To appear in the CVPR Workshop on Biometrics, June 2016

Grouper: Optimizing Crowdsourced Face Annotations*

Jocelyn C. Adams
Noblis
jocelyn.adams@noblis.org

Kristen C. Allen
Noblis
kristen.allen@noblis.org

Tim Miller
Noblis
timothy.miller@noblis.org

Nathan D. Kalka
Noblis
nathan.kalka@noblis.org

Anil K. Jain
Michigan State University
jain@cse.msu.edu

Abstract

This study focuses on the problem of extracting consistent and accurate face bounding box annotations from crowdsourced workers. Aiming to provide benchmark datasets for facial recognition training and testing, we create a 'gold standard' set against which consolidated face bounding box annotations can be evaluated. An evaluation methodology based on scores for several features of bounding box annotations is presented and is shown to predict consolidation performance using information gathered from crowdsourced annotations. Based on this foundation, we present "Grouper," a method leveraging density-based clustering to consolidate annotations by crowd workers. We demonstrate that the proposed consolidation scheme, which should be extensible to any number of region annotation consolidations, improves upon metadata released with the IARPA Janus Benchmark-A. Finally, we compare FR performance using the originally provided IJB-A annotations and Grouper and determine that similarity to the gold standard as measured by our evaluation metric does predict recognition performance.

1. Introduction

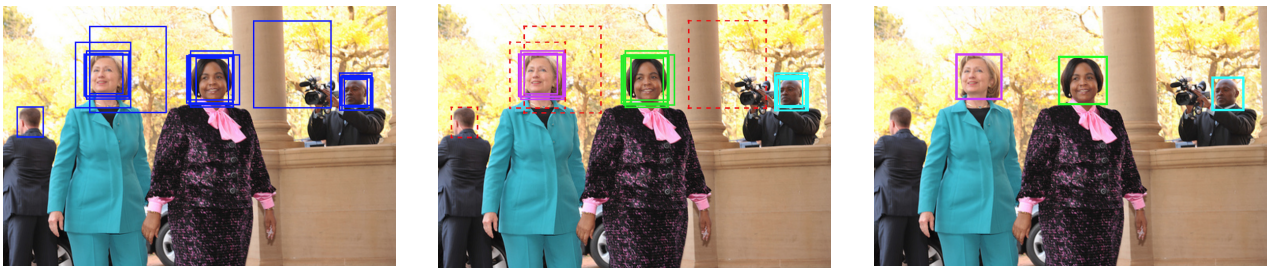
Advances in computer vision and facial recognition have led to dramatic performance improvements, boosted by availability of large-scale data sets from social media and other web scraping, along with the widespread implementation of deep learning methods that make best use of such imagery. In an increasingly saturated market, an algorithm's success has become more dependent on access to large-scale annotated databases. Crowdsourced work is frequently leveraged as a means to annotate large quantities of imagery scraped from the web [3][12][13][15][17]. This study is one of few to objectively verify consolidated face annotations from crowdsourcing against an expert-annotated dataset, which for the remainder of this paper we call the *gold standard*. The ultimate aim of this work is to facilitate consistently annotated datasets for facial recognition (FR) algorithm development.

Because crowdsourced workers have a potential for malicious or careless behavior, lack of understanding of instructions, and general inconsistency, crowdsourced annotations require redundancy and adjudication. Historically, consolidations of facial bounding box annotations have been verified by manual inspection and observations about worker annotation patterns; the original source for this data also estimated *consolidation accuracy* by the variance between different annotations [13]. Here, we define *consistency* by evaluating instead against an independent gold standard and develop an algorithm that creates consolidations most similar to that standard. By creating the gold standard, our methodology enables the objective evaluation of consolidation methods and a more consistent way to evaluate annotations and annotators. From the resulting findings, we lay out several criteria for crowdsourcing face annotations to maximize accuracy at a reasonable cost. The consolidation process and evaluation metric presented here can easily be extended to novel face datasets and image annotation applications.

2. Prior work

Numerous previous studies have analyzed the accuracy, cost, and efficiency of crowdsourced annotations. This study leverages knowledge gained from several prior works, described below, while pursuing the gold standard method proposed in [13].

