

Quality Assessment for Crowdsourced Object Annotations

SIRION VITTAYAKORN

JAMES H. HAYS

Computer Science Department
Brown University
Providence, RI, USA

svittayakorn | hays@cs.brown.edu

May 9, 2011

Abstract

As computer vision datasets grow larger, the community is increasingly relying on crowd-sourced annotations to train and test their algorithms. Since the capability of online annotators is considered varied and unpredictable, many strategies have been proposed to “clean” crowd-sourced annotations. However, these strategies typically require *more* annotations, rather than using the annotation or image content itself. In this paper we propose and evaluate several strategies for automatically estimating the quality of an object annotation. Finally, we show that we can significantly outperform simple baselines by combining multiple image-based annotation assessment strategies.

1 Introduction

In recent years, there has been interest in more diverse and larger object data sets such as LabelMe [3] and ImageNet [6]. These data sets contain spatial annotations of thousands of object categories in hundreds of thousands of images. In order to build these kinds of data sets, researchers must *crowdsource* the annotation effort to non-experts. Due to the unpredictable capability of online annotators, these annotations can not be trusted.

For this reason, numerous strategies are used to “clean” or filter the raw annotations. Sorokin and Forsyth [1] have suggested the use of the following strategies; “Gold standard”, “Grading”, and “Consensus”. These strategies are found to be effective, but they also incur additional necessary costs and ultimately add complexity to the database creation.

LabelMe, on the other hand, uses a simple but surprisingly effective heuristic to rank user annotations – the number of control points in each annotation. This heuristic is an instance of what we will refer to as the *annotation scoring functions*. Given a user annotation and an image, the *annotation scoring functions* will return a real-valued score, which estimates the quality of the given annotation. These functions are useful because they do not dispense with any user annotations and allow a database to be filtered in accordance with the demands of a specific algorithm.

We believe this paper is the first to explore *annotation scoring functions*. We hope these methods will help researchers utilize large-scale, crowdsourced data sets in both computer vision and computer graphics.

An ideal annotation scoring function would rank annotations in accordance with a “ground truth” quality ranking. We will discuss how we define ground truth later in this paper. The scoring function should be unbiased. For instance, one might consider using a pedestrians detector to assess the quality of person annotations, but this circularity would bias the user toward using person instances,

which are *already easy to recognize* and thus defeat the purpose of using larger datasets. To help avoid such biases, we examine only category agnostic scoring functions. If a scoring function only expresses confidence for “easy” instances, our evaluation method will penalize it because its ranking will diverge from the ground truth ranking.

Although the LabelMe [3] baseline scoring function, which counts control points, is surprisingly effective, it does not consider the image itself. Therefore, we should be able to score annotations more intelligently by considering whether an annotation has likely been given the local or global image content. While there is little previous work on “annotation scoring functions” per se, there is considerable research on edge detection, image segmentation, and more recently generic object detection which one might expect to be informative on the subject of annotation quality. In Section 3 we operationalize such techniques for use in annotation assessment and compare their performance to simple baselines. A combined scoring function leveraging multiple cues significantly outperforms any existing baseline.

2 Definitions and Dataset

By taking an image and user’s annotation, an annotation scoring function returns a real-valued score indicating the quality of that annotation. By our definition, we do not consider the situation where an annotation is determined to be poor quality because the annotator gave it the wrong label. In this paper, we investigate five annotation scoring functions and have each method ranked by hundreds of crowdsourced user annotations. We then compare those rankings of annotation quality to a ground truth ranking standard using Spearman’s rank correlation coefficient [11] and Kendall’s rank correlation coefficient [9]. As a result, it is first necessary to build a dataset containing pairs of crowdsourced annotations with their corresponding ground truth annotation.

2.1 User Annotations

We collect spatial object annotations from LabelMe [3] which are closed polygons entered from a web-based interface. All of the scoring functions, except for the LabelMe baseline, take a binary mask as input rather than a polygon. Thus our functions are suitable to any labeling interface that produces either segmentation or a bounding box.

We collect 200 random images and 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.

2.2 Ground truth annotations

We need to establish a ground truth spatial annotation in order to know how good each user annotation is. The red contour in Figure 1 shows a sample of the ground truth annotation for five objects in our dataset. Since we are assessing LabelMe annotations, our ground truth is specified in strict accordance with the LabelMe guidelines. Although different datasets have different rules about how to annotate occlusions, we do not believe these rules had a significant impact on user annotation quality, except for the building category where user annotations violated these rules more often. The actual ground truth segmentation was manually performed. Since our annotation protocol requires a high level of consistency, we verify the consistency of every image several times. We hope that this time consuming dataset might be useful for the community and we will release it in the future.

