An Empirical Study of Uncertainty in Polygon Annotation and the Impact of Quality Assurance

Eric Zimmermann*, **Justin Szeto***, **Frederic Ratle**

Sama
6795 Rue Marconi, Montréal, Québec H2S 3J9 Canada
{ezimmermann,jszeto,fratle}@samasource.org

Abstract

Polygons are a common annotation format used for quickly annotating objects in instance segmentation tasks. However, many real-world annotation projects request near pixel-perfect labels. While strict pixel guidelines may appear to be the solution to a successful project, practitioners often fail to assess the feasibility of the work requested, and overlook common factors that may challenge the notion of quality. This paper aims to examine and quantify the inherent uncertainty for polygon annotations and the role that quality assurance plays in minimizing its effect. To this end, we conduct an analysis on multi-rater polygon annotations for several objects from the MS-COCO dataset. The results demonstrate that the reliability of a polygon annotation is dependent on a reviewing procedure, as well as the scene and shape complexity.

1 Introduction

Over the past decade, the emergence of deep learning in computer vision has enabled applications to interpret and understand visual information with increasing accuracy. At the heart of this progress lies the crucial role of annotations in training, validating, and fine-tuning machine learning (ML) models through supervised learning. Annotations provide the necessary context and labelled data that empower algorithms to recognize and extract meaningful insights from images and videos.

In order to acquire annotations, human annotators meticulously label various visual elements and attributes of a scene. These annotations can be provided as class labels, semantic descriptors, bounding boxes, or dense contours described by masks or polygons. These annotations serve as ground truth references and are fundamental to the development and deployment of reliable ML systems.

A major challenge in computer vision annotations is the complexity and diversity of visual data. Images may contain a wide variety of dense and occluded objects at various resolutions and lighting conditions. These factors make it difficult to discern objects and lead to the notion of uncertainty, since it may not be possible to assign certain types of labels to ambiguous segments in an image (see example in Figure 1). Annotation quality relates to model quality, resulting in a need for perfection despite the task's overhead. The higher the quality of an annotation, the more challenging it may be to acquire the annotation.

Figure 1: Variability in annotating an object. The car tires are hidden by the car's shadow, thereby leading to large annotation uncertainty around the tires.

In a production setting, each annotation task is performed in accordance with instructional guidelines, and an emphasis is placed on fast completion time. In these workflows, a range of annotators with mixed levels of experience are required to complete a task with low or zero error tolerances using the polygon format. There is an expectation that quality can be met with guaranteed perfect pixel precision. However, these expectations fail to account for the constraints of the task, the ambiguity of the instructions, and the complexity of the scene. If a workflow is subject to a quality assurance (QA) review, these factors may result in additional reworks and wasted resources, which pose a risk to the project as a whole. It is possible to mitigate risk by better understanding the types of errors that are acceptable or unavoidable. This can be accomplished by synchronizing expectations across all parties involved in a project, by understanding the limits and uncertainties of a task, and by adapting reviews accordingly.

2 Overview

We study the presence of uncertainty and the effects of an additional quality assurance stage in a multi-rater annotation workflow. We summarize a group of polygons using the notion of a consensus shape and leverage it to describe global and local variability for sets of shapes. The choice of consensus model therefore dictates how reliable the downstream analysis may be.

*These authors contributed equally.

Presented at the Works-in-Progress and Demonstrations track, AAAI HCOMP 2023.

