TomatOD: EVALUATION OF OBJECT DETECTION ALGORITHMS ON A NEW REAL-WORLD TOMATO DATASET

V. Tsironis, S. Bourou, C. Stentoumis

up2metric P.C., 11521, Athens, Greece – (tsironisbi, stavroula.bourou, christos)@up2metric.com

Commission III, WG III/10

KEY WORDS: Object Detection, Benchmark, Precision Agriculture, Dataset, Deep Learning, Classification

ABSTRACT:

The integration of modern technologies in farming poses a challenging task to the research community. In this work, the task of selective cropping and treating is considered, whereas learning algorithms can provide essential assistance on crop growth and disease prediction, species recognition and fruit detection. In this paper, we introduce a highly specialized object detection (OD) and classification dataset of tomato fruits that contains class information for the ripening stage of each tomato fruit apart from its corresponding bounding box. With this dataset we aim to encourage the development of task-specific production ready object detection algorithms, as well as to evaluate and provide a baseline result of common state-of-the-art generic OD algorithms. In detail, a thorough presentation of the most common OD datasets takes place, where we discuss both generic OD and some highly specialized datasets. Our dataset contains more than 250 images and 2400 annotations in total. The dataset contains class information for three ripening stages of a tomato fruit provided by expert agriculturists, while providing views consistent with the targeted real-world use case scenario. Compared to other OD datasets our proposition differs in core areas such as the quality of the annotations, the object size distribution and the public availability. Evaluating the performance in our dataset for six object detection models we draw conclusions about the strength and weaknesses of each one's performance. Finally, we present a future roadmap of revisions and discuss some future research topics that could improve the performance of OD algorithms in our dataset.

1. INTRODUCTION

Agriculture plays a crucial role in humanity's everyday life as well as in the global economy. As world population increases rapidly and the natural resources are running out, the need for precision farming is becoming more and more evident. Advanced technologies can be used to provide automated solutions for manually performed tasks supporting, thus, precision farming. Hence, the integration of those modern technologies in farming poses a challenging task to the research community from a variety of scientific fields such as Remote Sensing, Machine Learning or Robotics. Especially regarding the field of cropping management, the integration of advanced IT technologies in farming can provide essential assistance on crop prediction, species recognition, disease prediction, etc.

Thus, there is an increased interest in the development of deep learning algorithms for various agricultural procedures. For instance, (Pantazi et al., 2016) present a novel method that focuses on crop prediction. This proposed method accumulates data from satellite imagery, crop growth characteristics and in-situ soil measurements for accurate wheat yield prediction. Regarding the species recognition task, (Grinblat et al., 2016) introduce an algorithm for identification and classification of legume species from leaf vein morphological patterns. Furthermore, concerning the disease detection diagnosis, (Ferentinos, 2018) proposes a CNN-based method, that classifies leaves as healthy or diseased in various plants from images.

The accurate detection of different crops as well as the identification of their ripening stage are considered challenging and essential tasks. For the achievement of accurate results at these tasks, the need of a high-quality labelled dataset is required. However, there is a lack in such publicly available benchmark datasets for precision farming, which restricts the efficient application of modern technologies, like machine learning algorithms, in greenhouses. To encourage and support the detection of specific crops, we introduce a dataset for tomato fruits detection.

Tomato is one of the most popular vegetables that plays a significant role in agricultural economy. Due to its planting characteristics, such as large planting area, the collecting of tomato is a time-consuming and intense activity. Moreover, its characteristics, like the short lifespan and sensitivity of the fruit, make tomato collection a difficult and delicate procedure that needs accurate timing. Thus, the development of an automated collecting mechanism, that can help tomato harvesting activity efficiently is particularly important. A critical part in the research of automated collecting mechanisms is object detection techniques based on images of tomato fruits in greenhouses. Therefore, the existence of a high quality and realistic tomato dataset is equally important, as can be used for the precise detection of those fruits.

Across literature there is a severe lack of such datasets, especially considering those freely available, while the few that exist differentiate over the proposed dataset in various aspects, such as the vastly different viewing angle or the very small quantity of available data. In the context of this work the task of selective cropping and treating is considered, where machine learning algorithms can provide essential assistance on crop growth and disease prediction, species recognition and fruit detection. The main objective of this work is to address the need for a task-specific object detection dataset for tomato fruits, namely "*Tomato Object Detection* (*TomatOD*)", for precision agriculture applications that typically require highly accurate localization.

In this paper, we introduce a highly specialized, novel object detection and classification dataset of tomato fruits, that contains class information for the ripening stage of each tomato fruit apart from its corresponding bounding box. The ripening stages of tomatoes can provide information about their harvest; in this dataset these stages are represented by three classes, unripe, semi-ripe and fully-ripe. Moreover, the *TomatOD* dataset is freely available for research purposes.