2.3 Comparing Annotations

To establish a ground truth quality score for user annotations we need to compare user annotations to our ground truth annotations. Although there are numerous methods in the vision literature for comparing such shapes, the most widely used is the PASCAL VOC overlap score [8]. However, the



Figure 1: Example of the ground truth dataset from 5 categories. The red bounding box is the ground truth bounding box of the object.

overlap score is not entirely satisfying for arbitrary shapes because it gives no special attention to boundaries or small protrusions. Therefore, we also use the *Euclidean distance score*, inspired by the “Shape Context” method [2], which will emphasize the boundary agreement of two annotations.

2.3.1 Overlap Score

In PASCAL VOC competition, algorithms are evaluated based on the overlap area of the two annotations. For two objects B_u and B_v , the overlap score is:

$$score_p = \frac{area(B_u \cap B_v)}{area(B_u \cup B_v)} \quad (1)$$

where $B_u \cap B_v$ denotes the intersection of two bounding boxes or binary masks and $B_u \cup B_v$ denotes their union.

2.3.2 Euclidean Distance Score

The Euclidean distance score measures the distance between two annotation boundaries. In our experiment, we sample $m = 300$ random points along each annotation boundary then calculate the Euclidean distance between all possible pairwise correspondences, resulting in an m by m matrix of Euclidean distances. Given this matrix, the Kuhn-Munkres algorithm [10] returns the m assignments which have the minimum total Euclidean distance. In Figure 2 left, the red and green points are the random points from the user and ground truth annotations, respectively, while the blue lines show the Euclidean distance between the assigned pairwise. Given these correspondences for annotations B_v and B_u where $(X_i, Y_i) \in B_u, (x_{i'}, y_{i'}) \in B_v$ we simply sum the pairwise distances according to Equation 2. Finally, we normalize these distances into $[0,1]$. Due to the fact that larger Euclidean distances are associated with lower annotation quality, we define our Euclidean distance score as 1 minus the normalized summed distance (Equation 3).

$$dist = \sum_i \sqrt{(X_i - x_{i'})^2 + (Y_i - y_{i'})^2} \quad (2)$$

$$score_e = 1 - \frac{dist}{max(dist)} \quad (3)$$

And $max(dist)$ is the maximum Euclidean distance of that category.

2.3.3 Ground Truth Quality Ranking

To create our per-category ground truth ranking, we compare each user and ground truth annotation using a combination of the overlap score and the Euclidean distance score:

$$score = (w_1 \times score_e) + (w_2 \times score_p) \quad (4)$$

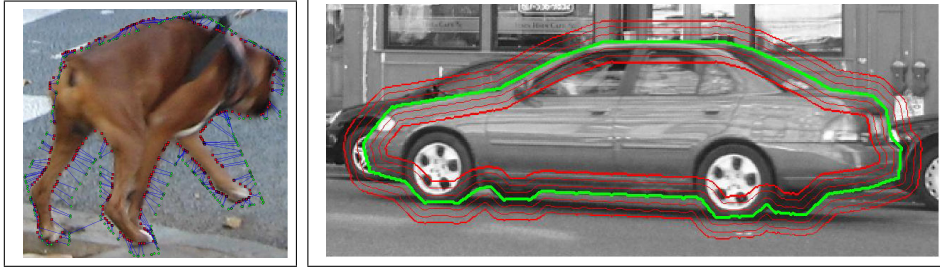


Figure 2: Left: The blue line shows the Euclidean distance between the ground truth bounding box (red points) and the user’s bounding box (green points). Right: The green contour is the user’s bounding box, while the red contours are the candidate object’s contour base on the user’s bounding box.

where $score_e$ is the Euclidean distance score, $score_p$ is the overlap score between the user and ground truth bounding box while w_1 and w_2 were empirically found to produce intuitive distance computations between annotations.

3 Annotation Scoring Functions

Given an image and annotation, we evaluate the quality of each annotation using the annotation scoring function and then rank all of the annotations within each object category. We investigate five annotation scoring functions based on different methods; number of annotation control points [3], annotation size, edge detection, Bayesian matting [5] and object proposal [7].

3.1 Baselines: Control points and Annotation Size

Starting with the simple baseline – the quality of an annotation is proportional to the number of control points in the bounding polygon. This baseline, used by LabelMe [3], works surprisingly well – a large number of control points usually indicates that the user made an effort to closely follow the object boundary.