*This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via FBI Contract # GS10F0189T-DJF151200G0005824. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government.



(a) Raw bounding box annotations. (b) Clustered bounding box groups. (c) Bounding boxes after consolidation.

Figure 1: The Grouper consolidation process. The raw bounding boxes in (a) are clustered by overlap into the three groups shown in (b). Dashed red lines indicate outliers. Those groups are then averaged to produce the final consolidation in (c).

Yuen et al. survey several tactics for improving quality and lowering costs, including specialized algorithms to compile crowdsourced results, questions to screen only qualified workers to complete a certain type of human intelligence task (HIT), pay incentives to workers who performed well in HITs, and filtering out workers who are believed to be cheaters or spammers [17]. Although the methods are not compared against each other, the study's results indicate that each of these techniques can improve the quality of HIT responses. Ipeirotis et al. [5] build on the work of Dawid and Skene [1] in identifying error-prone and biased workers and demonstrate that, on average, accuracy for a crowdsourced task begins to saturate at five labels per object. In addition, Snow et al. found that it took less than 10, and in some cases only 4, non-expert annotations to achieve the same quality as an annotation created by an expert [11].

As an alternative to redundancy, Dawid and Skene propose a system of estimating worker quality based on combining worker annotations. Other algorithmic quality control methods which verify the worker, instead of verifying the work, are shown by [15], [14], and [9] to be effective at increasing accuracy. Raykar and Yu use an empirical Bayesian algorithm to assign scores to annotators by determining the probability that their annotation was made randomly [9]. Their algorithm simultaneously removes spam annotations and consolidates the rest. In Vondrick et al.'s approach, AMT workers are handpicked based on performance and reliability metrics gathered about each worker [15]. Snow et al. have evaluated the use of gold standard annotations to assist in consolidation of categorical annotations on images [11]. All of these studies focus on simple tasks such as binary labeling which are less complicated to compare and consolidate than bounding box annotations.

For more granular tasks, another way to reduce annotation spam that has been explored in the literature is to require training or a qualification test for workers [12][8]. Both of these studies attempt steps to modify unsatisfactory annotations, with varying results. In [8], a freeform language generation task does not see any improvements from worker edits to annotations. Su et al. claim 98% accuracy in a task where a single worker draws a box and others verify its accuracy, as well as cost savings from consensus approaches [12].

The PASCAL Visual Object Classes (VOC) Challenge [3] sets the current precedent for evaluating bounding box annotations. Workers provided bounding boxes for particular objects in images. Then the authors used the overlap between these bounding boxes and ground truth boxes to determine true/false positives in the workers' annotations. While our paper presents a similar paradigm of comparing worker annotations to a ground truth, here termed the "gold standard," our work uses a more granular and comprehensive evaluation metric than does [3].

3. Methodology

We acquired the original Amazon Mechanical Turk annotations that were consolidated into inferred truth on a 477-image subset of IJB-A, along with the consolidations themselves [13]. Additionally, we created a new set of annotations termed the 'gold standard' by manually boxing all faces found in each of the images along a tightly-defined set of guidelines. Building from the consistency in this gold standard set, we define a new evaluation metric to describe the attributes of successful and unsuccessful annotations; this metric considers box size, shape, and location as well as false positives and false negatives. Comparing against this gold standard with our evaluation metric, we investigate the best methods to consolidate disparate user annotations into a single, accurate bounding box for each face in the image.

3.1. Annotation collection

Images in the IJB-A dataset were selected manually from Google Image search results on a predetermined set of 500 subjects, then annotated by Amazon Mechanical Turk (AMT) workers [6]. A number of annotations were collected on each image in order to create a set with all faces boxed; boxes containing the 500 subjects were labeled as

Figure 2: One of the image annotations included in our gold standard.

such, and three facial landmarks (eyes, nose base) were labeled in those boxes. Around five annotations of each type were performed per subject sighting. For the purposes of this paper we do not consider the facial landmark annotations, only the face bounding box. In addition to the publicly available consolidated bounding box annotations in the IJB-A dataset, we obtained the original annotations in order to evaluate various consolidation strategies.

