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

Maximillian G. Chen, David J. Stracuzzi, Michael C. Darling, and Matthew G. Peterson

Sandia National Laboratories, Albuquerque, NM

Sandia National Laboratories

Main Takeaways

- Datasets of ground truth images and benchmarks: features are marked by ensembles of human labelers performing manual segmentation in selected image products
- 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 for image segmentation, feature classification, and anomaly detection tasks.
- Enhanced ground truth datasets containing **multiple annotations for each image, and multiple images**, allow for better algorithmic evaluations in many application areas.

Example Challenge: Detecting Ships in Satellite Imagery [2]

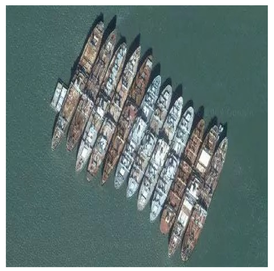


Figure: Google Maps satellite photos of US government's reserve fleet in Suisun Bay, California [4]

Current State of Challenge:

- Training dataset requires labeling of images as "ship" or "no-ship" classification, which is much less effort than ground-truthing images with segmentations of parts of ships.
- This is suitable for the task of detecting ships in the imagery.
- Limitations:
  - Cannot **accurately and precisely detect** the **constituent parts** of the ships, which limits information exploitation (What kind of ships are in the imagery? What are the ships carrying?)
  - Cannot assess the **uncertainty** of the algorithm's performance, where the margin of error for the model performance is quantified
  - Cannot be used for challenges requiring informed and justified decision making (Are the ships a threat to the port?)

Kaggle Challenge Description:

- Kaggle challenge "Ships in Satellite Imagery: Classify ships in San Francisco Bay using Planet satellite imagery" [2] (Due to contest rules, we display an example image of satellite imagery of ships from an alternate source [4].)
- Aim of dataset: address the common task of detecting the presence of large ships in satellite images and automating this process.
- The dataset consists of image chips extracted from satellite imagery collected over the San Francisco Bay area in Planet's Open California dataset [1]. It includes 2800 80×80 RGB images labeled with either a "ship" or "no-ship" classification. Image chips were derived from PlanetScope full-frame visual scene products, which are orthorectified to a 3 meter pixel size.

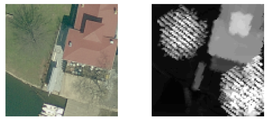
New Challenge Potential With Uniform Ground Truth Dataset:

- Improved image exploitation: detect **constituent parts** of ships, determine the **types of ships** and **what ships are carrying**
- Precision of detection: how **tightly** can an algorithm detect and segment the above features in the imagery?
- Distribution of "acceptable" annotations allows evaluation of uncertainty of model's results: how **confident** are we that not only there are ships in the imagery, but if there are ships, what kind of ships they are and what they are carrying?
- Make decisions regarding whether or not ships are a **legitimate** threat (See Stracuzzi et al [6] poster for more information on data to decisions.)

Another Potential Challenge: Multi-Modal Image Analysis

Multimodal Image Analysis: analysis of multiple types of images that describe the same scene

Example: images that cover a small region in Philadelphia that contains trees, grass, water, pavement, a building, and a variety of small, undetermined objects



(a) (b)

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

- Optical image (a): Each pixel contains red, green, and blue values scaled from 0 to 1.
- Lidar (Light Radar) image (b): Each pixel contains a height map (lighter colors indicate taller data points).

Research Problems:

- Strong need for uniform annotation criteria: these images often present difficulties such as poor collection conditions (such as clouds and haze), different resolutions and look angles for different sensors
- Ground truth dataset can serve as either training data or an evaluation standard for an image segmentation challenge task.
  - 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
  - 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.
- After quantifying the uncertainty of the segmentation models for the individual images, determine the significance of each image to our understanding of the scene.
  - Using the uncertainty results from each image to assess the overall uncertainty of the segmentation and classification of the region
  - Challenging problem: each image type (optical and lidar) could have different clustering models because the data values for different images have different structures and meanings