Moreover, we also propose an even simpler baseline relating annotation size and quality of annotation. We believe that the larger an object is, the better the annotation is. Because it is easier for a user to annotate a large, high resolution object than a smaller one. We calculate the size baseline score as the percentage of image area occupied by the user annotation.

3.2 Edge Detection

Observantly, a good annotation will tend to follow image edges. But this is not strictly true of most user annotations, even those which are qualitatively quite good. Although users have a tendency to annotate *outside/inside* the actual object boundary by a few pixels even, they are tracing the shape quite accurately. To accommodate for this we consider not just the user bounding polygon, but also several dilations and erosions of this bounding region as shown in Figure 2, right.

For each such dilated or eroded user boundary we measure the overlap with filtered image edges. We first run Canny edge detection [4] on the image. Any short fragments which often correspond to texture are discarded. Then, we group the resulting edge fragments into contours. Moreover, we also discard fragments whose dominant orientation differs from the orientation of the nearby user bounding box. Finally, we find the boundary which has the most edge overlap and assign a score to the user annotation in proportion to how much this best boundary was dilated or eroded.

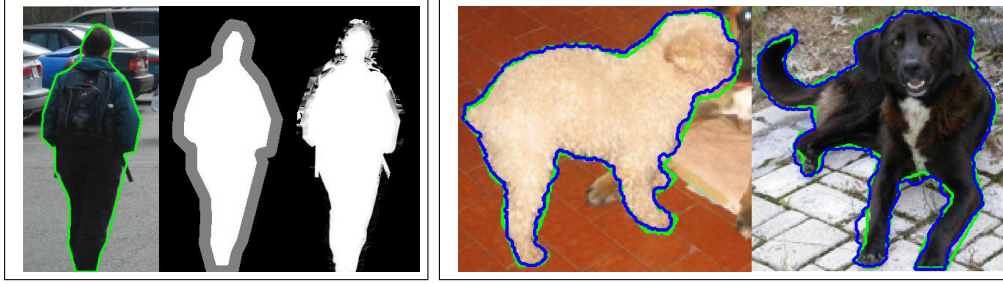


Figure 3: Left: The input image with annotation, its trimap and its alpha map from Bayesian Matting approach. The black, gray and white regions in the trimap are the background, unknown and foreground region, respectively. Right: The blue contours are regions, which are similar to the user’s bounding box (green contour) from the object proposal approach.

3.3 Bayesian Matting

The fourth annotation scoring function is based on the Bayesian matting algorithm [5]. Bayesian matting models the object and background color distributions with spatially-varying mixtures of Gaussians and assumes a fractional blending of the object and background colors to produce the final output. It then uses a maximum-likelihood criterion to estimate the optimal opacity, object and background simultaneously.

Given a *trimap* with three regions: “background”, “object” and “unknown” where the background and object regions have been delineated conservatively. We construct these three regions based on the user annotation. The region outside the user annotation is constrained as background while the inside region is constrained as object. All pixels within a small, fixed distance of the user boundary are marked as unknown (see Figure 3, left). Bayesian matting returns an *alpha map* with values between 0 to 1 where 0 means background and 1 means object. The fractional opacities, α , can be interpreted as a confidence of that pixel belonging to the object. Since, we believe that a good annotation is expected to agree with localized image segmentation, we then use these α values to compute an annotation score as follows:

$$score = \sum \alpha_{in} / area_{in} - \sum \alpha_{out} / area_{out} \quad (5)$$

where α_{in} and α_{out} are the opacity of each pixel within the unknown region inside and outside the annotation respectively while $area_{in}$ and $area_{out}$ are the area of the unknown region inside and outside the annotation.

3.4 Object Proposal

The last annotation scoring function is based on the object proposal approach by Endres and Hoiem [7]. In their work, they propose a category-independent method to produce a ranking of potential object regions. Given an input image, their approach generates a set of segmentations by performing graph cuts based on a seed region and a learned affinity function. Then, these regions are ranked using structured learning based on various cues. Top-ranked regions are likely to be good segmentations of objects. Thus, we assume that if the user annotation is similar to one of the top ranked object proposals, then that suggests the annotation may be of high quality. To score a user annotation, we first find the object proposal with the smallest Euclidean distance and overlap score to the user annotation. The score is the rank of that object proposal divided by the total number of object proposals. In addition, we also try to have the object proposal method [7] directly score the segmentation provided by the user, and not explore multiple segmentations, but the performance became worse.