3.2. Annotating a gold standard

In order to compare against representative "ground truth," we first randomly selected 500 images from the dataset. Then one of the authors annotated each image with bounding boxes according to pre-defined guidelines listed below. The annotations were reviewed by another of the authors and 23 image annotations were removed due to the potential for inconsistency with the guidelines, leaving a set of 477 images. The original instructions given to the AMT annotators were outlined in [13], and were based on the understanding that facial recognition (FR) algorithms perform best when boxes are consistent. One factor leading to inconsistency in annotations is varying sizes of boxes around the same face; Taborsky et al. found the most efficient way to keep box size consistent was to advise workers to annotate with boxes that align with the edges of the subject's head as closely as possible [13].

During the gold standard annotation, we followed the same instructions as provided to AMT workers and devised a handful of internal guidelines to deal with situations that the original instructions did not cover (for example, if only a single facial feature is clearly visible because the rest is covered, do not box that face). Figure 2 shows an example of an image annotation included in the gold standard.

3.3. Consolidating bounding boxes

The consolidation process in the proposed Grouper method consists of two main steps: (i) associating bounding boxes into groups that likely refer to the same face, and (ii) averaging the bounding boxes within each group. See Figure 1. After the initial consolidation, we filter out annotations by aberrant annotators and reconsolidate the results.

Associating bounding boxes. The simplest method for associating bounding boxes into groups, each representing a face, is to aggregate the bounding boxes into groups based on overlap. At each step, a new box is compared to already inferred groups of boxes. If no groups exist yet or the box does not overlap sufficiently with any of the groups, it forms a new group. A threshold parameter θ defines the minimum average pixel overlap that the box in question must have with a group of boxes in order to be added to that group. Once all boxes have been considered, any group with fewer users than some specified threshold is removed and the boxes in that group are considered outliers. The rest of the groups are passed along to the next stage of consolidation. A similar aggregative method was used to create the consolidated bounding box annotations included with IJB-A dataset [6]; see Taborsky et al. [13].

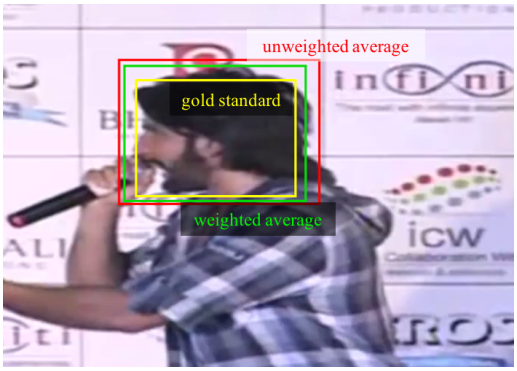
The aggregative method is simple but greedy. Considering pairs of bounding boxes individually ignores information about annotation density that can be useful for associating bounding boxes. For example, if a relatively tight box and a relatively loose bounding box around the same face are compared early in the process, they may be put into different groups even if many bounding boxes exist that bridge the gap between the original two boxes being compared. One solution to this problem is to use a density-based clustering approach to associate boxes on the same face.

The DBSCAN algorithm, first introduced in [2], was developed to cluster spatial databases and is thus designed to perform well on location data. In particular, unlike clustering methods such as k-means, DBSCAN does not explicitly require knowledge of the number of clusters. Instead, the number of clusters is determined by an algorithmic parameter (threshold) while outliers are identified based on relative density. Grouper runs a Python-based implementation of DBSCAN [7] on a similarity matrix representing the percentage of pixel overlap between each pair of boxes.

Averaging bounding box groups. Once the bounding boxes have been sorted into groups, each group must be condensed into a single box, creating what is essentially an average bounding box. Consider that each bounding box is defined by the points of its top right and bottom left corner, (x_1, y_1) and (x_2, y_2). By definition, $x_1 < x_2$ and $y_1 < y_2$. The simplest method for averaging a set of bounding boxes is to average each of these four coordinates and use the results as coordinates for a new bounding box. This is the method that was used to produce the consolidated bounding box annotations included with IJB-A [6].

