Proceedings, The Fourth AAAI Conference on
Human Computation and Crowdsourcing

Investigating the Influence of Data Familiarity to Improve the Design of a Crowdsourcing Image Annotation System

Danna Gurari
Dept. of Computer Science
University of Texas at Austin

Mehrnoosh Sameki
Dept. of Computer Science
Boston University

Margrit Betke
Dept. of Computer Science
Boston University

Abstract

Crowdsourced demarcations of object boundaries in images (segmentations) are important for many vision-based applications. A commonly reported challenge is that a large percentage of crowd results are discarded due to concerns about quality. We conducted three studies to examine (1) how does the quality of crowdsourced segmentations differ for familiar everyday images versus unfamiliar biomedical images?, (2) how does making familiar images less recognizable (rotating images upside down) influence crowd work with respect to the quality of results, segmentation time, and segmentation detail?, and (3) how does crowd workers' judgments of the ambiguity of the segmentation task, collected by voting, differ for familiar everyday images and unfamiliar biomedical images? We analyzed a total of 2,525 segmentations collected from 121 crowd workers and 1,850 votes from 55 crowd workers. Our results illustrate the potential benefit of explicitly accounting for human familiarity with the data when designing computer interfaces for human interaction.

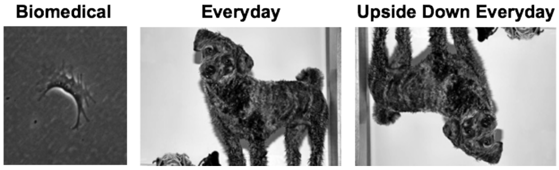


Figure 1: How does familiarity of content influence crowd workers who annotate images? This work extends well-known psychology studies about the effects of image familiarity and flipping on human perception. We explore the influence of these factors on crowdsourced workers asked to 1) demarcate the boundary of an object in an image (segmentation) and 2) judge the ambiguity of the segmentation task. Our findings suggest that "tuning" the familiarity of the content may be an important factor to consider to more effectively employ crowd workers.

Introduction

In a 2013 study, researchers discarded 33,508 crowdsourced image segmentations of everyday content, i.e., 32% of collected data, because the results were not "deemed to be good" (Bell and et al 2013). Conversely, a 2015 study (Gurari and et al 2015) demonstrated that crowdsourced image segmentations on biomedical image content nearly matched the quality of segmentations from domain experts. These contrasting findings are surprising. Why are there differences in the quality of crowd work for the two studies?

One interpretation of the contrasting findings is to infer the crowd workers in the 2013 study (Bell and et al 2013) were not to be trusted. This perspective has been popularized by Bernstein et al. (Bernstein and et al 2015) who posited that crowd workers are often either "Lazy Turkers" or "Eager Beavers." So, as "a rule-of-thumb, roughly 30% of the results from open-ended tasks are poor." However, in the 2013 study (Bell and et al 2013) workers were restricted to the best 26 out of 530 workers whereas the 2015 study (Gurari and et al 2015) employed less stringent worker filtering.

In this work, we hypothesize that the quality of crowd work is not simply a consequence of the industriousness of a crowd worker but also largely a consequence of task design. In particular, we were inspired to examine whether the hidden secret for success for biomedical images lied in the unfamiliarity of the data (Figure 1). To the best of our knowledge, we are the first to study the influence of content familiarity in a crowdsourcing environment. We focus on the open-ended segmentation problem of delineating the boundary of a single object in an image. Our findings are surprising. Not only did crowd workers make fewer egregious errors on unfamiliar biomedical content than familiar everyday content, but also rotating familiar images so content is less recognizable led crowd workers to produce higher quality segmentations significantly faster!

The interest in crowdsourcing the collection of object segmentations for everyday images spans research communities as diverse as computer human interaction (Hara, Le, and Froehlich 2013), computer vision (Gurari et al. 2016), computer graphics (Bell and et al 2013), and multimedia (Galli and et al 2012). Crowdsourced segmentations are currently valuable for offering run-time computations in final image analysis system designs; e.g., navigation systems leverage crowd workers' demarcations of sidewalk obstructions in Google Street View imagery to decide whether sidewalks are inaccessible for wheelchairs (Hara, Le, and Froehlich 2013). In addition, crowdsourced segmentations are exploited at design-time to build better automated methods; e.g., object segmentations serve as training data to teach machine learning systems to identify (i.e., classify) the type of observed object (Jain and Grauman 2013).

The demand for collecting crowdsourced image segmentations is growing as more novel image analysis systems are proposed. This is because object segmentation is a critical precursor for many downstream applications that aim to leverage and interpret the rich abundance of visual data, including to perform:

- *Image Retrieval:* find images in a database that are similar to user-submitted images (Bell and et al 2013).

- *Classification:* differentiate between types of objects such as flowers, cars, and boats (Jain and Grauman 2013).

- *Tracking:* follow objects over time (Paletta and et al 2014).

- *Behavior Analysis:* characterize how objects' shapes or trajectories change over time (Mancini and et al 2014).

The remainder of the paper is organized into seven sections. Related work is reviewed in the next section. Then, we describe our image sets. In the subsequent three sections, we describe three crowdsourcing studies that investigate: 1) How does the quality of crowdsourced segmentations compare for biomedical versus everyday images? 2) How does crowdsourcing segmentation collection compare when images are upright versus upside down? and 3) How do crowd workers judgements differ when they assess the ambiguity of the segmentation task for biomedical versus everyday images? We finish with a discussion and concluding remarks. The key contributions of our work are:

- Analysis of segmentations collected from the crowd for familiar everyday and unfamiliar biomedical images revealing that "mistakes" predominantly arise from task ambiguity rather than worker reliability.
- Analysis showing crowd workers produced higher quality segmentations with less effort when familiar everyday images were upside down and so less recognizable.
- Experiments demonstrating that crowd workers predicted task ambiguity to be greater than observed in practice for unfamiliar biomedical images and less than observed in practice for familiar everyday images.

Related Literature

Commonly, researchers address poor quality crowd work by introducing run-time machinery to improve the results. For instance, mechanisms exist to weed out workers with insufficient training qualifications (Lin and et al 2014), edit/validate crowd work (Bernstein and et al 2015), or mitigate the influence of poor quality work through redundancy (Hara, Le, and Froehlich 2013). Filtering workers has the undesirable consequence of limiting the crowd worker pool which, in turn, reduces the degree to which such a crowdsourcing solution can scale. The remaining aforementioned approaches introduce extra monetary costs as well as delays to acquire results, making such approaches less amenable to "real-time" applications. Unlike these methods, we address concerns about the quality of crowd work by modifying the task at design time rather than at run time.