4 Rank Correlation

At this point we have a ground truth ranking for the user annotations within each category as well as a ranking from each scoring function. We measure how well each ranking agrees with the ground truth using Spearman’s rank correlation coefficient or Spearman’s rho, ρ [11] and Kendall’s rank correlation coefficient or Kendall’s Tau, τ [9].

4.1 Spearman’s rank correlation

The n raw scores X_i, Y_i are converted to ranks (x_i, y_i) and the differences $d_i = (x_i - y_i)$ between the ranks of each observation of the two variables are calculated. If there are no tied ranks, then ρ is given by:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (6)$$

4.2 Kendall’s rank correlation

Let (x_i, y_i) be a set of joint observations from two random variables of n raw scores X_i and Y_i respectively, such that all the values of (x_i) and (y_i) are unique. Any pair of observations (x_i, y_i) and (x_j, y_j) are said to be

- **concordant** if the ranks for both elements agree: that is, if both $x_i > x_j$ and $y_i > y_j$ or if both $x_i < x_j$ and $y_i < y_j$
- **discordant** if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$
- **neither concordant, nor discordant** if $x_i = x_j$ or $y_i = y_j$

The Kendall’s rank coefficient is defined as:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)} \quad (7)$$

An increasing rank correlation coefficient implies increasing agreement between rankings. The coefficient is inside the interval $[-1, 1]$ and can be interpreted as follows:

- 1 if the agreement between the two rankings is perfect; the two rankings are the same.
- 0 if the rankings are completely independent.
- -1 if the disagreement between the two rankings is perfect; one ranking is the reverse of the other.

5 Results

Given the input image and its corresponding annotation, the annotation scoring function returns a score range between $[0, 1]$, which estimates the quality of that annotation. Using these scores, we then generate the overall annotation ranking from each scoring function and evaluate these rankings by calculating the rank correlation against the ground truth ranking. Table 1 shows the rank correlations for each annotation scoring function broken down by category. “Building” stands out as the most difficult category for most scoring functions. We believe this is because LabelMe users did not reliably follow the annotation rules – they sometimes included multiple buildings in a single annotation, or handled occlusion incorrectly. This led to a significant disparity between some user and ground truth annotation pairs.

These results show that the number of control points is indeed a good predictor of annotation quality except for the building category. We think that this metric is uniquely bad for the building

Category	Rank correlation	Annotation Scoring function					
		Points	Size	Edge	Bayesian	Proposal	Final
Car	Spearman's	0.5216	0.4356	0.5972	0.3848	0.0817	0.5999
	Kendall's	0.3469	0.3029	0.4191	0.2296	0.0443	0.4222
Chair	Spearman's	0.6758	0.6519	0.6132	0.6780	0.0190	0.6947
	Kendall's	0.4954	0.4406	0.4231	0.4820	0.0161	0.4996
Building	Spearman's	-0.3874	0.4271	0.4055	0.2030	0.0386	0.5214
	Kendall's	-0.2628	0.2901	0.2837	0.1380	0.1462	0.3688
Person	Spearman's	0.5503	0.4386	0.5716	0.7036	0.0394	0.7072
	Kendall's	0.3853	0.3060	0.4025	0.5163	0.0255	0.5217
Dog	Spearman's	0.6070	0.2367	0.6932	0.6503	0.0468	0.7689
	Kendall's	0.4129	0.1859	0.5077	0.4508	0.2359	0.5481
Average	Spearman's	0.3935	0.4380	0.5761	0.5239	0.0232	0.6584
	Kendall's	0.2755	0.3051	0.4072	0.3697	0.0936	0.4721

Table 1: The rank correlation between ground truth ranking and other rankings



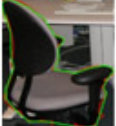


Bounding box	Number in the ranking (out of 200 images)						
	Ground truth	#control points	Bounding box's size	Edge Detection	Bayesian Matting	Object proposal	Final
	118	143	70	84	109	3	117
	85	106	107	73	101	96	92
	21	6	35	38	32	113	29
	1	5	3	19	77	99	18
	186	25	166	155	60	43	159

Figure 4: Selected user annotations and their rank according to each scoring function.

category and other categories such as window or door which the objects are usually in the rectangular shape. Thus many accurate annotations have only 4 control points. Interestingly, the even simpler annotation size baseline performs about as well as the control point baseline and is more directly applicable to annotation protocols that don't use polygons such as freehand lassoing.

Both the Bayesian matting and edge detection scoring functions have a high rank correlation. We believe the Bayesian matting approach handles foreground occlusions poorly and thus does not perform well in the building category.

However, 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 so as to contain no object. The object proposal methods may be intentionally *invariant* to the annotation quality, because the automatic segmentations that it considers cannot be expected to be particularly accurate.