The rest of the paper is structured as follows. In section 2, a thorough review of typical object detection dataset is presented, while the introduced *TomatOD* dataset and its statistical analysis is described in Section 3. In Section 4, some state-of- the-art detection algorithms are presented and, in Section 5, those algorithms are trained and evaluated on our proposed dataset. Lastly, the main impact points of *TomatOD* dataset are revised in Section 6, while we draw some conclusions and propose further future work.

	Pascal VOC2012	***ILSVRC2017 (ImageNet)***	***COCO***	***Open Images V6***
Images	12K	476K	328K	1.7M
Annotations	28K	534K	2.5M	15M
Classes	20	200	91	600
Instances / Image	2.3	2.7	7.7	8.3

Table 1. Comparison between common OD datasets

2. OBJECT DETECTION DATASETS OVERVIEW

Object Detection (OD) is one of the most challenging problems in computer vision, aiming to determine the location of certain objects on images and videos as well as to classify them among specific classes. Typically, the localization of object is described by a bounding box. OD can be applied in a wide range of applications like autonomous driving, object/people tracking, security and transportation field, etc. Recently, the introduction and free distribution of huge OD datasets has been the most contributing factor in the wide use of deep learning algorithms in the field of object detection. Some of the most common OD datasets for various tasks are descripted in this section.

One of the most well-known benchmarks in object classification and detection tasks is the PASCAL Visual Object Classes (VOC) dataset introduced in (Everingham et al., 2010). The PASCAL VOC challenge had been released annually from 2005 to 2012 and it is considered a reference dataset in object detection tasks, even though it contains only 20 basic categories. The first version was released in 2005 and it contained only 4 categories. In 2007 version of the PASCAL VOC dataset, the fixed final number of 20 classes was introduced. In the last iteration, VOC2012 dataset, it has over 12K images, with more than 28K annotated objects of 20 classes. Moreover, the PASCAL VOC dataset is not exposed to systematic bias in its data, like image centered objects, good illumination or non-occluded objects.

Following PASCAL-VOC, the first large-scale dataset is the ImageNet (Deng et al., 2009), containing over 14 million images organized into over 1000 classes. The images of this dataset collected from various online sources, which have variable appearances, positions, viewpoints, poses, background clutter, occlusions and lighting conditions. Focusing on generic object localization and detection tasks, the ImageNet Detection (ILSVRC17 DET) dataset uses a subset of 476K images of ImageNet Dataset with 534K bounding box annotations from 200 categories.

One of the most used datasets in OD is the Common Objects Context (COCO) dataset (Lin et al., 2014), published by Microsoft. It contains more than 2.5M labelled instances of 91 object categories in a total of 328K images, where 82 out of 91 categories of this dataset have more than 5K labelled instances. The categories are selected in order to be a representative set of all real-world categories and to be relevant to practical applications. The dataset is targeted at the detection of objects that can be found in everyday life in their natural environments.

Additionally, the Open Images dataset contains diverse images with complex scenes and several objects per image. Since its introduction in 2016, the creators of the dataset have published several updated versions, with the 2018 version, namely the Open Images V4 (Kuznetsova et al. 2018), to be among the most popular iterations. In 2020, the most contemporary version of this dataset, Open Images V6, was published, containing more than 9 million annotated images, while the annotations include object bounding boxes, image-level labels, object segmentation masks, visual relationships and localized narratives.

A numeric comparison of these four object-detection focused datasets can be found in Table 1. It is observed that the ImageNet provides much more categories and significant more images than

Figure 1. Sample images with annotations. Tomatoes below a certain size are considered out of scope. Green boxes are of class unripe, orange boxes of class semi-ripe and red boxes of class fully-ripe.

PASCAL VOC. COCO, on the other hand, contains fewer categories, but more instances per category than ImageNet. Moreover, the COCO dataset has more categories than the PASCAL VOC dataset as well as significantly more instances per category. However, Open Images dataset is the largest dataset of the four, given the number of total images, annotations and categories. Another crucial observation among these datasets is the number of labelled instances per image; the COCO dataset contains significantly more instances per image than both PASCAL VOC and ImageNet dataset. Specifically, it has an average of 7.7 instances per image instead of 2.3 and 2.7 instances per image respectively. Once again, the Open Images dataset is the most information rich of the four, since it contains 8.3 instances per image. Additionally, 10% of COCO images have one class per image, while the ImageNet contains more than 60% of images with a single object class. PASCAL and Open Images have over 60% and 20% of images with one category.
The datasets were designed to address the need for generic object detection; however, some particular classification and detection tasks demand more specialized, "task-specific", datasets. In the field of autonomous driving, for example, the KITTI dataset (Geiger et al., 2012) is one of the most well-known benchmarks for 2D and 3D object detection for this task. For human action understanding, the AVA (Gu et al., 2018) is a labelled video dataset with audiovisual annotations, aiming to improve the understanding of human activity. Regarding, the classification and detection of animal species and plants in the real world, the iNaturalist Species dataset (Van Horn et al., 2018) is among the most common choices.