Limited prior work discusses design-time improvements to yield higher quality crowdsourced segmentations. LabelMe (Russell and et al 2008), one of the earliest and still commonly emulated web-based user interfaces (Hara, Le, and Froehlich 2013; Jain and Grauman 2013; Sorokin and Forsyth 2008), sequentially connects user clicks on the image with straight lines to produce a closed polygon that demarcates the boundary of an object. Building on this framework, recent work has added three features to improve the quality of resulting segmentations: smooth zoom, undo/redo, and automatic pan (Bell and et al 2013; Lin and et al 2014). Alternatively, web-based segmentation annotation tools, such as interactive scissors (Little, Abrams, and Pless 2012) and Click'n'Cut (Carlier and et al 2014), augment the basic user interface with algorithms that interactively refine user-generated segmentations with the aim to clean up boundary imperfections. In contrast to methods that modify the basic segmentation interface, we instead only modify the presentation of images (i.e., rotating images) to yield significantly better segmentations.

Our experiments were partially inspired by psychology experiments that reveal human behavior differs when images are familiar versus unfamiliar and upright versus upside down (Balas and et al 2010; Bartonô and et al 2006; Gold, Mundy, and Tjan 2012; Murray 1997). Although, experimentally, we may understand that humans relate differently to content of differing levels of familiarity, to our knowledge, no work has quantified this impact either for the segmentation task or in the crowdsourcing environment.

More broadly, our work relates to human factors research aiming to improve crowdsourcing theories and methodologies (Lease 2011; Quinn and Bederson 2011). For instance, when choosing how to attract a crowd, important considerations are how crowds behave with different incentives (i.e., pay versus volunteer) (Mao and et al 2013) or cultural biases (Quattrone, Capra, and Meo 2015). In addition, one may make different inferences about results based on the demographics of the worker population (Ross and et al 2010) or perceptions of malicious intent (Gadiraju and et al 2015). Our findings complement existing human factors research by examining how crowd worker behavior relates to data of differing levels of familiarity for two types of tasks.

Finally, our work relates to image annotation crowdsourcing literature. Pioneering works include crowdsourcing the localization of objects with the Peekaboom (Ahn, Liu, and Blum 2006) game, while more recent works include coarse-grained object localization using a paintbrush (Welinder and et al 2010) or bounding box (Su, Deng, and Fei-Fei 2012) as well as pixel-accurate segmentations of specified types of objects (Lin and et al 2014). Our work differs by addressing the general segmentation problem, where the focus is on creating a pixel-accurate delineation of the most prominent object according to human perception, rather than requiring a crowd worker to draw a coarse bounding region (localization) or segment one of a pre-defined set of object categories (semantic segmentation).

Image Sets and Expert Annotations

We conducted our studies on a total of 405 images coming from two image libraries (Alpert and et al 2007; Gurari and et al 2015) that represent familiar everyday and unfamiliar biomedical content. To avoid the challenging problem of how to establish generalized image sets, we selected publicly-shared datasets intentionally designed to represent a diversity of objects in a variety of image conditions[1]. We also chose these datasets because they were designed to only include images that have a single, dominant object of interest and include expert-drawn, pixel-accurate delineations of the object of interest for each image.

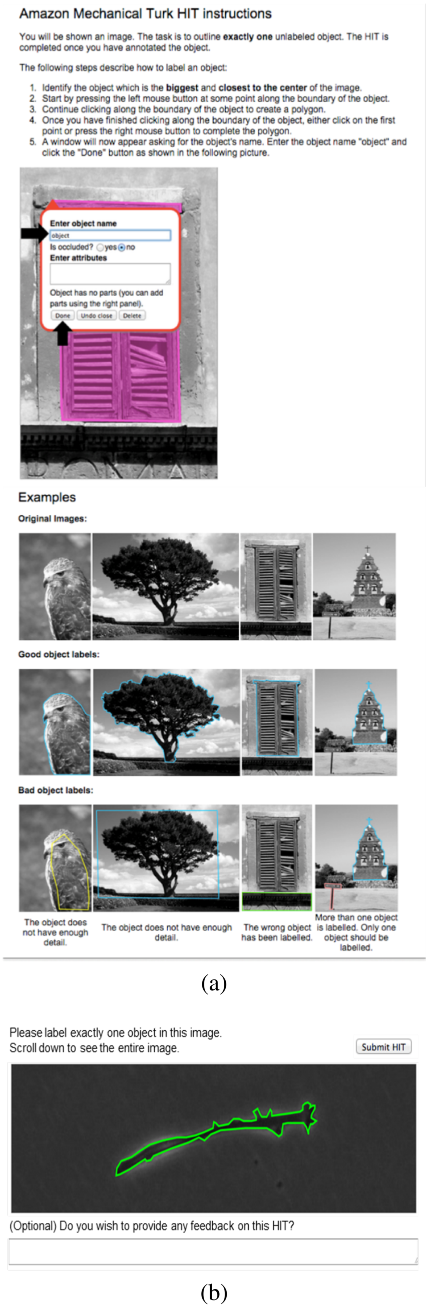
Familiar Everyday Images. We leveraged 100 images that were collected with cameras that detect visible light and so capture content detectable by the naked human eye (Alpert and et al 2007). The designers of the dataset chose images from royalty free databases that "avoid potential ambiguities" regarding the object of interest because the objects of interest differ from the "surroundings by either intensity, texture, or other low level cues." Images show objects such as animals, trees, buildings, and boats. We assume the objects are familiar to crowd workers from daily life experiences regardless of the workers' cultural backgrounds.

Unfamiliar Biomedical Images. We leveraged 305 biomedical images which represent content undetectable to the naked human eye (Gurari and et al 2015). These images come from six datasets created for biomedical research studies that target health care problems such as cancer and heart disease. Three datasets include phase contrast microscopy images that capture a variety of appearances of different types of cells. Two datasets include fluorescence microscopy images that show melanoma cells. One dataset includes magnetic resonance images showing aortas. We assume these images are unlikely to be familiar to a lay person.