Let $s \subset \mathbb{R}^2$ be a closed shape with boundary contour ∂s. Given a set of n shapes $S = \{s_i\}_{i=1}^n$, the asymmetric distance function $d(\partial s_i, \partial s_j)$ measures the total surface distance between a pair of contours. For any point p on a curve ∂s_i, we denote the corresponding point on ∂s_j as $y_{\partial s_j}(p)$. The segment between p and ∂s_j is the geodesic (Charpiat, Faugeras, and Keriven 2004; Boykov et al. 2006; Kervadec et al. 2019). The asymmetric squared distance is defined as:

$$d(\partial s_i, \partial s_j)^2 = \int_{\partial s_i} ||y_{\partial s_j}(p) - p||^2 dp. \quad (1)$$

For any set S of curves, the mean shape $\bar{\mu}_s$ minimizes the asymmetric distance between all curves in S. Given the set $\Omega(\partial s)$ of all curve boundaries, the mean curve $\bar{\mu}_s$ is defined as:

$$\bar{\mu}_s = \underset{\partial \bar{\mu}_s \in \Omega(\partial s)}{\operatorname{argmin}} \sum_{s_i \in S} d(\partial \bar{\mu}_s, \partial s_i)^2. \quad (2)$$

The mean curve is any curve that minimizes this partial differential equation, as per Eq 2. A gradient flow can be used to iteratively construct an optimal curve, where the quality of the curve depends on the chosen initial conditions as well as the complexity and variability of the shapes in S (Charpiat, Faugeras, and Keriven 2004). In practice, we may find an adequate approximation to the mean curve in a discrete setting under a few strict assumptions.

A1: The set of all curves in S all define the same underlying shape with minor variability.

A2: There exists an exact distance transform (EDT) (Elizondo-Leal, Parra-González, and Ramírez-Torres 2013) that approximates the asymmetric distance over a subsection of the curve.

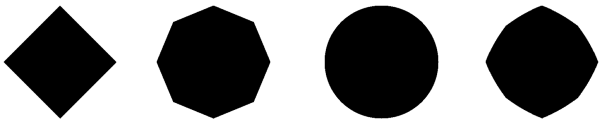
A3: The gradients of the distance function can be inferred from local measurements of the distance transform.

If all assumptions hold, we solve for the mean curve by finding a contour that corresponds to the zero crossing of the Laplacian. The mean curve is computed using marching squares on the mean signed distance map based on the EDT for all contours in S. This representation is useful as it may be used to identify regions of interest with high disagreement between elements in S. Given an EDT map $D_{T_{s_i}}(p)$ at spatial location p, the approximate lower bound to the average absolute difference between the boundaries of the mean contour and the set of shapes is computed as the average accumulation of distance over the mean contour. The expected boundary distance d_B between the mean shape μ with contour length $|\partial\mu|$ and S is therefore defined as:

$$d_B(\mu, S) = \frac{1}{|\partial\mu||S|} \sum_{s_i \in S} \int_{\partial\mu} D_{T_{s_i}}(p) dp. \quad (3)$$

We note that the distance map approximates geodesics asymmetrically and is robust to large spikes in curvature. This distance therefore plays a role in the underestimation of the true distance measure.

It is also possible to define the mode shape $\tilde{\mu}_s$ over the set of shapes S. The mode is defined at a point using a majority vote consensus over $\mathbb{R}^2$. The mode is not a particularly reliable consensus, as it may become jagged and discontinuous. It is also not clear how to then find corresponding distances between the mode curve and each curve that composes it. Moreover, it fails to capture distinct disagreement on subsections of a curve since it tends to converge to the majority trend. There are also other methods to compute consensus shapes that incorporate other priors and assumptions, such as Expectation-Maximization-based methods (Lee 2018; Warfield, Zou, and Wells 2004) and Reliability Aware Sequence Aggregation (Li and Fukumoto 2019).



(a) Diamond (b) Octagon (c) Circle (d) Consensus

Figure 2: Example of a mean consensus polygon generated from a fair aggregation of a diamond, octagon, and circle.