Regarding classification and detection tasks for Precision Agriculture, specially designed datasets also do exist. Relative to crop detection, the CropDeep dataset (Zheng et al., 2019) is a specialized dataset for species classification and detection of common vegetables and fruits. Specifically, CropDeep have classes of different parts and growth stages of vegetables and fruits, as well as similar parts of different different species, like flowers and leaves. It contains more than 31,000 images with over 49,000 annotated objects of 31 different classes. The images were collected by IoT visual cameras, autonomous robots and smartphones in greenhouses. The dataset contains different parts and growth periods of vegetables and fruits, but also similar parts of different species. Among the 31 categories, the CropDeep dataset contains four growth stages of tomato, which are ripe, unripe, early-blossom and full-blossom. The dataset is not publicly available, although it can allegedly be provided from the author by email.

Focusing on the detection of key organs of tomatoes using CNN architectures, the work of (Sun et al., 2018) introduces a dataset of over 5,000 images with annotated objects of tomato flowers, immature green and mature red tomatoes. The images are mostly close-up shots and were collected with high definition camera on greenhouse in different times and light conditions. It is not mentioned in the paper if the dataset is publicly available and whether the annotation process was done by experts or not. Another specialized dataset for grape detection is introduced in (Santos et al., 2020). This dataset is publicly available and contains 300 images with over of 4,000 boxed clusters from 5 different grape varieties, as well as binary segmentation masks for a subset of clusters.

3. TOMATOD DATASET

TomatOD is a novel dataset aiming to provide a realistic use case scenario for a highly specific task, whereas a robotic arm navigates across the corridors of a soilless tomato cultivation greenhouse and performs location mapping as well as ripening stage estimation of every tomato fruit in the greenhouse. The *TomatOD* dataset contains images of tomato fruits and high-quality expert annotations from a group of two agriculturists. The dataset contains 277 images with 2418 annotated tomato fruit samples of unripe, semi-ripe and fully ripe class, making it suitable as a benchmark in detection and classification of tomatoes in greenhouses. The *TomatOD* dataset can be found here (https://github.com/up2metric/tomatOD). Moreover, it is mentioned that the annotations of the introduced dataset are provided in a COCO compatible format. Our dataset is summarized in Table 2. In this section the data acquisition in the greenhouse, and the (manual) annotation procedures as well as the statistical analysis of the dataset are described in detail.

IMAGES	**ANNOTATIONS**		
277	2418		
	unripe	**semi-ripe**	**fully ripe**
	1592	395	431

Table 2. Images and annotations of tomatOD dataset

3.1 Data acquisition

Images of tomato plants containing unripe, semi-ripe and fully ripe tomatoes were collected from a soilless cultivation greenhouse in Crete, Greece. The data acquisition process took place in April 2019 and lasted 3 days. The camera that was used is a CMOS 12MP Ximea Machine Vision sensor. We selected this sensor to better simulate views of a camera mounted on top of a robotic arm, that navigates through the corridors of the greenhouse in order to capture the location and ripening stage of each tomato in a greenhouse. The final images depict rows of crops of tomato plants. The greenhouse provides a controlled environment regarding the lighting conditions, thus most images share ambient lighting characteristics. All in all, the set of images contained in our dataset were captured in a fashion that simulates a specific use case scenario.

Figure 2. Proportions of classes of tomatOD dataset

3.2 Annotation Procedure

The image annotation procedure includes the localization of tomatoes fruits in the image and the identification of corresponding class of ripening stage. Since the correct annotations play an important role in the accurate training of detection algorithms, this procedure is of great importance. The annotation of tomato dataset was done manually by two specialized agriculturists with deep knowledge on the tomato growing procedure. The online tool that was used for data annotation is our in-house annotation tool, namely the "ANNOTATOR", powered by up2metric. The annotated data follow some specifications. At first, only the relatively big tomato fruits on the first and second rows on the branches of plants were annotated, given that they appeared clearly and they were not blurred. Next, only the fruits, that are not occluded more than 50% of their size were annotated. In the case of a bunch, the

tomatoes were annotated one by one independently. Finally, all the blurred or very small fruit objects in the first two rows were manually erased from images. The final set of annotations include a bounding box for the location of the fruit and a list of labels of the corresponding class of ripening stage provided by the agriculturist experts. Tomato fruits of the third and over row are ignored.