Gold Standard Segmentations. To judge the quality of crowdsourced segmentations, we established gold standards using the multiple expert-drawn segmentations per image provided with the datasets. Specifically, for each image, we fuse the segmentations from the multiple experts into a single final segmentation, using majority pixel vote. Our intention was to reduce the impact of biases and mistakes from a single expert on performance analyses.

1: Segment Everyday vs Biomedical Image

Our first aim was to quantify the quality of crowd work for the segmentation task and investigate the impact of data familiarity. Towards this effort, we first describe our experimental methodology including the crowdsourcing system and approach to measure segmentation quality. Then, we describe our experiments to evaluate and compare crowdsourced segmentations collected for familiar everyday images and unfamiliar biomedical images. Our results uncover possible reasons for the contrasting quality findings discussed in the Introduction for everyday (Bell and et al 2013) and biomedical content (Gurari and et al 2015).



Amazon Mechanical Turk HIT instructions

You will be shown an image. The task is to outline **exactly one** unlabeled object. The HIT is completed once you have annotated the object.

The following steps describe how to label an object:

1. Identify the object which is the **biggest** and **closest to the center** of the image.
2. Start by pressing the left mouse button at some point along the boundary of the object.
3. Continue clicking along the boundary of the object to create a polygon.
4. Once you have finished clicking along the boundary of the object, either click on the first point or press the right mouse button to complete the polygon.
5. A window will now appear asking for the object's name. Enter the object name "object" and click the "Done" button as shown in the following picture.

Enter object name

object

Is occluded? ○ yes ● no

Enter attributes

Object has no parts (you can add parts using the right panel).

Examples

Original Images:

Good object labels:

Bad object labels:

The object does not have enough detail.

The object does not have enough detail.

The wrong object has been labelled.

More than one object is labelled. Only one object should be labelled.

(a)

Please label exactly one object in this image.
Scroll down to see the entire image.

(Optional) Do you wish to provide any feedback on this HIT?

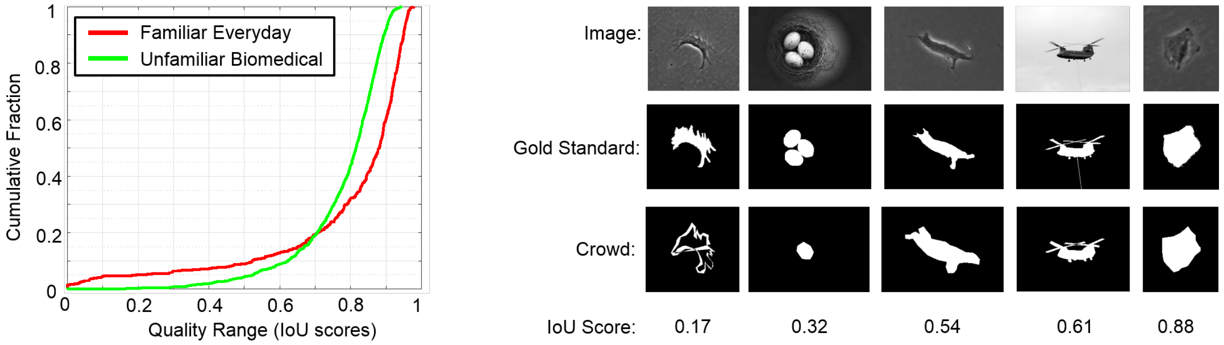
(b)

Figure 2: Crowdsourcing segmentation system (a) instructions and (b) user interface showing a completed annotation from a worker.

Segmentation Crowdsourcing System. Our crowdsourcing environment, described below, entails a two-step process where crowd workers are first shown instructions and then the interface they use for drawing.

Crowdsourcing Platform. We chose the Amazon Mechan-

[1] The image sets can be found at the following links: http://www.wisdom.weizmann.ac.il/∼vision/Seg_Evaluation_DB/1obj/index.html and http://www.cs.bu.edu/∼betke/BiomedicalImageSegmentation



(a)

(b)

Figure 3: Comparison of the quality of crowdsourced segmentations for the familiar everyday and unfamiliar biomedical images. (a) The cumulative distribution function shows for each IoU score on the x-axis the corresponding fraction of crowdsourced segmentations with IoU scores at most that value from the 500 crowdsourced segmentations for the 100 everyday images (red) and 1,525 crowdsourced segmentations for the 305 biomedical images (green). (b) Also shown are exemplar segmentation results for the range of IoU scores.

ical Turk (AMT) marketplace because of “easy access to a large, stable, and diverse subject pool” and “the low cost of doing experiments” (Mason and Suri 2012). In AMT, crowd workers browse among posted jobs, also called HITs (Human Intelligence Tasks), that are paired with a price that we paid upon completion of each HIT. We accepted all AMT workers that had previously completed at least 100 HITs and received at least a 92% approval rating.

Instructions. When a crowd worker reviews our segmentation HIT, (s)he is shown the instructions (Figure 2a). The instructions emphasize that a worker should segment the single object which is the largest and closest to the center of the image. Included are also pictures exemplifying desired and undesired segmentations to clarify that the aim of the task is to create a highly detailed boundary of the single, most prominent object in the image. Examples are intended to address various annotation concerns, such as the common complaint that crowd workers create coarse rather than detailed segmentations (Lin and et al 2014).

Segmentation Tool. After a worker accepts our HIT, the instructions embedded in the AMT webpage are replaced with the segmentation tool (Figure 2b). We employ the freely-available LabelMe code (Russell and et al 2008). With this tool, workers trace the boundary of an object by clicking points on the image which are connected with straight lines. A worker completes the segmentation by clicking on the first clicked point. Then, the worker is prompted with a message allowing him/her to delete the segmentation, in case (s)he made a mistake and so wants to redraw the object. Otherwise, the worker must specify a text label naming the object and click “Done” to submit the completed segmentation. The result is a file that records the sequence of (x, y) image coordinates the crowd worker clicked.

Segmentation Quality Evaluation. To measure the quality of each crowdsourced segmentation, we measured its similarity to the gold standard segmentation. To do this, we adopted the widely-used intersection over union (IoU) metric (Gurari and et al 2015; Hara, Le, and Froehlich 2013; Jain and Grauman 2013) which computes the pixel level similarity of each crowdsourced segmentation and the gold standard segmentation. Formally, this measure is represented as $\frac{|A \cap B|}{|A \cup B|}$ where A represents the set of pixels in the crowd segmentation and B represents the set of pixels in the gold standard segmentation. Resulting scores range from 0 to 1 with larger values indicating greater similarity between the two segmentations.

