@author: Isaac S. Robson
@class: COMP 776: Computer Vision, F18
@prof: Professor A. Berg
@date: 11 December 2018

PotatoNet

A Well-Rounded Introduction to Instance Segmentation in Precision Agriculture

Abstract. Instance segmentation, concerned with pixel-wise detection, identification, and classification of individual objects, is an increasingly understood problem in the AI field of computer vision. However, instance segmentation is not without its hazy boundaries, unseen verticals, and blindspots. Here we discuss one application of instance segmentation to the growing field of precision agriculture, which is concerned with augmenting and automating the ancient field of farming with the help of data science, robotics, and AI. Here we construct a new instance segmentation dataset of potatoes, a worldwide popular and semi-ubiquitous tuber, and in the process encounter many of the historical and current open problems in computer vision. Once constructed, we conduct an exercise in training an instance segmentation neural network using the recent Mask R-CNN architecture pioneered by He et al. (2017).

1 Introduction

Startups developing farm automation solutions like Sweeper, the "creepy-cute robot that picks peppers with its face" Simon (2018), and major technology and agricultural companies alike are rushing to develop innovative technologies for the agricultural sector. Fertilizer can only go so far, and now precision watering, pest identification, and robotic harvesting are allowing for higher, healthier yields at lower prices with fewer resource usage.

Figure 1: The Sweeper Robot, taken from official promotional material

Of course, many of these techniques rely on computer vision and 3D reconstruction, which has had many open source benchmarks and models, but these tools and architectures are oftentimes designed for facial recognition or multi-class object detection in clean, low-occlusion, well-lit, no crowd settings with high resolution cameras. Contrast this with the agricultural setting where internet connectivity and hardware availability may be constrained in addition to robots with moving parts that have to get their grippers dirty. While some open source and academic work exists regarding precision agriculture, many of the datasets developed for this type of work is a strong competitive advantage due to its practicality but high cost to acquire and maintain.

This project, with its emphasis on potatoes, seeks to be one of many investigations into intersection of computer vision and precision agriculture. While potatoes represent a major food staple around the world, the harvesting of potatoes is already largely mechanized unlike other crops like strawberries or grapes, and potatoes primarily grow underground, unlike wheat or rice. As such, segmentation datasets focusing on potatoes have less potential for commercialization, aside from downstream tasks like bagging, peeling, and sorting.

We thus feel that potatoes represent a more neutral 'Switzerland' of agricultural product to conduct agricultural computer vision on, and in the process, to further the field of computer vision by offering several interesting avenues of research.

For one, the availability of potatoes in both developed and devloping countries lends itself to the collection of potato images. Creating datasets from these images may prove to be a more challenging process, but researchers and potato fanatics worldwide can easily contribute their own data to develop and expand datasets. Secondly, the International Potato Center estimates some 4,000 distinct varieties of potatoes spanning over 180 species. This means that recognizing a potato requires a learning algorithm to focus on many different combinations of shapes, sizes, and potato skin colors, many of which are transparent and lightly marbled, not unlike human skin. Thirdly, potatoes not only ripen with time, they also grow eyes and new roots as they develop into a plant, but unlike many other plants, this growth can be examined without specialized equipment or even planting (as it occurs to some extent in open air).

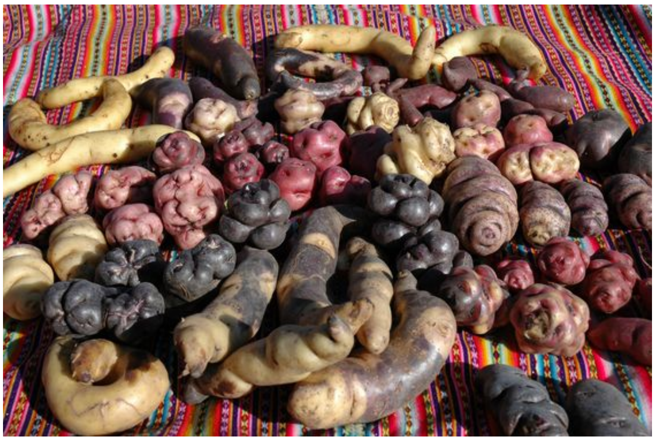

Figure 2: Several varieties of potatoes, taken from the website for the International Potato Center

Additional fun properties of potatoes are the 'swarm emergence' of piles when stacked together, frequent crowding of potatoes in piles, transformation from raw potato to cooked food, the abundance of both potatoes covered in dirt and potatoes wet from washing (good luck getting all of the humans in Microsoft COCO: Common Objects in Context, Lin et al. (2014), covered in dirt or soaked in water), and relatively long shelf life of a potato compared to many other fruits and vegetables (which allows for the same potato to be photographed in many different settings). Many potato carvings, animations, and cartoons also exist, so converting from CGI potato sketches to real life potatoes is also a potential development.