In an attempt to mitigate the effect of imprecise annotators drawing bounding boxes too loosely, Grouper implements a weighted average method which gives preference

(a)



(b)

Figure 3: An example of unweighted average bounding box and weighted average bounding box as compared to the gold standard. Original annotations can be seen in (a).

to tighter bounding boxes. Let bbs be a group of bounding boxes to be averaged. Let b_a be the number of pixels in box b in bbs. Let b_{x1} be the x_1 coordinate of box b in bbs, and so on. The average x_1 coordinate is calculated using the weighted average defined in Equation 1. Each coordinate's value is divided by ${b_a}^2$, so that the larger the box's area, the less influence the coordinate has on the average. This equation may be generalized for the other three coordinates.

$$\frac{\sum_{bbs} \frac{b_{x1}}{{b_a}^2}}{\sum_{bbs} \frac{1}{{b_a}^2}} \tag{1}$$

See Figure 3 for an illustration of the effect of using a weighted average. Note that the weighted average box is tighter and closer to the gold standard box. While the weighted average implemented in Grouper is effective in producing tighter and thus more precise bounding boxes, it would not be appropriate for all use cases. We present our weighted averaging strategy as an example of how our specific evaluation metrics allowed us to identify and ameliorate a problematic pattern in this set of bounding box annotations.

Reconsolidation. In this step, the averaged bounding boxes are considered a de facto ground truth in the absence of a gold standard. Each worker's complete annotation for an image is evaluated against the consolidation and the worker receives a similarity score based on bounding box overlap. If a worker strays from the norm on the image as a whole, we exclude that worker's annotations from consideration for the image. Once the aberrant workers' annotations have been removed, the consolidation process is repeated on the remaining bounding boxes. Both Grouper and the consolidation strategy used to create the metadata included with IJB-A employ a reconsolidation step [6].

3.4. Evaluating annotations and consolidations

The evaluation metric compares two sets of bounding boxes for a given image: the *ground truth*, most often the gold standard annotation, and the *candidate*, most often a consolidation. In some cases, the consolidation is the ground truth and/or an individual worker's annotation is the candidate. To evaluate a candidate box for an individual image, the overlap scores between each possible pair of boxes from the ground truth annotation and the candidate annotation are collected in a score matrix. The optimal pairing of bounding boxes that maximizes total overlap between the two sets is extracted from this matrix. Any ground truth boxes that are not paired off are considered false negatives and any unpaired candidate bounding boxes are likewise considered false positives. Once boxes are matched between the two sets of annotations for comparison, five different metrics are extracted and an overall score is created by averaging the five scores; see Figure 4 for examples of the first three.

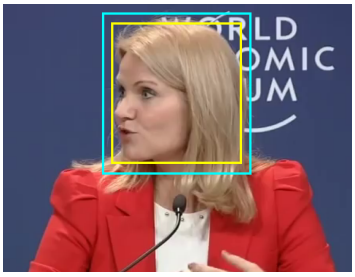
Percent overlap is a prerequisite for several of these scores; the method here differs from typical approaches [3] in that it computes the percentage only with respect to the larger box's area. If overlap is less than θ, the boxes are deemed too dissimilar, and the size, shape, and location scores are 0. In our system, $\theta = 0.5$.

Overlap: Let A_{ij} be the total pixel area of overlap between boxes b_i and b_j. Let a_i be the pixel area of whichever box is smaller, and a_j be the pixel area of the other box. The overlap score for boxes b_i and b_j is A_{ij} divided by a_j.