Berkeley Segmentation Data Set and Benchmarks 500 (BSDS500) Example [5, 3]

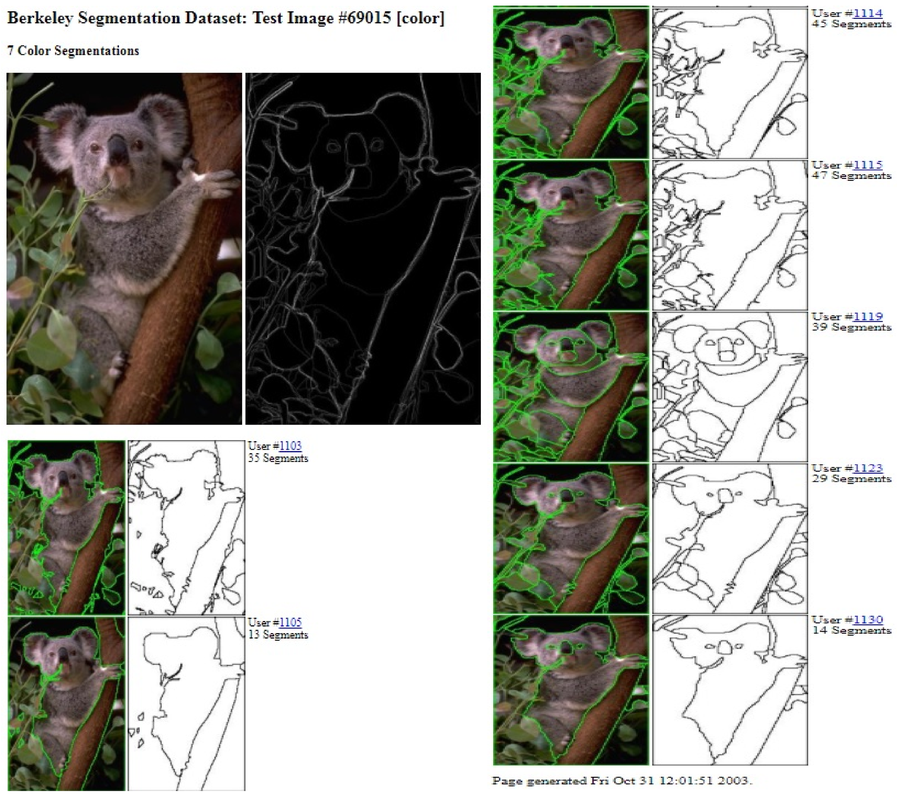


Figure: Seven user segmentations of Berkeley Segmentation Dataset: Test Image 69015 [color]

Analysis of Koala Example:

- With no established ground truth protocol, seven annotators yield seven different annotations.
- Varied segmentation of koala's body parts (ears, eyes, feet, hands, legs, mouth, nose)
- Varied segmentation of branches and eucalyptus leaves

Example Ground Truth Guidelines:

- Segment individual body parts: ears, eyes, feet, hands, legs, mouth, nose
- Segment branches and individual eucalyptus leaves
- Allows for evaluation of precise feature detection and segmentation in algorithms

References

- Open california.
  https://www.planet.com/products/open-california/.
- Ships in satellite imagery: Classify ships in san francisco bay using planet satellite imagery.
  https://www.kaggle.com/rhammell/ships-in-satellite-imagery.
- 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.
- Andrew Hough.
  Google maps satellite photos of us government's reserve fleet.
  http://www.telegraph.co.uk/technology/google/7301167/Satellite-photos-of-US-Navy-reserve-fleet-on-Google-Maps.html, November 2017.
- 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.
- David J. Stracuzzi, Maximillian G. Chen, Michael C. Darling, and Matthew G. Peterson.
  From data to decisions: Placing machine learning challenges in context.
  CiML 2017: Machine Learning Challenges as a Research Tool.

Acknowledgements

The authors thank Aleksander Bapst, Laura Matzen, and Laura McNamara for helpful comments and discussion.

Contact: mgchen@sandia.gov

U.S. DEPARTMENT OF ENERGY

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.