2 Dataset

Images were scraped from Google Images and Flickr, before individual instances were annotated with VGG Image Annotator (VIA) Dutta et al. (2016). Images were primarily selected for having a low to medium number of potatoes present for ease of segmenting. In order to ensure individual instances did not overlap in the final dataset, binary erosion was performed on the resulting mask arrays with OpenCV between overlapping instances. Non-overlapping portions of instances were kept, while overlapping portions were replaced by their erosions (eroded until no remaining overlap). The resulting dataset consists of 97 images with 1007 distinct instances, ranging from a single instance per image (11 images) to one image with 37 instances. A broad range of image resolutions are present, from few hundred pixels by a few hundred pixels to several thousand by several thousand.

No distinctions or additional annotations were made for different varieties of potatoes. Note that while sweet potatoes were included in the potato dataset, African yams were not.

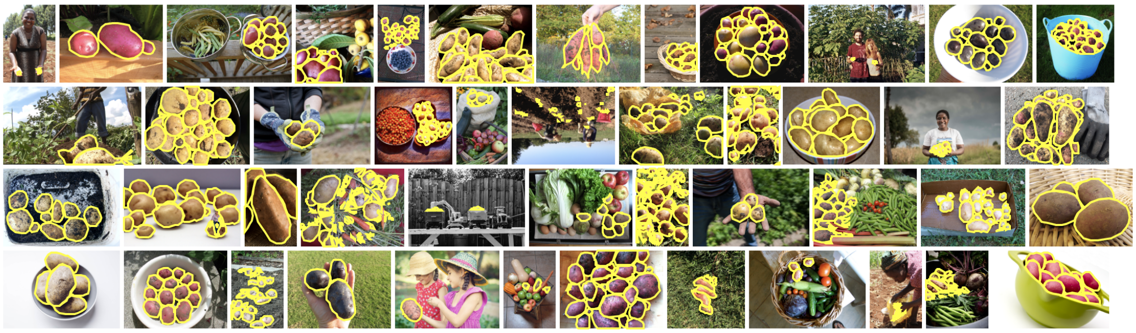

Figure 3: Sample annotations from VIA

Two different techniques were used to handle occlusions resulting in discontiguous polygons, either narrow 'bridges' between portions of the same instance or post-mask array creation merging manually. Some eyes and sprouting roots were included in the instances and some occlusions from leaves, branches, and dirt were included as well (others are excluded). The segmentations are mostly consistent as they were prepared by the same single annotator, but some variations certainly occur. No analysis was conducted here to determine an 'optimal' inclusion or exclusion criteria for the area for an instance.

Dataset visualizations were created by overlaying a transparent version of the binary mask for a unique RGB color. Due to the nature of our selected random color generator for the procedure, many masks are displayed as either red, blue, or green, although instances of similar colors are not necessarily related. Instances split into multipled polygons, however, are the same color (as the polygons are from the same instance).

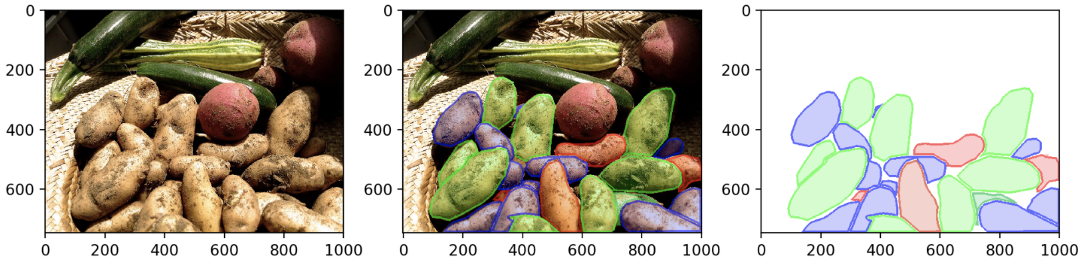

Figure 4: Sample segment masks post-erosion.
Left: Original Image. Center: Overlap of masks and image. Right: Instance Masks

3 Model

A Mask R-CNN model was trained in PyTorch 1.0 using the `maskrcnn_benchmark` Github repository created by Facebook Research (https://github.com/facebookresearch/maskrcnn-benchmark). The specific architecture used was a Mask R-CNN RPN using an ImageNet-trained ResNet-101 backbone. A base learning rate of 0.005 and the default learning rate scheduler were used for 3000 iterations. Of the original 97 images, 87 images were used in a training set, 8 were used as a development set, and 4 were left out as a test (evaluation) set (which are approximately the percentages used, respectively).