3 Dataset

A set of images is acquired by manually selecting samples from MS-COCO (Lin et al. 2014). Samples are selected based on a diversity criterion to ensure a mix of common classes composed of humans, cars, animals, and other miscellaneous common objects. A dataset sample is generated from a crop of an image based on a bounding box of interest. The crop is performed with a 10% margin around the bounding box. Samples are selected to have no holes and minimal occluders in order to avoid inconsistencies in annotations due to scene ambiguities. In total, the dataset is composed of 24 samples that are each paired with a unique instruction set that states which areas should or should not be annotated. The instruction set is reviewed by a team of quality assurance specialists to ensure that instructions are consistent with standard project workflows. The annotators are tasked with labeling the contour of the object in question using a polygon tool, with zero pixel error tolerance. Annotators are allowed to perform zoom operations, but image enhancement tools are restricted. Once completed, each annotator receives feedback on their tasks by a dedicated quality assurance specialist and are tasked with repeating the process. Quality assurance is accomplished via a calibration session between the specialist and the annotators to establish a common understanding of the annotation precision needed and semantic errors to avoid. Note that the QA may bias the result towards a specific interpretation of the image and instructions, which is not always the same as "the truth".

4 Methodology and Results

We analyze uncertainties aggregated across the dataset before and after the QA step and follow up with an analysis based on local curve segments with high variability. Analysis is done using the mean curve and an arc length parameterization is used to observe patterns on its subsections.

4.1 Aggregated Uncertainties

The mean shape is computed and used to calculate the mean distance (measured in pixels) for each sample in the dataset.

The process is repeated before and after a QA review. Average distances to the consensus are found to be 0.522 ± 0.650 and 0.407 ± 0.401 respectively. A Welch's t-test at a 95% confidence interval detects a statistically significant shift in the distribution of errors measured in pixels. The distributions of distances are presented in Figure 3.

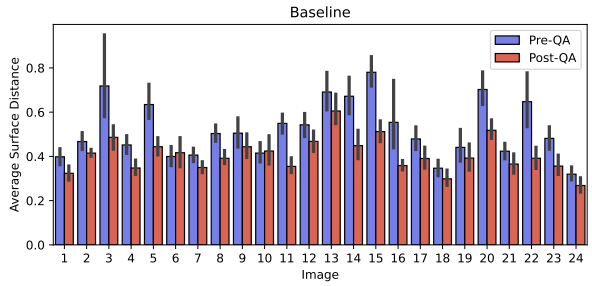


Figure 3: Per-image mean distance and variance, and differences between pre- and post-QA. Error is dramatically reduced after a QA intervention.

4.2 Local Contour Uncertainties

Since important regions of interest cannot be flagged using aggregated statistics along the entirety of a curve, we quantify local variations between the curves and the consensus. Local analysis is performed over subsections of the curve using an arc length parameterization of the consensus for each image. The accumulation of the signed asymmetric distance over a subsection of the consensus is used to represent the standard deviation of a family of curves relative to its midpoint. A rejection threshold is used to find regions on the curve whose deviation exceeds this value. Results are visually verifiable as per Figure 4 and 5 for a cutoff threshold of 0.5.

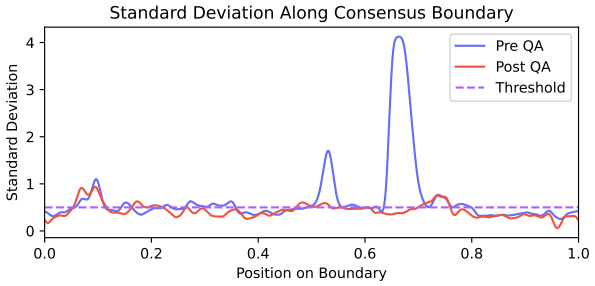
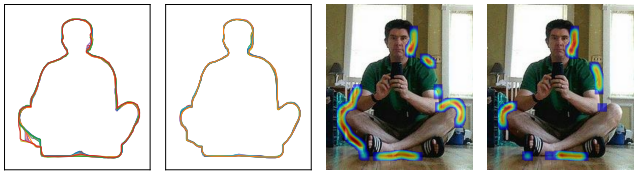


Figure 4: Contour variance over the arc length of the mean curve pre- and post-QA. The cutoff threshold is used to detect anomalous segments.