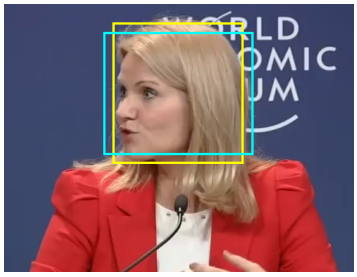
Size: Assuming overlap is greater than or equal to θ, the boxes are deemed too dissimilar, and letting a_i be the pixel area of whichever box is smaller, and a_j be the pixel area of the other box. Then the size score for these boxes is

$$1 - \frac{(1 - \frac{a_i}{a_j})}{(1 - \theta)}. \tag{2}$$

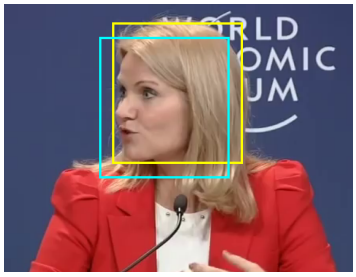
Shape: Let r_i be the ratio of width to height of whichever box is narrower, and r_j be the ratio of width to height of the other box. Necessarily, $r_i \leq r_j$. Because the overlap score of b_i and b_j exceeds θ, $\frac{r_i}{r_j} \geq \theta^2$, the shape score for boxes



(a) size: 0.65, shape: 1.0, position: 1.0



(b) size: 1.0, shape: 0.7, position: 1.0



(c) size: 1.0, shape: 1.0, position: 0.7

Figure 4: Yellow rectangles indicate the gold standard annotation and blue rectangles illustrate an example candidate annotation which would be scored as marked.

b_i and b_j is

$$1 - \frac{(1 - \frac{r_i}{r_j})}{(1 - \theta^2)}. \quad (3)$$

Position: Let X_{ij} be the horizontal distance between the centers of bounding boxes b_i and b_j. Let Y_{ij} be the vertical distance between the centers of bounding boxes b_i and b_j. Let W be equal to the greater of the widths of b_i and b_j and H be equal to the greater of their heights. Because the overlap score of b_i and b_j exceeds θ, $\frac{X_{ij}}{W} \geq \theta$ and $\frac{Y_{ij}}{H} \geq \theta$. The position or location score for boxes b_i and b_j is

$$1 - avg(\frac{\frac{X_{ij}}{W}}{1 - \theta}, \frac{\frac{Y_{ij}}{H}}{1 - \theta}). \quad (4)$$

False negatives: Defined as 1 minus the ratio of (number of ground truth boxes missed by candidate annotation) over (total number of boxes in the ground truth).
False positives: Defined as 1 minus the ratio of (number of boxes in candidate annotation that are not in ground truth) over (total number of boxes in candidate annotation).
Overall score: The size, shape, vertical position, and horizontal position sub-scores fall between 0 and 1. The overall score for the image is the average of its false negative score, false positive score, mean size score, mean shape score, and mean position score. Candidate annotations that are less similar to the ground truth receive lower scores, while annotations identical to the ground truth receive a score of 1.

4. Results and discussion

We will demonstrate the advantages of Grouper using a number of different experiments. First, we will employ the evaluation metric described in Section 3.4 to compare Grouper and other consolidation methods to the gold standard. This will measure the Grouper consolidation's adherence to the initial face annotation guidelines. Correlations between various factors and consolidation performance will be explored as well, providing evidence of the evaluation metric's potential to reduce the need for annotation redundancy. Finally, in order to evaluate the quality of the metadata produced by Grouper with respect to its ultimate use case, we will describe a methodology for comparing facial recognition performance on different metadata sets and present results for these comparisons.

4.1. Consolidation evaluations

Table 1 shows the breakdown in scores for four different consolidation attempts as evaluated against our gold standard. In addition to the IJB-A consolidation and Grouper, we tested a variation of Grouper which did not weight bounding boxes by size during the box averaging step and a variation that used aggregative bounding box association as opposed to clustering. Of all of the strategies tested, Grouper received the highest overall score.

The difference in overall score against the gold standard between the IJB-A consolidation and Grouper, which combines clustering, reconsolidation, and a weighted average bounding box, is statistically significant with p = 0.0084. Grouper thus represents a significant improvement over the strategy used to produce the initial IJB-A metadata in terms of producing annotations that closely resemble the gold standard.

4.2. Predicting consolidation performance

After evaluating our candidate consolidations in comparison to the gold standard, we examine particular factors that may contribute to the accuracy of consolidations. The goal of this analysis is to identify various attributes that might exist within an image or a consolidation that could predict that consolidation's score against the gold standard. It would be desirable to have a method that could predict consolidation strength without the use of a gold standard.

