Towards Universal Evaluation of Image Annotation Interfaces

Andrew M. Vernier, Jean Y. Song, Edward Sun, Allison Kench, and Walter S. Lasecki
CROMA Lab | Computer Science & Engineering | University of Michigan, Ann Arbor
{amverni, jyskwon, edwsun, akench, wlasecki}@umich.edu

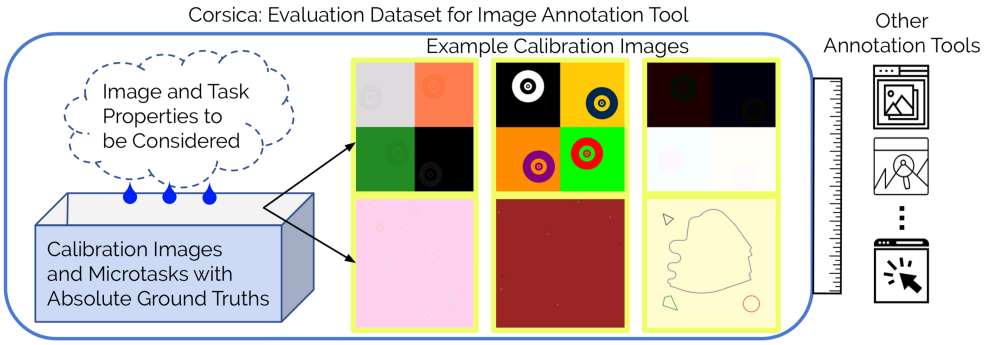

Figure 1. We introduce a test harness with non-domain specific evaluation dataset for calibration and verification of image annotation tools.

ABSTRACT

To guide the design of interactive image annotation systems that generalize to new domains and applications, we need ways to evaluate the capabilities of new annotation tools across a range of different types of image, content, and task domains. In this work, we introduce Corsica, a test harness for image annotation tools that uses calibration images to evaluate a tool's capabilities on general image properties and task requirements. Corsica is comprised of sets of three key components: 1) synthesized images with visual elements that are not domain-specific, 2) target microtasks that connects the visual elements and tools for evaluation, and 3) ground truth data for each microtask and visual element pair. By introducing a specification for calibration images and microtasks, we aim to create an evolving repository that allows the community to propose new evaluation challenges. Our work aims to help facilitate the robust verification of image annotation tools and techniques.

Author Keywords

Image Annotation; Tools; Evaluation; Crowdsourcing.

INTRODUCTION AND BACKGROUND

Image annotation tasks like image segmentation [1, 5, 6, 8], object bounding box annotation [3], or 3D object annotation [2, 4, 7], are of increasing interest for a wide range of applications. However, there has not been a standard evaluation method with absolute ground truths that can verify if a tool's performance gains generalize as intended across different types of images (small objects, poor color contrast, low resolution, etc). For example, polygon drawing tools [1, 5] are widely used for object segmentation, but it is known that the tools do not generalize to small or low-contrast objects when used alone.

We propose a test harness that consists of a calibration dataset of artificial images that simulate diverse natural challenges found in images, along with ground truth data, to address this gap. The calibration dataset supplements existing domain-specific datasets in the verification process, avoiding dataset bias in testing image annotation tools. We introduce Corsica, a proof-of-concept test harness for image annotation tools that consists of sets of three key components and the requirement specification on expanding the test sets. We hope these resources will guide the creation and verification of generalizable image annotation tools.

TEST HARNESS AND DATASET CREATION

The three components of Corsica are as follows: 1) synthesized calibration images with visual elements that are not domain-specific, which represent generalizable image properties, 2) target microtasks that represents generalizable task requirements and connects the visual elements and a tool together for evaluation, and 3) ground truths for each microtask and visual element pair. To create a set of calibration images, one has to consider two dimensions: image properties and task requirements. Image properties are the characteristics and elements of an image that must be dealt with during annotation (e.g., blurriness, curved lines, etc.). A task requirement is a capability that a tool must posess in order to complete the task correctly (e.g., path-following for organic lines). The datasets in Corsica present a range of challenges for better evaluation.

UIST '19 Adjunct, October 20-23, 2019, New Orleans, LA, USA.

ACM ISBN 978-1-4503-6817-9/19/10 ...$15.00.
http://dx.doi.org/10.1145/3332167.3357122

Examples of Image Properties and Task Requirements

Below we describe examples of image properties and task requirements, which were considered in building Corsica.

Image Property: Low Color Contrast

Low color contrast in images may cause different colors to appear similar increasing task difficulty. For example, images may have added black (e.g. objects covered by shadows) or added white (e.g. objects washed out by sun's light). Human factors (e.g. color blindness) can also cause low color contrast.

Image Property: Objects with an Organic Shape

Organic shapes arise in the images of many tasks because of their prominence in the world around us, e.g. annotating vehicles for training autonomous vehicles [7].

Task Requirement: Target Center Hitting

Some tasks require one to identify the center of an object even when it is not explicitly marked and hard to precisely pinpoint.

Task Requirement: Boundary Identification

Some tasks require users to identify the boundaries of an object or region. This may facilitate annotating the exact boundary of an object, creating a bounding box around an object, or counting the number of objects in a cluttered space. This requires correctly perceiving where that boundary is, which could be enabled or enhanced by a tool. A tool that performs poorly at this may provide noisy data to its applications (e.g., a machine learning model [1] or real-time robot vision [4]), preventing precise operation.