Experimental Design. We collected segmentations from crowd workers for the 405 familiar everyday images and unfamiliar biomedical images. To capture the variability of segmentation behaviors that may arise due to workers with differing skills, we collected five crowd-drawn segmentations per each image. For each batch of images (i.e., biomedical and everyday), we posted all HITs simultaneously while randomizing the order of jobs. We allotted a maximum of ten minutes to complete each HIT and paid $0.02 per HIT.

We next evaluated the similarity of each crowdsourced segmentation to the gold standard segmentation using the IoU metric. In total, we computed 2,025 IoU scores.

Results. Figure 3a shows the cumulative distribution of IoU scores for both datasets. Figure 3b exemplifies the quality of segmentations associated with the range of IoU scores.

The main distinguishing factor between the quality of crowdsourced segmentations for the two image sets is that crowd workers made egregious errors approximately three to five times more frequently for everyday images than biomedical images (Figure 3a; IoU scores < 0.4). As observed in Figure 3b, mistakes that lead to less than 40% pixel agreement with the gold standard segmentation (i.e., IoU score = 0.4) arise from task ambiguity (i.e., segment one egg versus three eggs?) as well as worker error (i.e., incorrect annotation protocol). From visual inspection of all outliers with IoU < 0.2 (4 from biomedical images, 25 from everyday images), we observe they arise primarily because of ambiguity; i.e., what is the appropriate object to annotate.

We observe more higher quality segmentations (i.e., IoU > 0.7) for everyday images than biomedical images. These results are consistent with our findings for expert annotators. Specifically, when evaluating segmentations from the multiple available expert annotations (included with the benchmarks) against the gold standard segmentations, the median IoU score is 0.85 for the biomedical images and 0.97 for everyday images. As observed in Figure 3b, differences in higher scores often arise due to difficulty in capturing the detail for complicated boundaries, and the biomedical images show objects with highly-complicated boundaries.

In total, 93 unique workers created the 2,025 segmentations with 3 workers in common across the two datasets. The 500 segmentations for the everyday images were created by 44 unique workers, with the average number of jobs completed per worker and its standard deviation being 11 and 23. The 1,525 segmentations for the biomedical images were created by 52 unique workers, with the average number of jobs completed per worker and its standard deviation being 29 and 82.

Discussion. Our findings offer promising evidence that crowdsourcing errors are more frequent for everyday images than biomedical images because crowd workers bring more conflicting opinions regarding how to interpret the segmentation task. For example, when segmenting an image of a basket with eggs, "experts" on the content may be focused on asking whether they should annotate one egg versus three eggs versus the basket holding the eggs. In contrast, when segmenting an image of a cell, crowd workers that are "not experts" on the content may be less distracted by the intricacies of the nucleus, membrane, and other internal structures that they could annotate within a cell. In the next section, we explore whether crowd workers are making more mistakes (i.e., perceiving ambiguity) on everyday images than biomedical images *because* the content is familiar. We design the next study to avoid the possible concern that findings arise due to differences in the studied datasets.

2: Segment Upright Vs Upside Down Image

Our next goal is to learn how making familiar image content less recognizable affects the quality of crowd work. Towards this effort, we evaluate and compare crowdsourced segmentations when the familiar everyday images are upright and upside down. An upside down image of a dog, for example, means the dog's feet will reside where the dog's head would be expected (Figure 1). We also quantify how upside down images influence crowd worker effort, with respect to segmentation time and detail. Finally, we quantify how crowd worker effort relates to the quality of his/her work.

Methods. Our methodology builds off of the crowdsourcing system and approach to measure segmentation quality discussed in Study 1. We describe below the measures we adopted to quantify a crowd worker's efforts to produce a segmentation. Then, we discuss our methodology to measure how a crowd worker's effort relates to the quality of his/her completed work.

Measuring Crowd Worker Effort. Our measures were inspired by the observation that workers with complete freedom in drawing may require more clicks and time in order to accurately capture the detail of objects (i.e., tree). In addition, workers may need to allocate extra attention (i.e., time) to decide with certainty how to separate an object from the background. Therefore, we quantify crowd worker effort with the following three metrics:

- Segmentation Time (T): We leverage logged values in the AMT system that report, for each completed HIT, the lapsed time between when the crowd worker clicked the "Accept HIT" button and the "Submit HIT" button.
- Number of Points (P): We count the number of (x,y) image coordinates recorded in the LabelMe result file describing the segmentation created by the crowd worker.
- Segmentation Speed (TpP): We compute the average time per point as $\frac{\text{SegmentationTime}}{\text{NumberOfPoints}}$.

Correlating Worker Effort with Work Quality. We next measure how a crowd worker's effort relates to the quality of his/her segmentation. We chose to model this relationship with a regression model to capture that segmentation quality ranges on a continuum from nearly perfect (i.e., IoU score close to 1) to seemingly meaningless (i.e., IoU score close to 0). In particular, we trained a multiple linear regression model with n crowdsourced segmentations to learn the model parameters, where each segmentation is described by the three worker effort parameters (T, P, and TpP) and an IoU score indicating the quality of the segmentation. More formally, the model is represented as $y = X\beta$ where y denotes an n-dimensional vector of segmentation quality scores, X denotes a n x k matrix consisting of n vectors that each contain k worker effort descriptors ($k \in \{1, 2, 3\}$), and β denotes a k-dimensional vector of the model parameters (e.g., β_T, β_P, and β_{TpP}) to be learned. At test time, given a new crowdsourced segmentation, the three learned model values (β) are multiplied with their respective worker effort parameters and then summed to establish the predicted segmentation quality (IoU) score. To evaluate how well our models generalize, we performed 10-fold cross-validation.

To evaluate how strongly correlated predicted IoU scores are to actual IoU scores, we computed the Pearson's correlation coefficient (CC). We used the combination of predictions on the 10 test sets from the 10 iterations in crossfold validation. CC values range between +1 and -1 inclusive, with values further from 0 indicating stronger predictive power of a model.

Experimental Design. We collected a total of 10 crowdsourced segmentations per image for the 100 familiar everyday images. Included were the five segmentations per image collected for Study 1. We collected five additional segmentations per image using the same crowdsourcing set-up as in Study 1, except each image was presented upside down (i.e., rotate image by 180 degrees).