First, we tested whether an annotation's score against the Grouper consolidation predicted score against the gold standard, and determined that the scores are highly correlated (r = 0.906 with $p < 2.2 \times 10^{-16}$ using the Pearson product-moment correlation). This finding establishes annotation

Strategy	Overall Score	Size	Shape	Position	False Neg.	False Pos.
IJB-A	0.926 ± 0.076	0.817 ± 0.147	0.925 ± 0.081	0.923 ± 0.077	0.975 ± 0.114	0.987 ± 0.092
Grouper	0.937 ± 0.052	0.854 ± 0.114	0.93 ± 0.066	0.924 ± 0.068	0.991 ± 0.063	0.984 ± 0.077
Unweighted Var.	0.934 ± 0.041	0.83 ± 0.13	0.934 ± 0.05	0.931 ± 0.045	0.992 ± 0.049	0.986 ± 0.075
Aggregative Var.	0.932 ± 0.08	0.85 ± 0.131	0.924 ± 0.1	0.917 ± 0.102	0.984 ± 0.105	0.985 ± 0.08

Table 1: Overall scores and score components for each consolidation strategy considered, as compared to the gold standard.

score against consolidation as a sound predictor of true annotation quality.

For each image, we determined *annotator concurrence* by calculating the average amount that each annotator differs from the combined consolidation. Let S_i be the overall score of box b_i against consolidation as described in section 3.4, and n be the number of annotations on the image. Then the annotator concurrence measure is

$$\frac{\sum_{i=1}^{n} S_i}{n}. \tag{5}$$

Leveraging equation 5, we then compare the concurrence score on a particular image to that consolidation's score against the gold standard. Intuitively, we would expect a consolidation with higher concurrence to perform better when compared to the gold standard. If a high-concurrence consolidation performs poorly, that would mean multiple annotators made the same error. While annotators may have the same misunderstandings which result in similar errors, such as drawing bounding boxes too loose or drawing boxes around the back of a person's head, we still expect workers to agree on correct annotations more often. Testing correlation between annotator concurrence and the consolidation's score against the gold standard, we find a moderate correlation, with r = 0.477 and $p < 2.2\times10^{-16}$. This result illustrates that some but not all prediction accuracy is maintained when annotations are consolidated.

We also hypothesized that the average size of bounding boxes in a consolidation might predict score: larger boxes should indicate larger faces, less likely to be missed by annotators and with more easily identified boundaries. A slight correlation does exist (r = 0.222 with p = 1.693 $\times10^{-6}$), with larger bounding boxes predicting higher consolidation scores. It is likely that some larger size averages merely come from loosely-drawn bounding boxes, which would score poorly against the gold standard; we conclude that the correlation between average box size and consolidation score is not stronger because we cannot differentiate these cases from images with genuinely larger faces based on raw annotations alone.

Further tests focused on number of bounding boxes per image in the Grouper consolidation. This variable has a strong negative correlation with overall consolidation score (r = -0.458, $p < 2.2 \times10^{-16}$) and a moderate negative correlation with the specific false negative score (r = -0.200, p = 1.739 $\times10^{-5}$). The latter result is somewhat intuitive: the more faces are in an image, the more opportunities the annotator has to skip a face and receive a lower false negative score. In the same vein, an annotator who encounters an image with many faces may also spend less time and effort per bounding box as they would on an image with only one or two faces, in order to complete the HIT as quickly as possible.

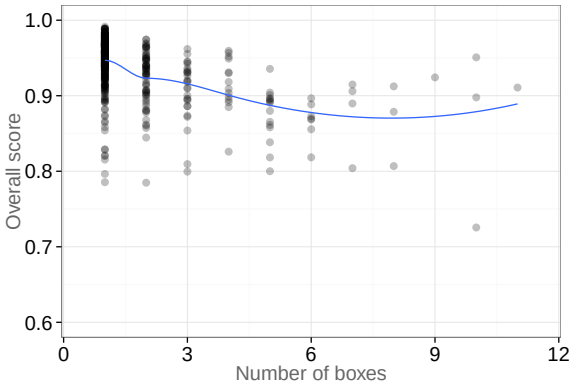


Figure 5: Number of boxes in consolidation against overall consolidation score.