Task Requirement: Path Following

Some tasks require one to annotate both linear and nonlinear paths in an image. This includes tracing features of objects (e.g. a person's smile in an emotion detection task). This is a particularly difficult task because it requires fine motor skills with hand-eye coordination. Thus, an image annotation tool should be designed to reduce the manual and cognitive loads.

Task Requirement: Sub-pixel Annotation

Some tasks require higher levels of precision and accuracy than the source image can provide (i.e., potentially sub-pixel accuracies [7]) in order to accomplish their goals.

Requirement Specifications on the Dataset

Corsica is intended to facilitate in the verification of the accuracy of image annotation tools now and in the future. The images and microtasks in Corsica could be in a repository that is community-contributed. Therefore, we propose the requirement specification on the dataset to guide future contributors:

1) The synthesized images should contain visual elements that are "abstract objects" so that the images are not biased by the frequency of occurence in certain domains or applications.

2) The evaluation microtasks must come with instructions and have a defined "correct" answer (no ambiguity in answers).

3) The ground truth data accompanying the images and microtasks must be accurate down to the pixel or sub-pixel.

4) Documentation on how the image was created should be specified. An example description is "additive noise with Gaussian distribution with $\sigma = 20$ was applied".

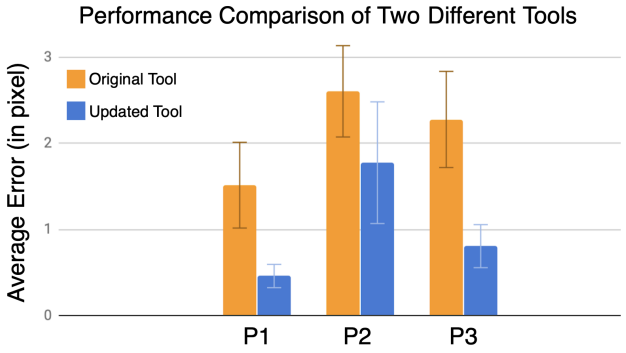

Figure 2. Corsica enables to verify the improvement of an updated tool for the targeted tool capabilities. The orange bars show the error of the original tool, while the blue bars correspond to the updated tool.

PILOT STUDY RESULTS

As a guideline, we demonstrate how to use Corsica to evaluate a tool's expected capability using images targeting the combination of low color contrast and target center hitting. For the demonstration, we compared a baseline tool, Popup [7], with a updated version that we created. We added the following features to the tool: the ability to zoom, the ability to change the contrast of the image, a hollow circle marker as the cursor. As a result of these changes, we expected the updated version to be better at accurately annotating target centers for images with low color-contrast.

We ran two annotation sessions (one on the original version of Popup and one on the updated version) with a subset of calibration images and target tasks in Corsica. The pilot study was designed as a within-subject test, comparing the performance of using the original tool with that of using the updated tool. Three annotators familiar with the original tool participated in the study. A total of 84 small microtasks per session were completed, but one task was dropped for the performance comparison because the participant did not understand the instructions. Our results showed that the updated tool outperformed the original tool in accuracy ($p < .001$). Figure 2 shows the accuracy comparison results. By testing tools with Corsica, we were able to verify the improvement of an updated image annotation tool without having to concern of the generalizability of the results.

CONCLUSIONS AND FUTURE WORK

We introduce a test harness, Corsica, that utilizes both image properties and task requirements which should be considered in the design of general image annotation tools and in the creation of datasets for the verification of such tools. Corsica uses calibration images and evaluation microtasks to use in the verification process. Future work will expand Corsica's evaluation dataset to create a rich set of images to test diverse combinations of image properties and task requirements. Using Corsica, the accuracy of image annotation tools can be evaluated to better understand where they succeed and struggle, leading to more generalizable tools and wider use.

ACKNOWLEDGMENTS

This work was supported in part by funding from the Toyota Research Institute and the University of Michigan.

REFERENCES

[1] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. OpenSurfaces: A richly annotated catalog of surface appearance *(ACM TOG '13)*.

[2] Aleksey Boyko and Thomas Funkhouser. Cheaper by the dozen: Group annotation of 3D data *(UIST '14)*.

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite *(CVPR '12)*.

[4] Sai R Gouravajhala, Jinyeong Yim, Karthik Desingh, Yanda Huang, Odest Chadwicke Jenkins, and Walter S Lasecki. EURECA: Enhanced Understanding of Real Environments via Crowd Assistance *(HCOMP '18)*.

[5] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. LabelMe: a database and web-based tool for image annotation *(IJCV '08)*.

[6] Jean Y. Song, Raymond Fok, Alan Lundgard, Fan Yang, Juho Kim, and Walter S. Lasecki. Two Tools Are Better Than One: Tool Diversity As a Means of Improving Aggregate Crowd Performance *(IUI '18)*.

[7] Jean Y. Song, Stephan J. Lemmer, Xieyang Liu, Shiyan Yan, Juho Kim, Jason J. Corso, and Walter S. Lasecki. Popup: Reconstructing 3D Video Using Particle Filtering to Aggregate Crowd Responses *(IUI '19)*.

[8] Yan Yan, Chenliang Xu, Dawen Cai, and Jason J Corso. 2017. Weakly supervised actor-action segmentation via robust multi-task ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1298–1307.