Then, for each crowdsourced segmentation, we computed the segmentation time, number of points, segmentation speed, and IoU score. In total, 4,000 computed values were the foundation for subsequent analyses.

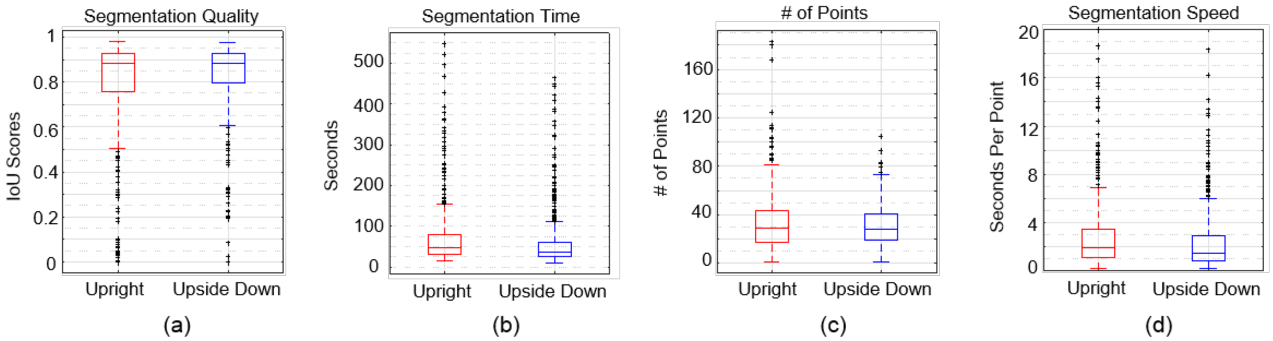


Figure 4: Analysis of 1,000 crowdsourced segmentations collected on 100 everyday images where five crowdsourced segmentations were collected per image when it was upright as well as rotated 180 degrees. For each plot, the central marks of the boxes denote the median values, box edges denote the 25th and 75th percentiles values, whiskers denote the adjacent value to the data point that is greater than one and a half times the size of the inter-quartile range, and black cross-hairs denote outliers. Overall, when images were upside down, (a) segmentation quality was higher, (b) crowd workers took less time to annotate, (c) crowd workers denoted the boundary of objects with more points, and (d) crowd workers annotated at the same speed.

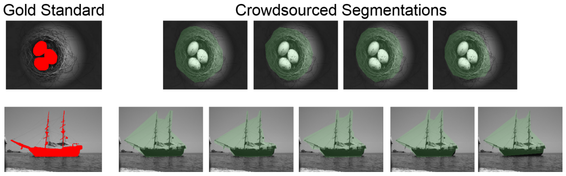


Figure 5: When images are upside down, the majority of egregious errors (i.e., IoU < 0.2) arose from crowd workers (green overlays on right) because they consistently disagreed with experts (red image overlays on left) regarding the true segmentation. (Best viewed in color.)

Next, we analyzed whether there were significant differences in crowd performance for upright and upside down images. For instance, do crowd workers differ in the amount of time they take to annotate when images are upright versus upside down? Inspired by previous work (Smucker, Allan, and Carterette 2007), we chose the distribution-free bootstrap test to compare the 500 segmentations on the upright images and 500 segmentations on the upside down images. The test returns a p-value which indicates the probability of obtaining the two sets of observed results by chance. We infer, with high probability, that observed differences are reflective of a true difference between the two sets of results when the computed two-sided p-value is less than 0.05. We performed four t-tests with respect to each of the four descriptors (T, P, TpP, IoU score).

Finally, we evaluated a total of eight regression models that indicate how crowd worker effort relates to segmentation quality. For both upright and upside down images, we evaluated four models where we analyze how segmentation quality relates to each of the three worker effort cues independently (T, P, TpP) as well as in combination. We used the freely-shared data mining software Weka (Hall and et al 2009) to train, test, and evaluate our models.

Results. Figure 4 illustrates the observed distribution of values for segmentation time, number of points, segmentation speed, and IoU score for both the upright and upside down images. Upside down images led to *higher quality* results and *reduced efforts* from crowd workers, as observed by comparing median scores. Upside down images also led to less variability in quality and effort, as evidenced by smaller interquartile ranges as well as ranges excluding the outliers.

We found that the observed improvement of segmentation quality on upside down images from upright images was significant ($p < 0.05$). Interestingly, as we hypothesized from Study 1, differences were predominantly isolated to “poor quality” segmentations. Specifically, comparing upside down to upright images, the median and top 25th percentile scores are similar while scores demarcating the 75th percentile score and outliers differ by 5% and 10% respectively (Figure 4a). We visually inspected the most egregious segmentation outliers (i.e., IoU < 0.2) for the upside down images and found they arose predominantly because crowd workers disagreed with experts regarding the true delineation of the object of interest (Figure 5).

We also found that the observed *reduction* in effort to create the higher quality segmentations on upside down images was significant ($p < 0.05$). This finding was true for two of the three effort measures: segmentation time and number of points. Crowd workers took 16% less time with an average of 73 seconds for upright images and 61 seconds for flipped images. Crowd workers marked 7% fewer points to create each segmentation for upright images than upside down images (i.e., 33.9 and 31.4 number of points respectively).

Overall, we observed similar trends for how worker effort correlated to segmentation quality on upright and upside down images (Table 1). Comparing CC scores (Table 1, rows 1-4), we observed that the segmentation quality could best be predicted by considering all three worker effort metrics followed by relying exclusively on the segmentation speed, number of points, and segmentation time. In addition, when examining the learned correlation models (Table 1, rows 1-

	Upright Images		Upside Down Images	
Parameters	*IoU =*	*CC*	*IoU =*	*CC*
Time (T)	0.0002T + 0.7767	0.03	-0.0001T + 0.8368	-0.21
# Points (P)	0.0022P + 0.7173	0.23	0.0024P + 0.7592	0.22
Speed (TpP)	-0.0121TpP + 0.8362	0.33	-0.0052TpP + 0.8498	0.33
All	0.0007T + 0.0006P - 0.0146TpP + 0.7726	0.40	0.0002T + 0.0016P - 0.0051TpP + 0.7847	0.34

Table 1: We compare worker effort to the quality of a segmentation when images are upright and upside down. We report learned linear regression models with respect to 1) time to segment, 2) number of points, 3) segmentation speed, and 4) the combination of the three parameters. We also report correlation strengths (CC) for each model. Larger CC scores indicate greater correlation between worker effort and segmentation quality. The main difference between models for upright and upside down images is that comparable segmentation speeds (TpP) leads to considerably worse segmentation quality on upright images.

