Insight GT: a public, fast, web image ground truth authoring tool

Barrantes-Garro Joel[1], Rodríguez-Morales Hellen[2], Garnier-Artiñano Adrián[1], Calderón-Ramírez Saúl[1], Porras-Jiménez Fabian[2], Corrales-Arley Luís Carlos[3], and Brenes-Camacho Ricardo[3]

[1] PAttern Recognition and MAchine Learning Group (PARMA-Group), Computing Engineering School, Instituto Tecnológico de Costa Rica, Cartago, Costa Rica.
ibarrantes@ic-itcr.ac.cr, agarnier@ic-itcr.ac.cr, sacalderon@itcr.ac.cr,
https://www.tec.ac.cr/en/grupo-investigacion/parma
[2] Industrial Design School, Instituto Tecnológico de Costa Rica, Cartago Costa Rica.
herodriguez@estudiantec.cr fporras@tec.ac.cr
[3] Huli luis.corrales@huli.io ricardo.brenes@huli.io

Abstract. This paper proposes the community the development of a public web tool for fast image Ground Truth Authoring Tool (GTAT). Image ground truth authoring tools are key to generate training and validation data for image segmentation and classification systems. The paper does a short review of similar publicly available GTAT's, its features and short-comings, in order to spot the key features missing for a public GTAT to the community. Based in the concluded wished features, we aim to develop a free and open GTAT in the future.

Keywords: Ground Truth Authoring Tools · Machine Learning · Labeling · Deep Learning

1 Introduction

Over the last few years, image analysis has taken an important role in multiple applications and fields, such as robotics, medical imaging, botany and microbiology. The sheer amount of images and videos produced for further analysis demands instruments and tools that ease the work that has to be done to obtain results. The generation and analysis of such data is increasingly assisted by several techniques associated to computer vision, pattern recognition, machine and deep learning, for instance image segmentation and classification tasks. Image segmentation refers to the pixel wise classification in an image or frame into different categories [1], which is also referred as semantic segmentation. If the classification aims to distinguish different instances of the same object category, the task is known as instance segmentation. Images and videos can be segmented to track objects like cells or distinguish relevant items from the background, as in [2–5].

Image segmentation is a common problem in computer vision, well addressed in literature. To measure metrics for an image segmentation algorithm, proper

ground truth data is needed, which means that a human must assist the segmentation of a set of images. Generally, it is required that the generated ground truth has few to no flaws, and ideally must be statistically relevant, thus several subjects must build ground truth data. Nonetheless, ground truth generation can make use of machine learning models and segmentation techniques to speed up or semi-automate the process, an useful feature increasingly available in modern ground truth authoring tools. For instance, sophisticated ground truth authoring tools implement automatic region initialization to speed the ground truth authoring process. An example of a technique which could be used to initialize ground truth masks is superpixels. In [6], pancreas computerized tomography (CT) scans are sliced into 2D images. These images are segmented using a superpixels based technique, which are later forwarded into a deep convolutional neural network to aid image classification in computer assisted diagnosis. The result is a faster and more reliable method to achieve pancreas image segmentation.

Superpixels typically over-segment the image, with groups of pixels clustered into regions or segments. The segments generated contain enough information to produce a valid initial segmentation and also, a segmentation easier to optimize compared to an initial segmentation based on a grid of pixels [7]. For example, a superpixels based web tool was also used by [5] to generate input for the training of a convolutional neural network model.

In this paper, we propose a public available ground truth authoring tool (GTAT) to validate image segmentation and object tracking algorithms. In Section 2 we address previous similar GTAT tools, to identify missing and useful GTA functionalities. Later we address our proposed GTAT, Insight GT, and perform a set of experiments to compare existing GTATs with the proposed prototype in section 4, to finally reach the conclusions and future work in Section 5.

2 State of the art

In this section, we review several tools and available software that addresses ground-truth generation for image/video segmentation and object tracking.

FAST-GT (FAst Semi-automatic Tool for Ground Truth generation) [8] is a generic framework for semiautomatic generation of ground truth, which allows different implementations of building blocks. FAST-GT implements a detection layer, which applies multiple object detectors to a frame or image, and takes into account previous annotations (called trackers). Manual intervention is then applied to the output of the previous layer. Finally, the trackers of the detection layer are updated according to the resulting annotation of the manual step. This tool is available to the general public. Its most important drawback is the need of installing and compiling the source code, which depends on the OpenCV and Eigen libraries.