In order to train such a model on this specific dataset, a `PotatoDataset` class was created and inheriting from the Python `cocoApi` (https://github.com/cocodataset/cocoapi/tree/master/PythonAPI) to be used with the `maskrcnn_benchmark` training and testing scripts. Additional evaluation, dataset factory, configuration, and dataloader scripts were also created. To make the segmentations compatible with the `maskrcnn_benchmark` pipeline, we converted the eroded instance masks into bounding boxes and polygons in a `cocoApi`-friendly format. The data and API were then loaded onto a virtual machine on Google Cloud with an NVIDIA Tesla K80 GPU attached running CUDA 10.0.

Much of the model configuration file was unmodified from the original ResNet-101 FPN configuration. This means the model was trained with stochastic gradient descent plus momentum and weight decay. Multiple initializations were not used as training a model was primarily for exercise and proof-of-concept purposes rather than for competition or benchmark results.

4 Results and Discussion

The predictions from the fitted model on the 4 test set images can be seen in Figure 5. Bounding boxes and segmentations with scores below 0.8 are filtered out. We see that many potatoes are recalled correctly, although there is a high false positive rate such as on the wicker in the first image and the apples in the second image. Negative sampling from other agricultural datsets would likely easily remedy the false positives, and considering the small training size, the pixelwise segmentation of true positves seems remarkable.

Numerical scores for average precision and intersection over union are omitted, although the small test set size means visual inspection is a fairly reliable estimate of performance on the test set.

Many interesting future extensions and investigations can be imagined. Additional annotations such as species/variety of potato and age/disease state may be interesting subclasses that may improve the accuracy of the model and generalizability across unseen varieties. Analyzing the usefulness of distinct techniques for handling partial or near-total occlusion and crowding would also be interesting. Translating images of dirty or wet potatoes to clean or sprouting equivalents would also be an interesting avenue of research, and a comparably easy dataset to procur compared to covering humans, animals, or even other fruits (due to the relatively durable nature of potatoes).

Additionally, while the images chosen for this project were focused on individual potato instance segmentation a low to medium number of potatoes, it is not unreasonable to wonder if this approach to instance segmentation is a good approach to the task at hand. Large piles of potatoes may be better characterized with a volumetric approach and intermediate clusters may be better characterized as a 'cluster' of potatoes with a regression learning algorithm to predict the number of potatoes present. Regardless of exact implementation, some hybrid of what are now called semantic segmentation and instance segmentation may prove more efficient or useful for actual precision agriculture tasks.

A particularly promising avenue may be to acquire large datasets on a particularly abundant agricultural product such as potatoes, and leverage this dataset into a transfer learning scenario for building more customized segmentation algorithms for agricultural products that are more rare. This may either be a single class case, where a simpler architecture may suffice, or a multiclass case, which may include separate classes for roots, stems, and leaves as refernce landmarks. Granted, many of these applications may be run on mobile GPUs so simpler and smaller architectures may be preferred.