Finally, we investigate the performance of combinations of scoring functions. It is simple to “average” multiple rankings by averaging the object's score, and then re-ordering. 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. We were surprised that the baseline scoring functions did not reinforce the more advanced, image-based functions. We also have visualized that specific annotation instances are ranked by each function in Figure 4 and the rankings that result from each scoring function in Figure 5-9.

6 Conclusion

In this paper we obviously 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. We will share this dataset in the future and we look forward to the development of better annotation scoring functions, which will make crowdsourced annotations easier to leverage.

References

- [1] D. F. Alexander Sorokin. Utility data annotation with amazon mechanical turk. *First IEEE Workshop on Internet Vision at CVPR 08*, 2008.
- [2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, April 2002.
- [3] K. P. M. C. Russell, A. Torralba and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77:157–173, May 2008.
- [4] J. Canny. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-8(6):679–698, nov. 1986.
- [5] Y.-Y. Chuang, B. Curless, D. H. Salesin, and R. Szeliski. A bayesian approach to digital matting. In *Proceedings of IEEE CVPR 2001*, volume 2, pages 264–271. IEEE Computer Society, December 2001.
- [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.

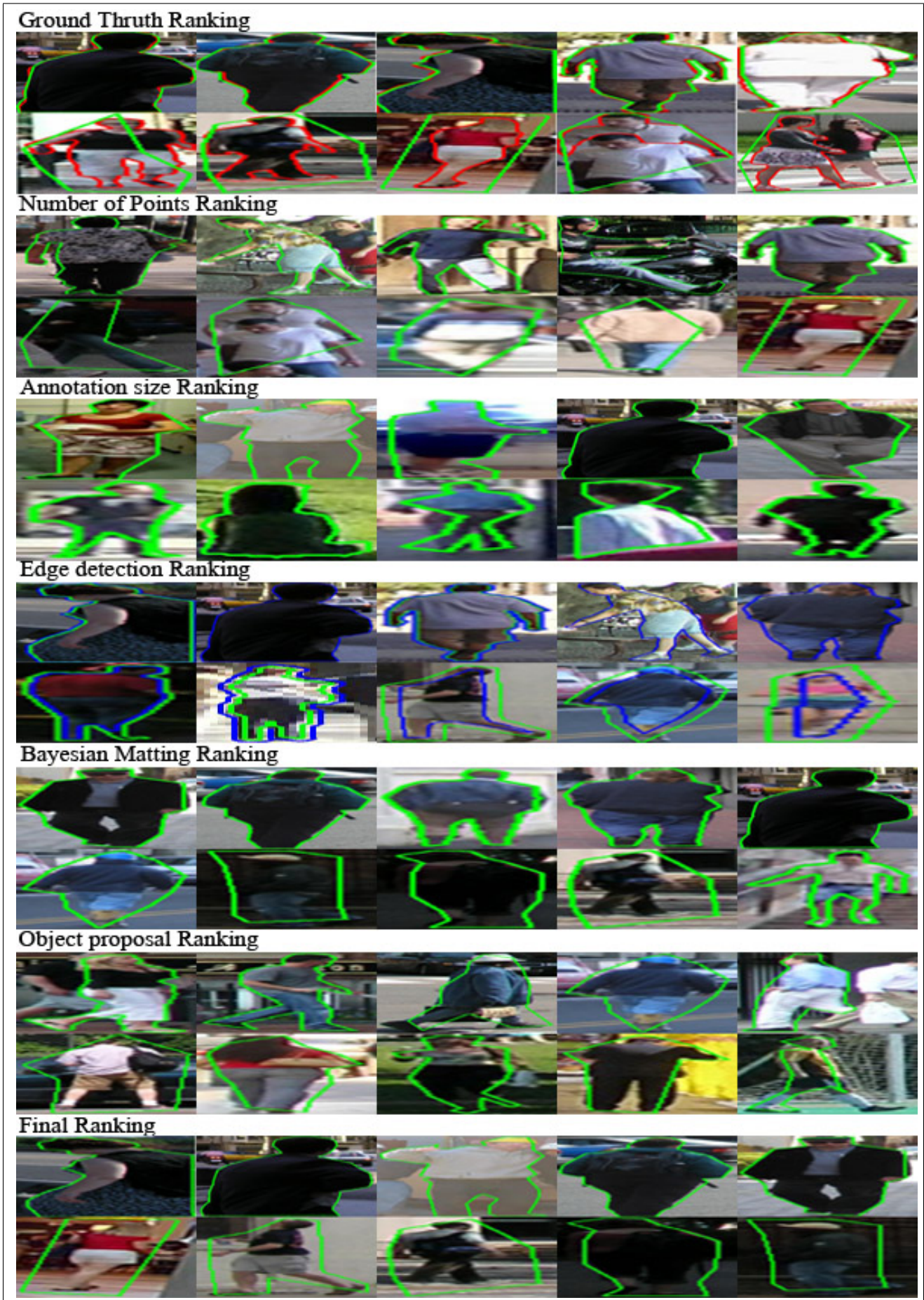


Figure 5: The five highest and lowest ranked user annotations according to the ground truth ranking and each annotation scoring function. The green and red bounding boxes are the user's and ground truth bounding boxes respectively, while the blue contours are the object's contour from the edge detection approach.



Figure 6: 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 the edge detection approach.

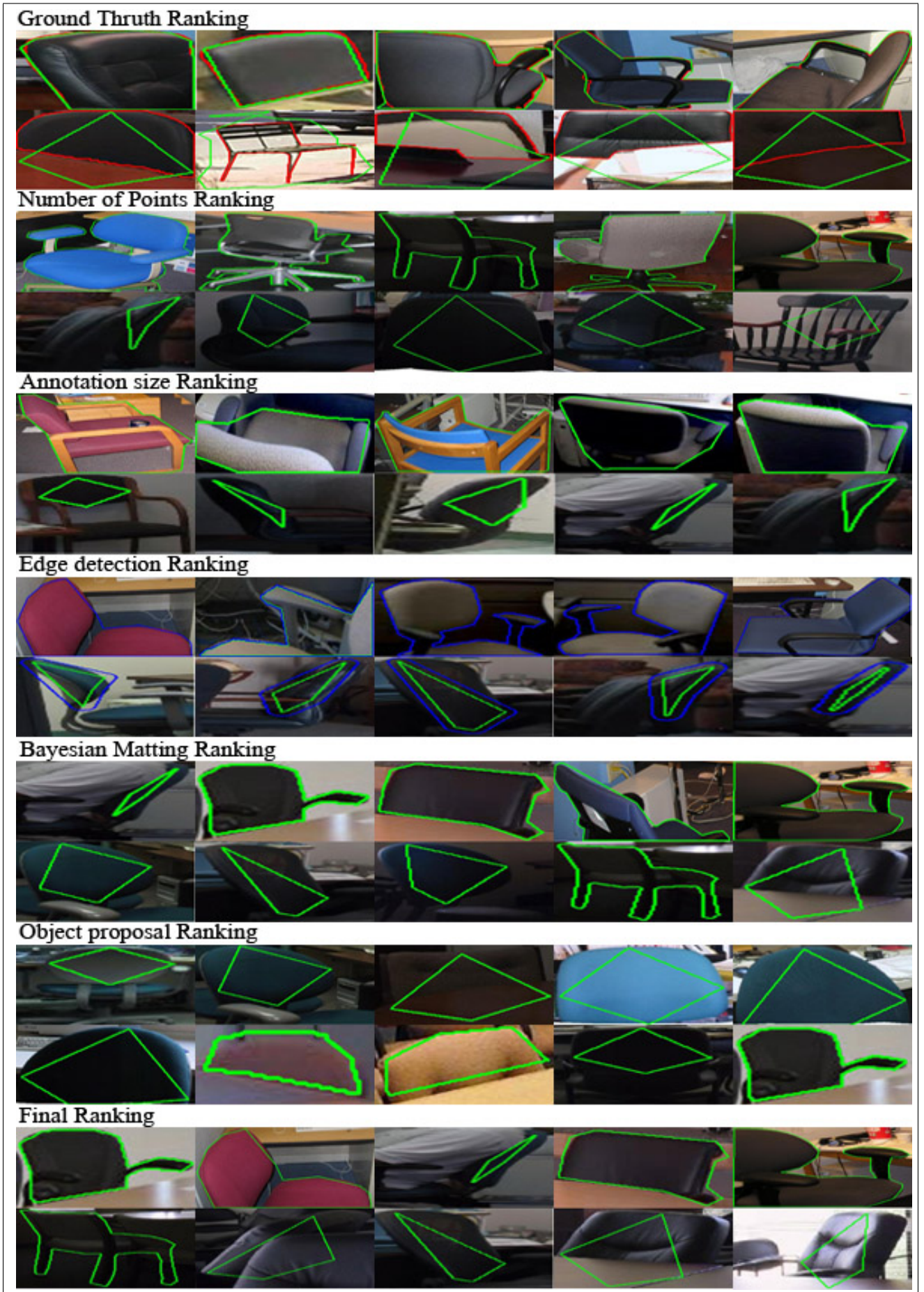


Figure 7: The five highest and lowest ranked user annotations according to the ground truth ranking and each annotation scoring function. The green and red bounding boxes are the user's and ground truth bounding boxes respectively, while the blue contours are the object's contour from the edge detection approach.

