Quality Assessment for Crowdsourced Object Annotations

Sirion Vittayakorn
svittayakorn@cs.brown.edu
James Hays
hays@cs.brown.edu

Computer Science Department
Brown university
Providence, RI, USA

As computer vision datasets grow larger the community is increasingly relying on crowdsourced annotations to train and test our algorithms. Due to the heterogeneous and unpredictable capability of online annotators, various strategies have been proposed to "clean" crowdsourced annotations. However, these strategies typically involve getting *more* annotations, perhaps different types of annotations (e.g. a grading task), rather than computationally assessing the annotation or image content. In this paper we propose and evaluate several strategies for automatically estimating the quality of a spatial object annotation. We show that one can significantly outperform simple baselines, such as that used by LabelMe, by combining multiple image-based annotation assessment strategies.

We believe this paper is the first to explore *annotation scoring functions*. An annotation scoring function takes an image and user's annotation and returns a real-valued score indicating the quality of that annotation. Such functions allow a database to be filtered to the degree required by a specific algorithm. For instance, a silhlouette-based detectors may require very strict filtering while fixed aspect ratio detectors like Delal and Triggs need only the spatial extent of each object. To learn contextual relationships one might only need the rough location of objects or one might only be concerned with whether objects co-occur in the same scene.

By our definition, an annotation scoring function does not consider the supposed *category* of an annotation. We do not consider the situation where an annotation is poor quality because the annotator gave it the wrong label (e.g. a user accurately segmented a dog but then labeled it a car). In our experience, this is a very rare situation.

We investigate five annotation scoring functions and have each method rank hundreds of crowdsourced user annotations.

An ideal annotation scoring function would rank annotations in accordance with a "ground truth" quality ranking. To quantify how good a ranking is, we must first build a dataset containing pairs of crowdsourced annotations with their corresponding ground truth annotation. We collect from LabelMe 200 pairs of object images and their user annotations from 5 categories – person, car, chair, dog, and building. These categories are picked because they are among the most common objects in LabelMe and because they are distinct in size, shape, and annotation difficulty. For these 1,000 total images we establish a ground truth spatial annotation in order to quantify how good each user annotation is. To establish a ground truth quality score for user annotations we need to compare user annotations to our ground truth annotations. We use a weighted combination of the PASCAL VOC overlap score and a "Shape Context" inspired total distance between annotation boundaries to compare pairs of annotations. We then sort the user annotations by this score to get a ground truth ranking.

We investigate the following annotation scoring functions:

- *Baseline: number of control points* – introduced by LabelMe, the score of an annotation is proportional to the number of control points in the bounding polygon.
- *Baseline: annotation area* – score is proportional to image area occupied by the user annotation.
- *Edge detection* – score is proportional to the amount of dilation or erosion of the user-specified boundary necessary to get the best alignment with image edges.
- *Bayesian Matting* – score encodes how well a matting operation agrees with the user-specified boundaries.
- *Object Proposal* – score is proportional to the ranking of the user segmentation with respect to other proposed objects found by a generic object detector.

We also investigate the performance of combinations of scoring functions. We tried numerous combinations and found that none could exceed the performance of the Bayesian matting and edge scoring functions together. This "final" combination scores the highest for every category.

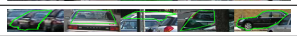
Ranking method	Example of best ranked	Example of worst ranked
Ground truth		
Number of points		
Annotation area		
Edge detection		
Bayesian Matting		
Object proposal		
Final		

Table 1: The five highest and lowest ranked user annotations according to the ground truth ranking and each annotation scoring function.The green and red bounding boxes are the user's and ground truth bounding boxes respectively, while the blue contours are the object's contour from edge detection approach.

Catergory	Rank correlation					
	Points	Area	Edge	Bayesian	Proposal	Final
Car	0.5216	0.4356	0.5972	0.3848	0.0817	**0.5999**
Chair	0.6758	0.6519	0.6132	0.6780	0.0190	**0.6947**
Building	−0.3874	0.4271	0.4055	0.2030	0.0386	**0.5214**
Person	0.5503	0.4386	0.5716	0.7036	0.0394	**0.7072**
Dog	0.6070	0.2367	0.6932	0.6503	0.0468	**0.7689**
Average	0.3935	0.4380	0.5761	0.5239	0.0232	**0.6584**

Table 2: The rank correlation between ground truth ranking and rankings produced by various annotation scoring functions.

Results. Given the input image and its corresponding annotation, the annotation scoring function returns a score corresponding to the estimated quality of that annotation. We then generate the overall annotation ranking from each scoring function and evaluate these rankings by calculating the Spearman's rank correlation against the ground truth ranking. We visualize the rankings that result from each scoring function in Table 1. Table 2 shows the rank correlations for each annotation scoring function broken down by category.

These results show that the *number of control points* is indeed a good predictor of annotation quality, although the *annotation size* performs equally well and is more broadly applicable (e.g. not restricted to polygonal annotations). Both the *Bayesian matting* and *edge detection* scoring functions have a high rank correlation. We expected the *object proposal* scoring function to perform better. But, in fact, the algorithm is answering a somewhat different question than what we are interested in. The object proposal evaluates *how likely is this segment to be an object?* and the question we are interested in is *how accurately annotated is this object segment?*. We know that all user annotations are object segments – none of the annotations are so adversarial as to contain no object. The object proposal method may be intentionally *invariant* to annotation quality.

Summary. In this paper we show that numerous annotation scoring functions, some very simple, can produce annotation rankings that are reasonably similar to the ground truth annotation quality rankings. We propose new annotation scoring functions and show that, in isolation or combination, they outperform the simple LabelMe baseline. To evaluate these scoring functions we produced an extensive database with 1,000 pairs of crowdsourced annotations and meticulous ground truth annotations. As a sanity check, we show that user annotations ranked highly by our annotation scoring functions are more effective training data than random user annotations in a simple object classification task. We share our dataset and look forward to the development of better annotation scoring functions which will make crowdsourced annotations easier to leverage.