4.3. Face recognition experiments

To justify the appropriateness of our gold standard bounding box guidelines and the validity of our bounding box similarity metric, we designed experiments to test how performance with a state-of-the-art face recognition algorithm compares using input generated from various consolidation strategies. We began by identifying all mutual face locations, defined as a group of bounding boxes (one from each metadata set being tested) which overlap with each other at least 60%. Only face locations that correspond to a subject in IJB-A are included and any unmated samples are removed. We then enrolled the imagery using a deep learning approach based on implementations of methods in [16] and [10]. This approach scores in the same range as the top-10 results on the LFW leaderboard [4]. When the templates are compared against each other, slight flaws caused by misaligned or overly loose bounding boxes may compound and

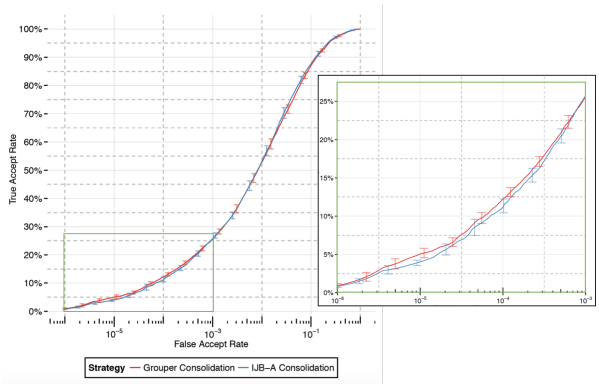


Figure 6: Face recognition performance on the IJB-A consolidations and Grouper.

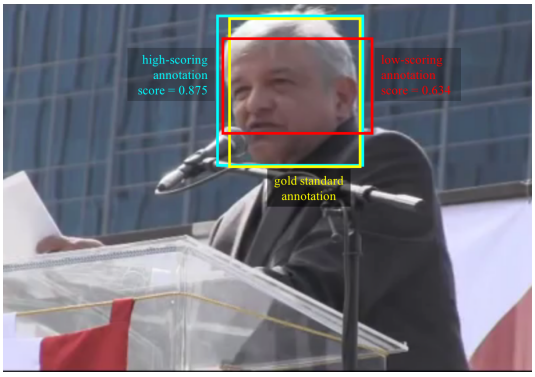


Figure 7: Example of an annotation that received a high score and one that received a low score for the same image.

increase errors.

Note that because the experiments described here only consider mutual face locations, any results are independent of the false positive and false negative rates of the various consolidation algorithms. Therefore, these results should not be the only consideration when evaluating consolidation success.

First, three different metadata sets were considered: the gold standard annotations, the original IJB-A consolidations, and Grouper. The differences in true accept rates among the three strategies were not statistically significant at $N = 370$ (the number of faces represented in all metadata sets). We also performed a larger scale FR experiment to compare Grouper to the consolidation that was included with IJB-A. The partial ROC curve in Figure 6 demonstrates that at operational false accept rates (FARs) of one in a thousand and below, using Grouper consolidations resulted in a significantly higher true accept rate (TAR) by approximately 1%. At higher FARs, the TAR did not differ significantly between the two strategies.

Finally, we identified a set of face annotations for which

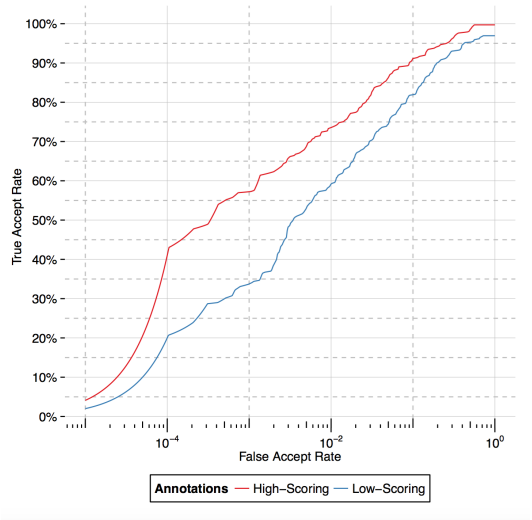


Figure 8: FR performs consistently better on the high-scoring annotations as compared to their low-scoring counterparts.