Figures 5 demonstrates the impact of a QA step and how it leads to a correction. When observing the legs of the individual, it is unclear that there may be pants in the region where contrast is low. This is corrected by an executive decision based on style guidelines provided by a reviewer. On the other hand, regardless of a review, the variance along



(a) Pre-QA (b) Post-QA (c) Pre-QA (d) Post-QA

Figure 5: A set of polygons acquired pre- and post-QA. Heatmaps provide visual cues in regions with high local variance. Regardless of review, regions of low contrast and high curvature remain uncertain.

the head of the individual and below the legs could not be avoided. It is observed that these regions had poor contrast and higher relative curvature. It is noted that even though the post QA polygon looks reasonable, a rework may still be requested in a zero pixel tolerance setting due to minor disparities in the interpretations of the result.

4.3 Discussion

Our analysis demonstrates that the QA process significantly enhanced style consistency and is a valuable step in the annotation process. Without an expert in the loop, stylistic variation can have a significant impact on the success of a project. The QA strategy is of utmost importance and cannot be understated. Rejections and reworks due to errors are the single largest factors limiting the success of a project. Small unavoidable issues may lead to flagging that greatly increase the total time spent working on a task and as a result, inflate the total cost of the venture. These costly interventions are associated with local variations. By understanding which errors can be tackled and which are inherent to an image, it is possible to reduce time, effort, and resources spent.

Due to the limited data and cohort size, it is not possible to provide a statistical breakdown of what confounders have effects on local uncertainties. However, intuitively, there is a relationship between curvature, contrast, and uncertainty. Regions of low contrast and high curvature may lead to more uncertainty. An annotator in a known domain has an unspecifiable shape prior and additional context provided by the entire scene that cannot be estimated. These factors allow them to perform well in regions that would otherwise be difficult without this contextual information.

5 Conclusion

We demonstrate that a quality assurance review phase plays an important role in ensuring consistency throughout an annotation project, and that it is possible to assess the quality of a set of annotations using a mean consensus. Furthermore, we provide empirical results illustrating that the average spread of contours typically spans at least one pixel on a global level and often vastly exceeds acceptable ranges on a local level. Regions with inherently high uncertainty should not be simply reworked as they will lead to additional costs in a project and improvements cannot be guaranteed.

References

Boykov, Y.; Kolmogorov, V.; Cremers, D.; and Delong, A. 2006. An Integral Solution to Surface Evolution PDEs Via Geo-cuts. In Leonardis, A.; Bischof, H.; and Pinz, A., eds., *Computer Vision – ECCV 2006*, 409–422. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-33837-6.

Charpiat, G.; Faugeras, O.; and Keriven, R. 2004. Approximations of Shape Metrics and Application to Shape Warping and Empirical Shape Statistics. *Foundations of Computational Mathematics*, 5(1): 1–58.

Elizondo-Leal, J. C.; Parra-González, E. F.; and Ramírez-Torres, J. G. 2013. The Exact Euclidean Distance Transform: A New Algorithm for Universal Path Planning. *International Journal of Advanced Robotic Systems*, 10(6): 266.

Kervadec, H.; Bouchtiba, J.; Desrosiers, C.; Granger, E.; Dolz, J.; and Ben Ayed, I. 2019. Boundary loss for highly unbalanced segmentation. In Cardoso, M. J.; Feragen, A.; Glocker, B.; Konukoglu, E.; Oguz, I.; Unal, G.; and Vercauteren, T., eds., *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, volume 102 of *Proceedings of Machine Learning Research*, 285–296. PMLR.

Lee, D. J. L. 2018. Quality Evaluation Methods for Crowdsourced Image Segmentation.

Li, J.; and Fukumoto, F. 2019. A Dataset of Crowdsourced Word Sequences: Collections and Answer Aggregation for Ground Truth Creation. In *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*, 24–28. Hong Kong: Association for Computational Linguistics.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, 740–755. Cham: Springer International Publishing. ISBN 978-3-319-10602-1.

Warfield, S. K.; Zou, K. H.; and Wells, W. M. 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7): 903–921.