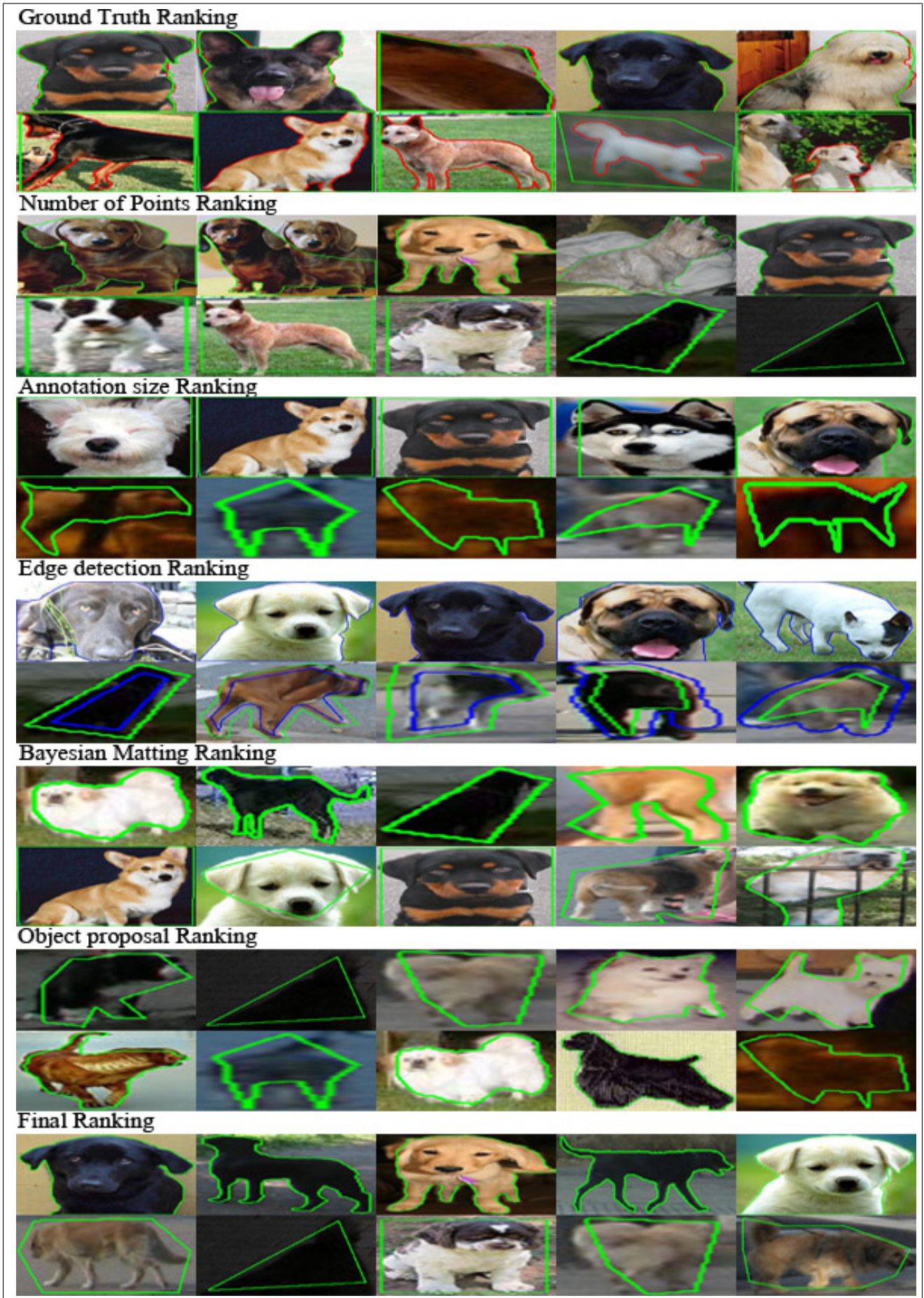


Figure 8: The five highest and lowest ranked user annotations according to the ground truth ranking and each annotation scoring function. The green and red bounding boxes are the user's and ground truth bounding boxes respectively, while the blue contours are the object's contour from the edge detection approach.