some annotator's bounding box had scored relatively low when evaluated against the gold standard (0.80 or lower) and some other annotator had scored relatively high (0.86 or higher). Examples of annotations in the two sets are shown in Figure 7. The final set contained 107 face images. Visible in Figure 8, the FR algorithm was significantly more accurate when the high-scoring annotations were used as opposed to the low-scoring annotations, with TARs improving by up to 20%. This indicates that our evaluation metric is relevant for predicting success of FR using bounding box metadata. In addition, this result indicates that the lack of variation in face recognition performance found in the other experiments performed in this paper does not indicate the irrelevance of bounding box quality to FR performance, but rather indicates that the metadata sets tested were of similarly superior quality.

5. Conclusions

This paper illustrated the benefits of specific analysis on bounding box annotations and presented Grouper, a consolidation method that produces better annotations than previously published methods. The clustering approach used by Grouper decreases the percentage of false positives and false negatives among consolidated face annotations, which is particularly critical if the data is to be used to evaluate face detection algorithms. Grouper's weighted averaging strategy reduces variation in bounding box tightness.

Furthermore, the analyses presented here allow the identification of high-performing consolidations. When annotators closely agree on bounding boxes, the consolidated result is closer to the ground truth. Additionally, images

with fewer boxes are more likely to have strong consolidations. Future work could leverage this information to identify consolidations that do and do not require further quality assurance processes, hence increasing overall collection efficiency. The metrics developed here could also be applied to evaluating particular workers' annotations against a gold standard or a suitably validated consolidation, since it has been established elsewhere that annotation quality for an individual worker is relatively stable. Such evaluation could identify particularly successful workers or reject workers who perform poorly, forming the basis of a qualification test to improve the quality of raw annotations before consolidation.

The use of Grouper-produced metadata does result in different FR templates and improved performance at low FARs, but not to an extent that notably impacts scores along the entire ROC. However, FR performance is significantly worse on consolidations that perform poorly against the gold standard, which underscores the need to enforce clear and consistent bounding box guidelines.

There is significant complexity inherent in creating and validating boxed region annotations. Thus, we recommend use of a delineated metric that provides supplementary information about annotation geometry. When monitored, the additional information can be used to demonstrably improve bounding box metadata quality.

References

[1] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, pages 20–28, 1979. 2

[2] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996. 3

[3] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 1, 2, 4

[4] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, 07-49, University of Massachusetts, Amherst, 2007. 6

[5] P. G. Ipeirotis, F. Provost, and J. Wang. Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 64–67, 2010. 2

[6] B. F. Klare, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1931–1939, 2015. 2, 3, 4

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 3

[8] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 139–147, 2010. 2

[9] V. C. Raykar and S. Yu. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *J. Mach. Learn. Res.*, 13(1):491–518, Feb. 2012. 2

[10] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *arXiv preprint arXiv:1503.03832*, 2015. 6

[11] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, 2008. 2

[12] H. Su, J. Deng, and L. Fei-Fei. Crowdsourcing annotations for visual object detection. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012. 1, 2

[13] E. Taborsky, K. Allen, A. Blanton, A. K. Jain, and B. F. Klare. Annotating unconstrained face imagery: A scalable approach. In *IAPR Int. Conference on Biometrics*, volume 4, 2015. 1, 2, 3

[14] L. Tran-Thanh, S. Stein, A. Rogers, and N. R. Jennings. Efficient crowdsourcing of unknown experts using bounded multi-armed bandits. *Artificial Intelligence*, 214:89–111, June 2014. 2

[15] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation - a set of best practices for high quality, economical video labeling. *International Journal of Computer Vision*, 101(1):184–204, 2013. 1, 2

[16] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 6

[17] M.-C. Yuen, I. King, and K.-S. Leung. A survey of crowdsourcing systems. In *2011 IEEE International Conference on Privacy, Security, Risk and Trust*, pages 766–773, Oct 2011. 1, 2