 2011. 2, 3, 5

[36] P. Welinder, S. Branson, P. Perona, and S. Belongie. The Multidimensional Wisdom of Crowds. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2424–2432. Curran Associates, Inc., 2010. 3, 5

[37] P. Welinder and P. Perona. Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 25–32. IEEE, June 2010. 2

[38] J. Whitehill, T.-f. Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, pages 2035–2043, 2009. 3

[39] H.-L. Yang and C.-Y. Lai. Motivations of wikipedia content contributors. *Computers in Human Behavior*, 26(6):1377–1383, 2010. 3