Establishing Uniform Image Segmentation Ground Truth Protocols for Uncertainty Quantification and Improved Model Evaluation

Maximillian G. Chen, David J. Stracuzzi, Michael C. Darling, Matthew G. Peterson
Sandia National Laboratories
Albuquerque, NM 87123
{mgchen|djstrac|mcdarli|mgpeter}@sandia.gov

1 Introduction and Motivation

Creating datasets of ground truth images and benckmarks, where features are marked by ensembles of human labelers performing manual segmentation in selected image products, gives rise to many challenge problems evaluating the performance of computer vision algorithms. These datasets have proven their fundamental importance in computer vision research by enabling targeted evaluation and comparison of algorithms for image segmentation, feature classification, and anomaly detection tasks. They also allow for algorithm performance to be evaluated under uncertainty, where the margin of error for the model performance is quantified. With a dataset that contains multiple annotations for each image, and multiple images, we can assess how well algorithms answer a number of machine learning questions:

- How well does a model segment an image and detect and classify features of interest?
- How confident are we that the model's results are reasonable?
- If we have multiple images of the same scene, does a particular image help us better understand what is happening in a scene?

Human visual judgment is the acknowledged gold standard for assessing the performance of computer vision algorithms. Unfortunately, most benchmark sets are not capable of supporting uncertainty assessments in algorithmic evaluation, because established practice is to recruit a single image expert to mark boundaries and features in imagery. When multiple experts are called upon to annotate the same image, we see corresponding variability in judgment about morphological boundaries, centroids and feature classes. For example, images in the Berkeley Segmentation Data Set and Benchmarks 500 (BSDS500) [5, 1] and the Caltech-256 Object Category Dataset [4] have multiple different annotations due to a lack of an established ground truth protocol for the human annotators to follow. Properly captured, this variability could provide useful information to algorithm developers interested in establishing performance envelopes for their exploitation techniques, as well as enable more robust algorithmic evaluations to determine model performance.

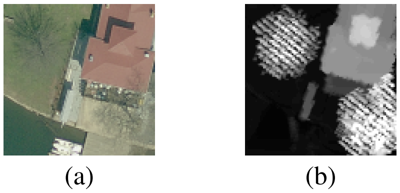
In order to minimize variability among human annotators, a strict ground truth protocol needs to be developed. This protocol must specifically define what features annotators should look for and account for the design and needs of the algorithms that will be tested on the ground truthed images. Protocol development requires interdisciplinary collaboration between individuals with expertise in multiple domain areas spanning algorithm design, uncertainty quantification, and cognitive/human factors. It is a challenging problem due to the need to develop a deep understanding of what defines "truth" in imagery analysis and how different assessments of data relate to each other both qualitatively and quantitatively. The protocol can include semi-automatic approaches that include support from clustering methods to shorten the amount of time it takes to ground-truth images [2, 6, 3].

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Effectiveness of the protocol can be evaluated by the variability in the annotations on the ground truthed images, as well as how quickly the images can be annotated.

2 Application to Multi-Modal Image Analysis

One problem that requires rigorously developed ground truth protocols is *multimodal image analysis*, the analysis of multiple types of images that describe the same scene. Figure 1 contains example images that cover a small region in Philadelphia that contains trees, grass, water, pavement, a building, and a variety of small, undetermined objects. Figure 1a shows the 100x100 pixel optical image of the target area. Each pixel contains red, green, and blue values scaled from 0 to 1. Figure 1b shows the same region imaged with lidar (Light Radar) which has been preprocessed into a height map (lighter colors indicate taller data points).



(a) (b)

Figure 1: 100x100 pixel images of target area captured by optical (a) and lidar (b) sensors.

To address the problems of segmenting the multiple images and determining if each image provides significant information about the scene of interest, the images need to be ground truthed for uncertainty quantification. Because these images often present difficulties such as poor collection conditions (such as clouds and haze), different resolutions and look angles for different sensors, and different collection times, this will increase the variability in human annotation. Therefore, a uniform criteria for drawing the bounding boxes around the objects in a scene will minimize the variability in human annotations and ensure the same objects are being identified by all annotators. After a dataset containing multiple ground truth annotations of the images using the established protocol is created, it can serve as either training data or an evaluation standard for an image segmentation challenge task. There will be a distribution of annotations, which can be used to evaluate the performance of a segmentation model under uncertainty. The ground truthed images can serve as a more informative training dataset towards learning an appropriate segmentation model by providing the data values in the image and the bounded features of interest in the image. After a segmentation model is fitted to the images, the performance of the model can be evaluated against the ground truthed images. The results of the model fit, such as the estimated data values or the estimated parameters of the clustering model, can be compared to the distribution of the ground truthed images. If the results are closer to the median of the distribution of the ground truthed images, then the model appears to segment the images well. However, if the results are in the tails of the distribution of the ground-truthed images, then the validity of the model needs to be questioned.

After quantifying the uncertainty of the segmentation models for the individual images, we need to determine the significance of each image to our understanding of the scene. An additional challenge problem is using the uncertainty results from each image to assess the overall uncertainty of the segmentation and classification of the region. This problem can be challenging, as each image type (optical and lidar) could have different clustering models because the data values for different images have different structures and meanings. (For each pixel, optical data is a three-dimensional vector with red, green, and blue values, while lidar data is a single point with height information.)

3 Conclusions

We argue the importance of establishing uniform ground-truth protocols that produce more consistent ground-truth images among human annotators and how creation of these ground truthed images gives rise to a new class of challenge problems. Enhanced ground truth datasets allow for better algorithmic evaluations in many application areas. We illustrate this need on the multimodal image analysis problem, where strict protocols are required to minimize the variability among annotations and produce better ground-truth datasets suitable for algorithmic evaluation and uncertainty quantification.

Acknowledgments

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. The authors thank Aleksander Bapst, Laura Matzen, and Laura McNamara for helpful comments and discussion.

References

[1] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):898–916, May 2011.

[2] Bastiian J. Boom, Phoenix X. Huang, Jiyin He, and Robert B. Fisher. Supporting ground-truth annotation of image datasets using clustering. In *21st International Conference on Pattern Recognition (ICPR 2012)*, pages 1542–1545, November 2012.

[3] Z. Chen and T. Ellis. Semi-automatic annotation samples for vehicle type classification in urban environments. *IET Intelligent Transport Systems*, 9(3):240–249, 2015.

[4] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[5] D. Martin, C. Fowlkes, D. Tal, and J. Malik. 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, pages 416–423, July 2001.

[6] J. Moehrmann and G. Heidemann. Semi-interactive image annotation with visual feedback. In *ECCV Workshop on Human-Machine Communication for Visual Recognition and Search*, 2014.