4), we observed that higher quality segmentations typically arose when a crowd worker took more time (positive valued multiplier β_T), marked more points (positive valued multiplier β_P), and segmented at faster speeds (negative valued multiplier β_{TpP}).

The key difference between learned models for the upright and upside down images lies in the multipliers learned for the segmentation speeds (β_{TpP}). The difference is approximately a factor of 2.5 (Table 1, rows 3 & 4). This means that workers annotating all upright images were observed to produce *worse* quality results than workers annotating all upside down images when spending the same amount of time on the annotation task.

In total, 75 unique workers created the 1,000 segmentations with three workers in common for both datasets[2]. The 500 segmentations for the upright images were created by 44 unique workers, with the average number of jobs completed per worker and its standard deviation being 11 and 23. The 500 segmentations for the upside down images were created by 34 unique workers, with the average number of jobs completed per worker and its standard deviation being 15 and 18.

Discussion. We continue to observe an advantage of fewer segmentation errors from crowd workers when data is less familiar. Moreover, our analysis of crowd mistakes on upside down images raises an important question for future work of what should the truth should be when the majority of the crowd disagree with experts regarding the truth.

Our findings also offer insight into why crowd workers may produce significantly *higher quality segmentations* with considerably *less effort* on upside down images than upright images. In general, we expect crowd workers to be more physically constrained and so annotate at slower speeds when annotating objects with highly-jagged boundaries, such as a tree. Moreover, we expect that crowd workers would find it more challenging to achieve pixel-perfect, high quality segmentations on objects with complicated object boundaries like a tree than rigid objects like a box. In other words, it makes sense that greater worker effort (slower segmentation speeds) could be correlated with lower quality segmentations. However, we hypothesize that this tendency is even more pronounced for everyday images because crowd workers are distracted by additional thoughts about the object, which could include analysis regarding the appropriate level of granularity to segment. In other words, we hypothesize that recognition leads to extra cognitive processing that causes crowd workers to be less focused and therefore effective at performing the segmentation task. We hypothesize that rotating an image upside down disrupts the holistic perception of what an object is (recognition) and so channels a worker's attention to more effectively focus on demarcating the prominent boundaries in an image.

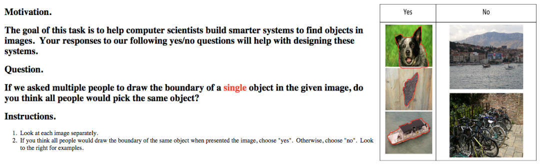
3: Assess Segmentation Ambiguity

In our final study, we ask crowd workers to predict which images would be ambiguous to segment among everyday and biomedical images. In practice, this task could be valuable as a preliminary step to establish when to forego the expensive segmentation task or to expect disagreement. More broadly, this study highlights how data familiarity may influence crowd workers' perceptions.

Methods. We prepared our crowdsourcing system as an internal HIT in AMT. When a crowd worker on AMT reviews one of our posted HITs, (s)he can see the task header and voting task on the same webpage before deciding whether to accept the HIT (Figure 6a).

Our task header includes the problem motivation, task question, and two steps instructing how to perform the task (Figure 6a). We ask workers to answer the following question about an image: "If we asked multiple people to draw the boundary of a single object in the given image, do you think all people would pick the same object?" We intentionally specify criteria that aligns with the segmentation task we used in practice. In an effort to help workers feel their contributions are valued, we state that the long-term aim of

[2] We found that our study results are not impacted by learning effects. Specifically, for both the upright and upside down everyday image experiments, we created a plot showing the quality of each worker's segmentations in the sequential order the jobs were completed (x-axis = job number, y-axis = IoU score). We then computed a best fit line for every worker who completed multiple jobs. A positive slope for a best fit line indicates that segmentation quality improved with more worker experience. However, the median slope for all best fit lines was negligible for the upright images (0.00175) and upside down images (0.000395), with roughly equal portions of slightly positive and negative slopes across all workers.



Motivation.

The goal of this task is to help computer scientists build smarter systems to find objects in images. Your responses to our following yes/no questions will help with designing these systems.

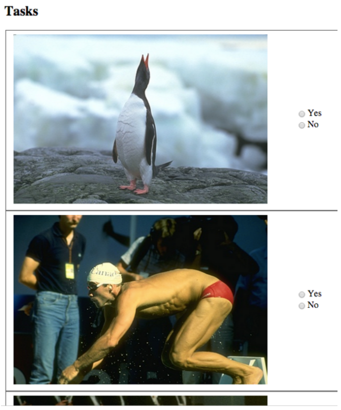
Question.

If we asked multiple people to draw the boundary of a single object in the given image, do you think all people would pick the same object?

Instructions.

1. Look at each image separately.
2. If you think all people would draw the boundary of the same object when presented the image, choose "yes". Otherwise, choose "no". Look to the right for examples.

(a)



Tasks

○ Yes
○ No

○ Yes
○ No

(b)

Figure 6: Crowdsourcing voting system (a) instructions and (b) user interface.

the task is to support computer scientists to build systems. Finally, to clarify the aim of the task, we include pictures exemplifying when to label an image with "Yes" versus "No."

To increase study efficiency, we present a set of five images per HIT. Each image is shown in a column on the left and the crowd worker casts a vote by selecting one of two radio buttons to the right of each image to indicate "Yes" or "No" (Figure 6b). Once a worker completes voting on the five images, the workers clicks a button to submit the results. AMT records the submitted results along with the lapsed time between when the crowd worker clicked the "Accept HIT" button and "Submit HIT" button.

Experimental Design. We collected crowd votes for the 100 familiar everyday images and 270 unfamiliar biomedical images (i.e., BU-BIL:1-5) used in Study 1. To minimize concerns about worker quality, we use the majority vote answer from five collected answers to assign the image label. To avoid concerns about voting biases related to the same groupings and orderings of images per HIT, for each dataset, we randomly assigned five groupings of five images per HIT.

