# Tile-Based Quality Evaluation Methods for Crowdsourcing Image Segmentation

**Authors removed for Anonymity**

**Abstract**

Abstract

## Introduction

- Most segmentation evaluation strategy people use are summarization/heuristic based, require ground truth, we come up with a more formal tile based model which does not require ground truth.
- Summarization based method says this BB is good, that BB is bad, but doesnt allow partial inclusion of a region
- In the limit of small number of workers with noisy output, that means that if no BB is very good, we would only get an okay-ish BB. (c.f. dumbell example)
- Tile based methods give us a way to allow partial inclusion of a region.
- This concept is similar to picking the majority voted region, but majority vote doesnt account for worker qualities
- With a worker error model + tile based scoring function, we can take make use of the best responses from many workers.
- Explain its important to beat :
  - summarization-based methods
  - tile based majority vote
  - Individual worker resposnes
  - Computer vision methods

## Related Works

Despite several large-scale efforts to collect image segmentation from crowds(Lin et al. 2014; Martin et al. 2001; Torralba, Russell, and Yuen 2010; Everingham et al. ), most have relied on summarization-based metrics to quantify their segmentation data quality. Summarization-based scores are functions that measures the quality of a worker's bounding box given the ground truth. Common summarization methods include precision, recall, area ratio or number of control points in the bounding box (Vittayakorn and Hays 2011). While these are fairly discriminative measures of bounding box, the scoring calculation requires the use of ground truth to compare against. Other heuristic approaches that doesn't require ground truth segmentation include quantifying the user types and their click-stream behaviour to determine work quality(Cabezas et al. 2015; Sameki, Gurari, and Betke 2015) and comparing the worker responses to features extracted from computer vision algorithms(Vittayakorn and Hays 2011; Russakovsky, Li, and Fei-Fei 2015).

(Dawid and Skene 1979) is the first work that used the EM algorithm to model an individual's biases and skills in the absence of ground truth data, by using a confusion matrix. (Welinder and Perona 2010) proposes a general model that separately models annotator quality and the biases applied to binary, multivalued and continuous annotations. (Welinder et al. 2010) develops a multidimensional concept of worker qualities and task difficulties by considering object-presence labelling as a noise generation process. The objective truth label is captured by a multidimensional quantity of task-specific measurements and deformed by worker and image related noise, the noisy vector obtained after this process is projected onto the vector of user expertise (which summarizes how well the user percieves each of these measurements), and finally the score is binarized into an inferred label. Many have extended this line of work beyond binary classifications by developing EM-like approaches that works on multiple-choice (Karger, Oh, and Shah 2013) as well as free-form responses (Lin, Mausam, and Weld 2012), but these have not been directly applied to the task of object segmentation.

However, while EM algorithms assign probabilities regarding *how good a worker's bounding box is*, for the task of object segmentation, we are ultimately more interested the end goal of *what is the best bounding box that we can get from these data*. Since these formal probabilistic models treats worker bounding box as the base quantity for modeling worker quality, the best bounding box that one could derive from such algorithm can only be as good as the best worker bounding box in the dataset. Even though the annotation probabilities are sufficient for determining the best binary-labels, image information such as overlapping areas would be useful and not account for in these algorithms. We suspect that this is why many area-based metrics are still more commonly used in practice than EM approaches.

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

## Model

### Tile Graph

Describe Tile based models. Briefly describe construction of tiles from BBs. Clarify that our search space is tile combination formed by all the worker's tiles not the space of all possible coordinates. (i.e. we assume that a region can not be inside the BB if no workers bounded that region)

### Definitions

### Worker Error Model

### Inference

Describe assumptions on pdfs and inference process. (E and M steps). How are parameters in the models determined empirically.

## Experiments

We collected data from Amazon Mechanical Turk where each HIT consisted of one annotation task for a specific prelabelled object in the image. There is a total of 46 objects in 9 images from the MSCOCO dataset (Lin et al. 2014). These tasks represent a variety of image difficulty (based on object clutter-ness), potential logical error and level of ambiguity. For each object, we collected annotations from a total of 40 independent workers.

- Data collection process
- PR curves
- show that our experiment works well on smaller, noisier datasets
- Beat Baselines:
  - summarization-based methods
  - tile based majority vote
  - Individual worker responses

## Conclusion

## References


Cabezas, F.; Carlier, A.; Charvillat, V.; Salvador, A.; and Giro-I-Nieto, X. 2015. Quality control in crowdsourced object segmentation. *Proceedings - International Conference on Image Processing, ICIP* 2015-Decem:4243–4247.

Dawid, A. P., and Skene, A. M. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. 28(1):20–28.

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Karger, D. R.; Oh, S.; and Shah, D. 2013. Efficient crowdsourcing for multi-class labeling. *ACM SIGMETRICS Performance Evaluation Review* 41(1):81.

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. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8693 LNCS(PART 5):740–755.

Lin, C. H.; Mausam; and Weld, D. S. 2012. Crowdsourcing control : Moving beyond multiple choice. *Uai* 491—-500.

Martin, D.; Fowlkes, C.; Tal, D.; and Malik, J. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, 416–423.

Russakovsky, O.; Li, L.-J.; and Fei-Fei, L. 2015. Best of Both Worlds: Human-Machine Collaboration for Object Annotation. 2121–2131.

Sameki, M.; Gurari, D.; and Betke, M. 2015. Characterizing Image Segmentation Behavior of the Crowd. 1–4.

Torralba, A.; Russell, B. C.; and Yuen, J. 2010. LabelMe: Online image annotation and applications. *Proceedings of the IEEE* 98(8):1467–1484.

Vittayakorn, S., and Hays, J. 2011. Quality Assessment for Crowdsourced Object Annotations. *Procedings of the British Machine Vision Conference 2011* 109.1–109.11.

Welinder, P., and Perona, P. 2010. Online crowdsourcing: Rating annotators and obtaining cost-effective labels. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010* 25–32.

Welinder, P.; Branson, S.; Belongie, S.; and Perona, P. 2010. The Multidimensional Wisdom of Crowds. *NIPS (Conference on Neural Information Processing Systems)* 6:1–9.