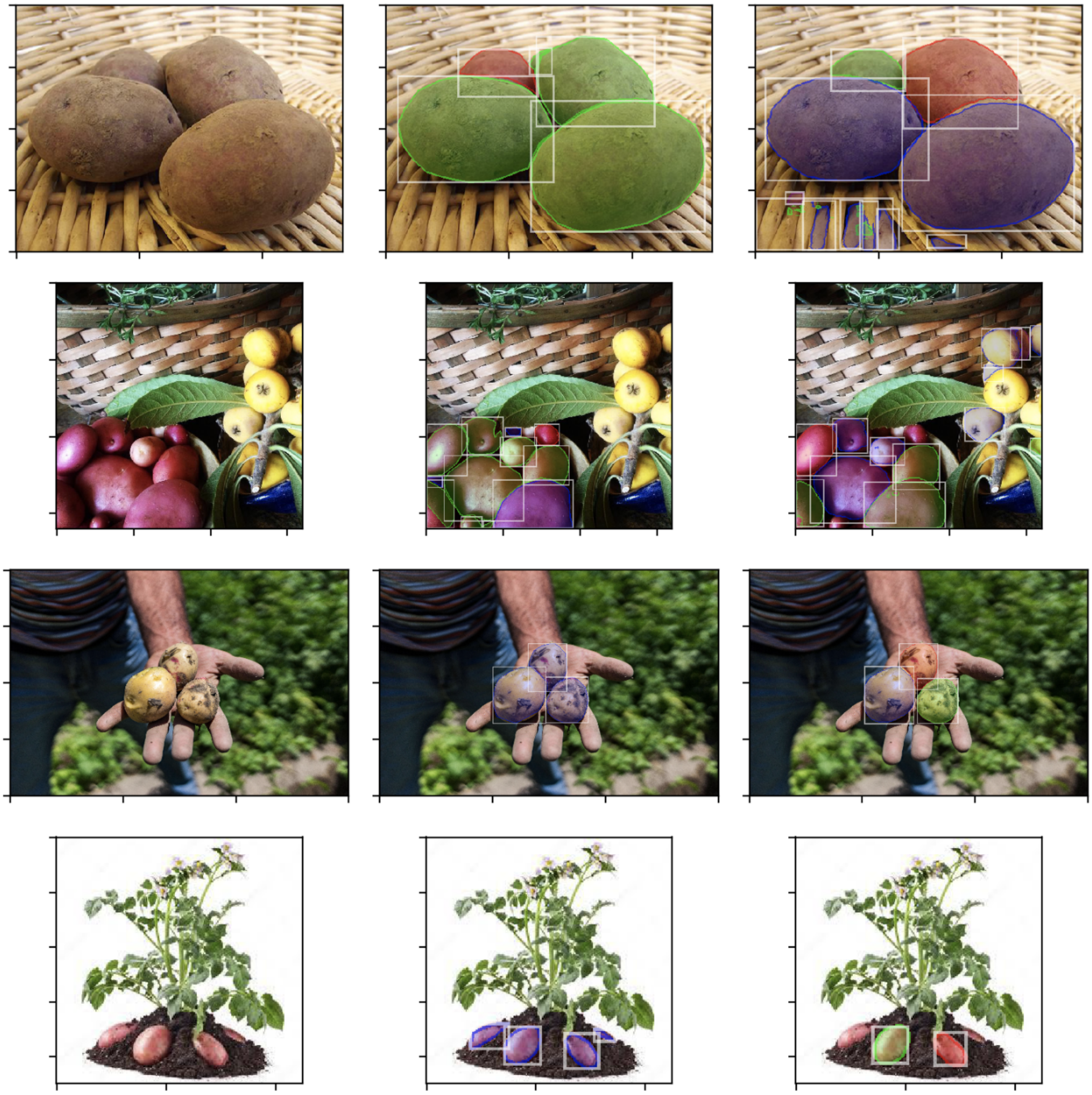

Figure 5: Test set predictions for scores above 0.8
Left: Original Image. Center: Ground truth masks + bounding boxes. Right: Predicted masks + bounding boxes.

5 Conclusion

The intersection of computer vision and precision agriculture is a broad but promising field. Here, we've created a new dataset focused upon instance segmentation of several varieties and species of potatoes in a number of different image resolutions. We've discussed how the annotations generated with VIA can then be intelligently eroded to preserve uniqueness of individual pixel labels, and these masks converted into a `cocoApi`-friendly format. A Mask R-CNN model capable of learning was generated by creating a `PotatoDataset` class and extending the the `maskrcnn_benchmark` pipeline API. The resulting convolutional neural network yielded a moderate degree of accuracy with the ability to detect and segment individual potato instances. As discussed previously, a simple yet effective way to improve this neural network would be via negative sampling on similar agricultural images that do not contain potatoes.

In the process of describing our work, we've presented and discussed some thoughts and findings about creating an instance segmentation dataset for agricultural products. Far more investigation into the handling and prediction of discontiguous instances, occlusions such as leaves, grass, and dirt, as well as the crowding of objects in many such images is warranted. Additional applications such as transfer learning to new agricultural products and the inclusion of additional annotations such as potato species or variety are also worth examining.

References

[1] International potato center. https://cipotato.org/. "Accessed: December 10, 2018.

[2] A. Dutta, A. Gupta, and A. Zissermann. VGG image annotator (VIA). http://www.robots.ox.ac.uk/ vgg/software/via/, 2016. Version: 2.0.2, Accessed: October 20, 2018.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[4] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[5] Francisco Massa and Ross Girshick. maskrnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. https://github.com/facebookresearch/maskrcnn-benchmark, 2018. Accessed: December 5, 2018.

[6] Matt Simon. The creepy-cute robot that picks peppers with its face. https://www.wired.com/story/the-creepy-cute-robot-that-picks-peppers/, Sep 2018.