We then quantified the correlation between perceived and actual "easy" and "hard" segmentation problems. We labeled images as *perceptually* difficult when the majority vote indicated there was not a clear object to annotate. We labeled images as *actually* difficult when at least two crowdsourced segmentations of the image (collected in Study 1) were incorrect detections. Incorrect detections are identified by IoU scores below 0.5, which means the crowdsourced segmentation shares less than 50% of pixels with the gold standard segmentation. We report the relationship between perceived and actual labels using the evaluation measures *precision* and *recall*. Precision indicates what fraction of images that were *perceived* as easy to segment were *actually* easy to segment. Recall indicates what fraction of images that were *actually* easy to segment were *perceived* as easy to segment. Both measures return scores ranging from 0 to 1 with better performance reflected by higher scores.

	Precision	**Recall**	**Avg Voting Time**
Everyday	90%	81%	23.4 seconds
Biomedical	97%	51%	20.2 seconds

Table 2: Results illustrating how closely crowd votes regarding the perceived ambiguity of segmenting images matched the actual observed segmentation task ambiguity for everyday and biomedical images. For biomedical images, crowd workers more frequently perceived the segmentation task as more ambiguous than in reality. For everyday images, crowd workers tended to perceive the segmentation task as less ambiguous than in reality.

Finally, we compared time to complete voting HITs for the everyday and biomedical images. As in Study 2, we measure the significance in time differences using a distribution-free bootstrap test and infer observed differences are significant when the computed two-sided p-value is less than 0.05.

Results. Overall, crowd workers typically voted images to be more ambiguous for unfamiliar biomedical image content and less ambiguous for familiar everyday image content (Table 2). Specifically, the absolute difference of 7% in precision (i.e., 90% vs 97%) indicates that crowd workers more often perceived images as providing an unambiguous segmentation task than actually observed in practice. In addition, the absolute difference of 30% for recall (i.e., 81% vs 51%) reveals that crowd workers more often perceived images that are unambiguous to segment as ambiguous for the biomedical images.

Our findings highlight the interesting question of why is popular *perception* of task ambiguity contrasting what is observed in practice regarding the ambiguity of the segmentation task? Exemplar results illustrate how crowd workers judgements are influenced by the type of content (Figure 7). We suspect that, for instance, crowd workers' understanding of the three eggs perhaps blocked them from recognizing the actual ambiguity regarding whether to instead annotate a single egg or vice versa (Figure 7, upper right quadrant). In contrast, we suspect crowd workers lack of recognition of the cell led them to infer the task is ambiguous despite the observation that there is a single, clearly-defined "blob" in the image (Figure 7, lower left quadrant).

Aligning with our findings from Study 2, we found crowd workers took more time for familiar everyday images than the unfamiliar biomedical images ($p < 0.05$). Similar to the previous study, we suspect that recognition of content adds extra cognitive processing, and so time, to complete the task.

In total, 55 unique workers completed all tasks. The 100 voting HITs for everyday images were created by 13 unique workers, with the average number of jobs completed per

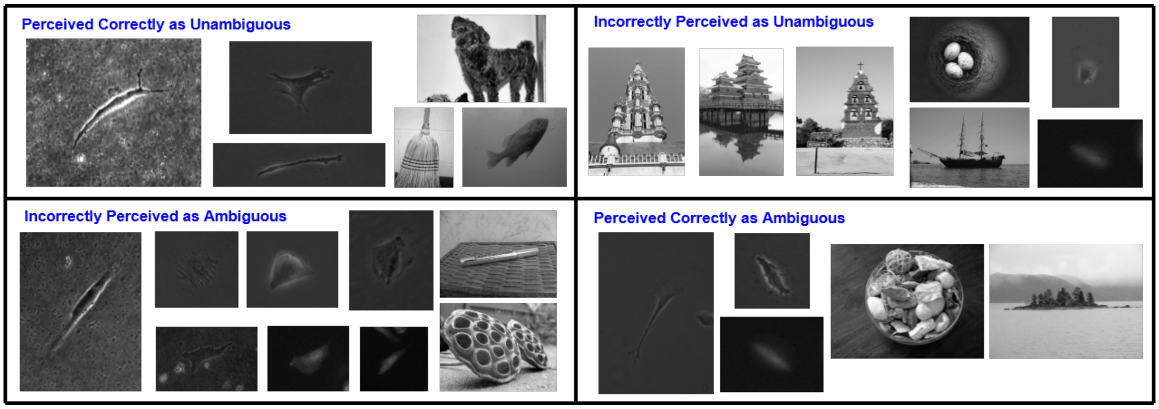

Perceived Correctly as Unambiguous
Incorrectly Perceived as Unambiguous
Incorrectly Perceived as Ambiguous
Perceived Correctly as Ambiguous

Figure 7: Examples illustrating when crowd workers' judgments of the segmentation ambiguity differ and match actual segmentation ambiguity observed for biomedical and everyday images.

worker and its standard deviation being 38 and 54. The 275 voting HITs for the biomedical images were created by 45 unique workers, with the average number of jobs completed per worker and its standard deviation being 34 and 75.

Discussion

While the reliability of humans depends on many factors, our findings offer promising evidence that familiarity of the data may be an important factor to consider when designing human computer interaction systems.

Segmentation Collection. By broadening our analysis of crowd work to include familiar and unfamiliar data, we were inspired to rethink generally held assumptions about how to collect segmentations for familiar image content. In absolute terms, the practical importance of our findings may be great. Rotating familiar images to make content less recognizable would yield higher quality results while eliminating over eight 40-hour work weeks and 7,000 user clicks at the scale of 100,000 crowdsourced segmentations, assuming the findings observed in Study 2. This simple image rotation step may yield great savings and quality improvement for individuals hoping to design novel systems that leverage the rich abundance of visual data.

Trustworthy Crowd Workers for Open-Ended Tasks. We found the 115 unique crowd workers in our studies were generally highly trustworthy. We were pleasantly surprised to learn that less than 0.2% of crowdsourced segmentations (4 out of 2,525) in Studies 1 and 2 appeared to be "malicious" poor quality results. Our findings suggest that most of the remaining egregious outliers were avoidable problems through, for example, better gold standards used for evaluation. Our results offer hints that poor crowd performance may be due to workers' cognitive overload from a complicated task rather than lack of sufficient effort in accomplishing the task.