LabelMe [9] is a GTAT which consists of a database of labeled images and a web tool to manually annotate images. The LabelMe web tool provides an

unsophisticated canvas that allows the user to annotate an image using a toolbox consisting of a polygon tool and a paintbrush to generate a mask. LabelMe presents the disadvantage that only accepts images in JPG format and the toolbox available to annotate images is quite restricted.

In [10], VATIC is proposed as a tool that helps video annotation tasks by providing a web based platform for crowd-sourced video labeling. VATIC allows the user to annotate and delimit objects by surrounding them with a rectangular shape. The user only needs to manually annotate a subset of the frames of a given video, called the key frames. Then, VATIC performs an annotation for the remaining frames, using interpolation methods. VATIC is web based, however there is no free server running it, and it does not allow pixel wise tagging.

Sensearea, a public video editing tool, provides users with interactive tools to perform video editing and effect generation, as also object tracking in videos and ground-truth authoring. Before the object tracking process, the user can annotate the first frame of the video, using basic tools such as paintbrush, polygon and ellipse tools, for example. Then, the user can start the object tracking process to readjust the mask automatically, enabling the user to correct it afterwards. We experienced major performance shortcomings when drawing many masks, while using Sensarea. Another drawback of Sensarea is its low portability given the need of a Windows based installation.

In [11], it is presented interactive Video Annotation Tool (iVAT) as a tool that aid and ease annotation tasks. iVAT provides manual, semi-automatic and automatic annotation for videos. In the manual mode, the user must generate a ground-truth mask for each frame on the video. The semi-automatic approach requires ground-truth annotation for a given frame, to later calculate an annotation mask for the next frames automatically. The automatic annotation involves supervised detectors where a learning step has taken place, making the automatic approach domain dependent. The tool does not allow image GT authoring, and is not web based.

Ilastik [12] is a GTAT proposed as an easy-to-use tool to perform image segmentation and classification. Ilastik makes use of user manual annotation to begin a learning phase, which usually consists on mouse strokes across a canvas to label regions of pixels. The tool implements the following steps: first, ilastik calculates a generic basis to represent general image features. Later, a random forest classifier is trained using user-labeled data to initialize pixel labels. Ilastik is not web based, and its GUI usability is limited.

In addition, the tool Supervisely [13] allows to mark both images and videos in a semi-automatic way, selecting the desired area to create the marking and automatically generating the required shape. Supervisely also has tools to draw the masks in a completely manual way. It has quick access commands to make the marking process more efficient. Another important function is the ability to make brightness and contrast adjustments to improve the marking process.

Another tool analyzed is LabelBox [14], it is a platform that allows to mark and classify images and video manually. Labelbox enables collaborative work, making possible to see and review the markings made by other people. A de-

ficiency found is that to mark video it must be converted into a sequence of images before uploading it. Also it requires the user to define the objects before they start labeling the images and if there is a new or different object, the user must go to the project menu, settings and add the new object.

Table 1 summarizes a feature comparison of the analyzed GTATs.

3 Proposed tool

Given the presented state of the art, we propose Insight GT, a public web tool consisting in a canvas that allows manual image annotation and a semi-automatic segmentation approach to generate ground truth data from 2d images and videos.

The following the proposed functionalities implemented so far in our prototype available at `https://insight-gt.hulilabs.xyz/canvas/`. For region initialization, SLIC superpixels are computed [15]. A javascript implementation of the algorithm can be found at [16].

- A canvas that allows manual image annotation with different brushes and tools, as seen in [9, 17, 18], aided by region initialization, using super pixels or a similar algorithm.
- Web tool: The GT authoring tool must be web based, to increment the tool portability, and open to the community.
- Flexible storing format: The format must allow its usage with GT compromising several (thousands) of GT masks, avoiding performance degradation.
- A semi-automatic algorithm for mask initialization: The tool must implement a technique for initializing the foreground masks, and allow user to make adjustments to the masks.
- A semi-automatic algorithm for mask readjustment: For video segmentation, the masks drawn for the first frame can be readjusted for next frames. An algorithm which automatically estimates such readjustment is useful for a GTAT, as implemented in Sensarea.
- Collaborative support and crowd-sourcing support: The tool must allow several concurrent users working on the same project.
- Evaluation module: The tool must provide means to evaluate metrics for the segmentation algorithms, including but not limited to: sensitivity, specificity, accuracy, F-score, among others.
- Change history: the changes made by other users in the markings of the images are shown in the tool.