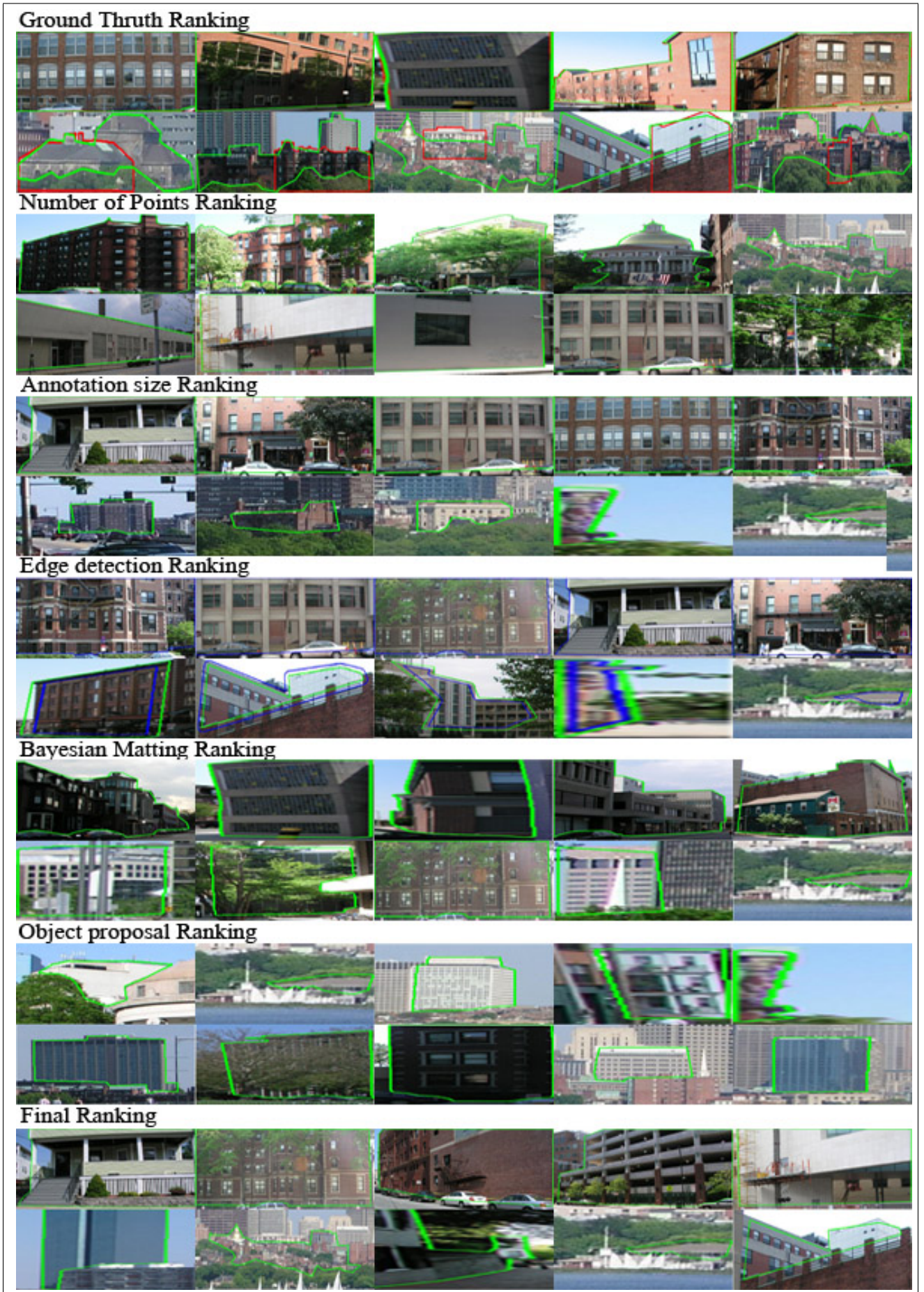


Figure 9: The five highest and lowest ranked user annotations according to the ground truth ranking and each annotation scoring function. The green and red bounding boxes are the user's and ground truth bounding boxes respectively, while the blue contours are the object's contour from the edge detection approach.

- [7] I. Endres and D. Hoiem. Category independent object proposals. *Computer Vision – ECCV 2010*, 6315:575–588, 2010.
- [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.
- [9] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, June 1938.
- [10] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5:32, 1957.
- [11] J. Myers. *Research Design and Statistical Analysis*. Lawrence Erlbaum Associates, Hillsdale, 2003.