We offer our study as a meaningful example for the value of trusting crowd workers when designing systems. Teasing out richer information as to why crowd workers make "mistakes" can offer valuable feedback regarding our blind spots where tasks may be ambiguous. Analogous experiments in other domains could include examining how crowd workers perform in text-based or audio-based tasks that are in English when their first language is English versus is not English. Additional experiments could include investigating how a crowd worker's behavior changes over time as (s)he becomes more experienced and perhaps begins to see and/or become blind to task ambiguities.

Influence of Image Familiarity on Humans. Our findings augment existing psychology studies that examine how image familiarity influences human behavior (Balas and et al 2010; Bartonô and et al 2006; Gold, Mundy, and Tjan 2012; Murray 1997). Our findings are surprising in that workers perform the given task *faster* when images are upside down than upright. We posit that our contrasting finding to prior work is due to the distinction in our studied task as, to our knowledge, our work is the first to examine the image segmentation task.

We hypothesize that humans that work with image content that is familiar accrue extra mental processing related to recognition that can both lead to better and worse crowd results. While removing familiarity may eliminate extra cognitive processing competing for a human's attention and so lead to faster and better segmentations (Study 2), removing familiarity may also lead workers to misjudge the ambiguity of tasks (Study 3). Future work will investigate the impact of differing levels of content familiarity for other image analysis tasks, such as object detection, counting, and tracking.

Free-Hand Drawing. After completing our studies, we were surprised to discover that it is a well-established understanding in the art community that individuals draw better when looking at images flipped upside down. Although a different drawing scenario than ours, in the New York Times bestselling book "Drawing on the Right Side of the Brain" the authors recommend to beginners to flip a photograph upside down and then try to draw the contents of the photograph on a blank piece of paper (Edwards 1997). Possible future research would be to collaborate with members of the art community to gain inspiration for additional insights for how to improve the design of human computer interaction systems for image annotation tasks.

Conclusion

We examined the segmentation task and how crowd workers' skills and judgments relate to data of differing levels of familiarity. Our results highlight familiarity can lead to better *and* worse crowd results. Study 1 shows crowd workers make more egregious errors when segmenting familiar than unfamiliar data due to greater perceived task ambiguity. Study 2 reveals better results are obtained when familiar data is artificially made less familiar. Study 3 demonstrates crowd workers predict task ambiguity more accurately when data is familiar than unfamiliar. We hope our segmentation studies will encourage rethinking generally held assumptions that one should *expect* large fractions of poor quality work when crowdsourcing open-ended tasks. Our studies offer promising evidence that researchers can improve designs of crowdsourcing systems by explicitly studying the influence of content familiarity on human behavior.

Acknowledgments

We thank crowd workers for participating in our experiments and gratefully acknowledge funding from the National Science Foundation (IIS-1421943).

References

Ahn, L. V.; Liu, R.; and Blum, M. 2006. Peekaboom: A game for locating objects in images. In *CHI*.

Alpert, S., and et al. 2007. Image segmentation by probabilistic bottom-up aggregation and cue integration. In *CVPR*.

Balas, B. J., and et al. 2010. Personal familiarity influences the processing of upright and inverted faces in infants. *The Developing Human Brain*.

Bartonô, J. J., and et al. 2006. Information processing during face recognition: The effects of familiarity, inversion, and morphing on scanning fixations. *Perception*.

Bell, S., and et al. 2013. OPENSURFACES: A richly annotated catalog of surface appearance. *TOG*.

Bernstein, M. S., and et al. 2015. Soylent: A word processor with a crowd inside. *Communications of the ACM*.

Carlier, A., and et al. 2014. Click'n'Cut: Crowdsourced interactive segmentation with object candidates. In *Workshop on Crowdsourcing for Multimedia*.

Edwards, B. 1997. *Drawing on the Right Side of the Brain*.

Gadiraju, U., and et al. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *CHI*.

Galli, L., and et al. 2012. A draw-and-guess game to segment images. In *PASSAT*.

Gold, J. M.; Mundy, P. J.; and Tjan, B. S. 2012. The perception of a face is no more than the sum of its parts. *Psychological Science*.

Gurari, D., and et al. 2015. How to collect segmentations for biomedical images? A benchmark evaluating the performance of experts, crowdsourced non-experts, and algorithms. In *WACV*.

Gurari, D.; Jain, S.; Betke, M.; and Grauman, K. 2016. Pull the plug? Predicting if computers or humans should segment images. In *CVPR*.

Hall, M., and et al. 2009. The weka data mining software: an update.

Hara, K.; Le, V.; and Froehlich, J. 2013. Combining crowdsourcing and Google street view to identify street-level accessibility problems. In *CHI*.

Jain, S. D., and Grauman, K. 2013. Predicting sufficient annotation strength for interactive foreground segmentation. In *ICCV*.

Lease, M. 2011. On quality control and machine learning in crowdsourcing. In *HCOMP*.

Lin, T., and et al. 2014. Microsoft COCO: Common objects in context. In *ECCV*.

Little, J.; Abrams, A.; and Pless, R. 2012. Tools for richer crowd source image annotations. In *WACV*.

Mancini, M., and et al. 2014. How is your laugh today? In *CHI*.

Mao, A., and et al. 2013. Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In *HCOMP*.

Mason, W., and Suri, S. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods*.

Murray, J. E. 1997. Flipping and spinning: Spatial transformation procedures in the identification of rotated natural objects. *Memory & Cognition*.

Paletta, L., and et al. 2014. Attention in mobile interactions: Gaze recovery for large scale studies. In *CHI*.

Quattrone, G.; Capra, L.; and Meo, P. D. 2015. There's no such thing as the perfect map: Quantifying bias in spatial crowd-sourcing datasets. In *CSCW*.

Quinn, A. J., and Bederson, B. B. 2011. Human computation: A survey and taxonomy of a growing field. In *CHI*.

Ross, J., and et al. 2010. Who are the crowdworkers?: Shifting demographics in mechanical turk. In *CHI*.

Russell, B. C., and et al. 2008. LabelMe: A database and web-based tool for image annotation. *IJCV*.

Smucker, M. D.; Allan, J.; and Carterette, B. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *CIKM*.

Sorokin, A., and Forsyth, D. 2008. Utility data annotation with Amazon Mechanical Turk. *Urbana*.

Su, H.; Deng, J.; and Fei-Fei, L. 2012. Crowdsourcing annotations for visual object detection. In *Workshops at the AAAI Conference on Artificial Intelligence*.

Welinder, P., and et al. 2010. Caltech-UCSD birds 200. Technical report, California Institute of Technology.