4 Experiments and results

We selected Supervisely and LableBox in order to compare it to Insight GT. This applications where selected because they are all web applications. Additionally, they allow to label different objects in order to track them.

Table 1. Comparison of GTATs

	Proposed Tool	FAST-GT	ITK-SNAP	ROXAS	LabelME	Sensarea	iVAT	TWS	ilastik
Platform	Web based	Desktop	Desktop	Desktop	Web based	Desktop	Desktop	Desktop	Desktop
Cross-platform	●	○	●	○	●	○	●	●	●
Image batch segmentation	●	○	●	●	○	●	○	●	●
Video segmentation	●	●	○	●	○	●	●	○	○
Manual segmentation	●	○	●	○	●	●	●	○	●
Available toolbox	Brush, polygon	Rectangle	Brush, polygon	?	Brush	Brush, Polygon	?	Scribble	Scribble
Semi-automatic segmentation	●	●	●	●	○	●	●	●	●
Automatic segmentation	○	○	○	●	○	●	●	●	●
Mask correction	●	●	?	●	○	●	●	○	●
Region initialization	●	○	○	○	○	○	○	○	○
Metric Evaluation	●	○	○	●	○	○	●	○	○
Restricted Domain	○	○	●	●	○	○	○	○	○

4.1 Quantitative experiments and results

To quantify user experience for each selected tool, we defined the following tests:

1. Open an image and with the rectangle tool mark the same sample image with car that is in it. The input image is stored in a known location.
2. Mark with the brush tool one of the figures in the image. For this second task, the user will modify an image that is already open.
3. With the eraser tool, correct the edge that is left of the image. The user must open a mask that has been previously marked with a flaw easy to detect.
4. Save a previously created mask. The file containing the mask representation must be in a previously known route by the user.
5. Open a previously created mask again. The file containing the mask representation must be in a previously known route by the user.

Table 2. Mean and Std. time in seconds for each test in each tool

Tool		Test 1	Test 2	Test 3	Test 4	Test 5
Insight GT	**Mean**	45.653	19.503	18.023	10.188	38.418
	Std.	25.698	15.413	7.703	5.959	36.495
LabelBox	**Mean**	83.995	19.069	11.185	103.97	20.454
	Std.	13.654	6.838	7.055	24.636	12.531
Supervisely	**Mean**	25.33	26.805	29.740	57.044	13.905
	Std.	7.835	35.094	78.022	17.748	5.353

The following are key aspects taken into account during the execution of the proposed tests.

- Only initial and simple instructions were given to the user. We wrote down the users' mistakes.
- We performed simple questions to understand the user's thought process.
- If the subject performs a faulty action several times, the test is finished. We aim to understand how the person performs a task without knowledge of the application, if the user tries many times to perform an action, it is likely that the user will learn to do it by trial and error, and not by intuition.
- We collected a set of observations and suggestions per user for each tool.

We tested 30 university engineering students, between the age of 18 and 25 that had no previous experience with any of the tools presented to them. Then they were split into 3 groups of 10 for each tool. This was made in order to prevent the user to learn from the similarities from the other tools.

The first test measured the speed of opening an image for a project. Here is the first difference between the platforms. In Insight GT user performed this action in an average of 46 seconds, but it only loads one image, for LabelBox and Supervisely the is able to select a folder or a group of images for the labeling project.

The process in Supervisely was very straightforward and simple for loading a group of photos. Labelbox was also simple to load images, but it required a couple extra steps like defining the objects and color for labels, the tools to use, and there a couple of screens that let you pick the objects that you will mark on the images before you get to the marking screen.

The second and third tests the results were similar, with Insight GT yielding the lowest times by a low margin. The resulting similar times for these tests are likely to be explained by the simplicity of the tests. In the case of LabelBox, the process to enable the necessary functionalities for tests 2 and 3 required to manipulate a JSON file, however we did not include the time to perform this task.

In the fourth test, LabelBox was the slowest, since it required the user an extensive search for the functionality interface. Additionally, Labelbox only allows to download the whole data set. As for Supervisely, the process was rather

simple, however it downloads a JSON file, not the mask in an image format. Insight GT yielded the lowest average time for the fourth test.

For the last test, corresponding to opening a previously saved mask, it was faster in average for Supervisely and LableBox, since both tools implemented a cloud based recently saved mask option access.

5 Conclusions

The test showed that in the core functionalities of the tools, they are similar, but upon further inspection of the work flow, it can be seen that the Supervisely and LabelBox need some extra effort in learning the tool for some parts of the process that require editing JSON files. Insight GT keeps the core functionalities simple and easier to learn.

We aim to make available Insight GT publicly available with its core functionalities, along its source code. We think its important to build an user friendly and web based GTAT with a powerful set of customizable mask initialization algorithms to speed up image and pixel labeling.

As future work, we think it is possible to improve the accuracy of the usability tests, with more accurate and modern tools like eye tracking, which would allow us to validate and compare more accurately the implemented functions of the proposed Insight GT tool with existing tools.

Thus, as future work, an user experience research should be develop to find the requirements of the users with the objective of enhance the learning curve and add features that improve the user experience and learning curve of the web tool Insight GT.

References

1. L. G. Shapiro and G. C. Stockman, *Computer vision*. Prentice Hall, 2001.
2. J. Liu, X. Tong, W. Li, T. Wang, Y. Zhang, and H. Wang, "Automatic player detection, labeling and tracking in broadcast soccer video," *Pattern Recognition Letters*, vol. 30, pp. 103–113, 1 2009.
3. L. Grady and G. Funka-Lea, "Multi-label Image Segmentation for Medical Applications Based on Graph-Theoretic Electrical Potentials," pp. 230–245, Springer, Berlin, Heidelberg, 2004.
4. J. S. Suri, S. K. Setarehdan, and S. Singh, *Advanced Algorithmic Approaches to Medical Image Segmentation : State-of-the-Art Applications in Cardiology, Neurology, Mammography and Pathology*. Springer London, 2002.
5. A. S. Aydin, A. Dubey, D. Dovrat, A. Aharoni, and R. Shilkrot, "CNN Based Yeast Cell Segmentation in Multi-Modal Fluorescent Microscopy Data,"
6. A. Farag, L. Lu, H. R. Roth, J. Liu, E. Turkbey, and R. M. Summers, "A Bottom-Up Approach for Pancreas Segmentation Using Cascaded Superpixels and (Deep) Image Patch Labeling," *IEEE Transactions on Image Processing*, vol. 26, pp. 386–399, 1 2017.
7. X. Ren and J. Malik, "Learning a Classification Model for Segmentation,"

8. F. Comaschi, S. Stuijk, T. Basten, and H. Corporaal, "A tool for fast ground truth generation for object detection and tracking from video," in *2014 IEEE International Conference on Image Processing, ICIP 2014*, 2014.
9. B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 157–173, 2008.
10. C. Vondrick, D. Patterson, D. Ramanan, C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently Scaling Up Crowdsourced Video Annotation A Set of Best Practices for High Quality, Economical Video Labeling," *International Journal of Computer Vision manuscript*, 2012.
11. S. Bianco, G. Ciocca, P. Napoletano, and R. Schettini, "An interactive tool for manual, semi-automatic and automatic video annotation," *Computer Vision and Image Understanding*, vol. 131, pp. 88–99, 2015.
12. C. Sommer, C. Straehle, U. Kothe, and F. A. Hamprecht, "Ilastik: Interactive learning and segmentation toolkit," in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 230–233, IEEE, 3 2011.
13. Supervisely, "Overview," 2019. [Online]. Available: `https://docs.supervise.ly/cluster/overview/`, [Accessed: May. 05, 2019].
14. LabelBox, "Overview," 2019. [Online]. Available: `https://support.labelbox.com/docs`, [Accessed: May. 11, 2019].
15. R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels," 2010.
16. P. Tangseng, Z. Wu, and K. Yamaguchi, "Looking at outfit to parse clothing," Mar 2017.
17. P. Bertolino, "Sensarea: an Authoring Tool to Create Accurate Clickable Videos,"
18. P. A. Yushkevich and G. Gerig, "ITK-SNAP: An Intractive Medical Image Segmentation Tool to Meet the Need for Expert-Guided Segmentation of Complex Medical Images," *IEEE Pulse*, vol. 8, pp. 54–57, 7 2017.