3.3 Statistical Analysis

The statistical analysis of the dataset is very crucial, since it gives valuable insights about the data, which can lead in a deeper understanding and thus the design of better "task-specific" detection algorithms. Firstly, the appearance frequency of each category of *TomatOD* dataset is presented. In Figure 2 the three classes of the *TomatOD* as well as their relative appearance frequencies are shown. The class unripe is the most frequent with 1592 instances, the fully-ripe is the second among the three classes with 431 instances and the semi-ripe with 395 instances is the last frequent one. The classes of the *TomatOD* dataset are clearly not balanced, however their relative proportion is in line with the actual appearance frequency of each class in a realistic scenario. Another major point of interest, and a huge differentiating factor for our dataset is the size distribution of bounding boxes. Thus, the percentile relative size of each bounding box is calculated, that indicates the proportion of the diagonal length of each box over the diagonal length of the image. In Figure 3, the histogram of the percentile relative size distribution of the *TomatOD* bounding boxes is presented. The graph of Figure 3 is skewed right. Most of the bounding boxes have size between 3% to 15% relative to the image size. Additionally, the graph shows that the maximum bounding box in size is only 23% relative to the whole image size, thus the bounding boxes of tomato fruits cover a small area on images in general, adding extra difficulty to the OD task.

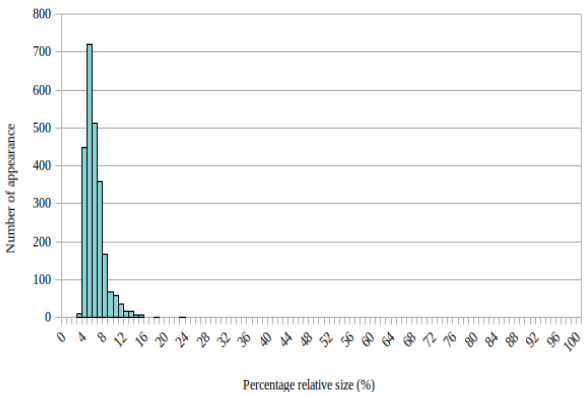

Figure 3. Histogram of relative size distribution (%) of bounding boxes in the *TomatOD* dataset

Moreover, the number of labeled instances per image is computed for the *TomatOD* dataset. As it is shown in Figure 4, only 1% of images have one category per image and 11% of images include 8 instances, while the maximum number of instances per image, which is 20, is found only in 0.72% of the images. Finally, the *TomatOD* dataset has an average of 8.7 instances per image. Figure 5 presents some further insight regarding the total number of categories found in each image. This type of information describes the variety of the object in each image. It is observed, that more than half of the *TomatOD* images contain objects of all 3 categories, while less than 8% of the images have objects of a single category.

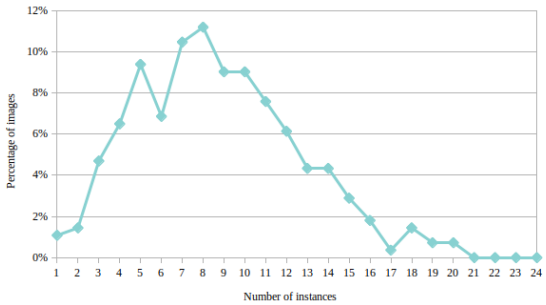

Figure 4. Histogram of the number of annotation instances per image

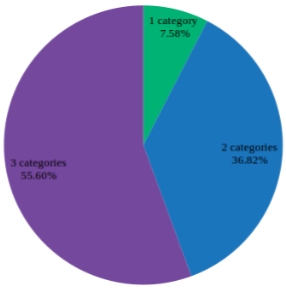

Figure 5. Distribution of the number of categories appeared in *TomatOD* images

Finally, according to established protocols of machine learning methods, we propose an official train-test split for the *TomatOD* to ensure that any algorithms evaluated on our dataset can present easily comparable results in a straightforward manner. We adopted a 80%/20% train-test split ratio, which in our estimation represents a fine trade-off between the amount of training data and the adequacy of test data to provide a representative subset of a realistic use case scenario. Please, note that the selection of the training and test data respectively was conducted in a semi-random manner; we exploited our statistical analysis in order to constraint our random selection algorithm to come up with a train-test split that maintains about the same split ratio both on the image level (19.9% test size) and the total annotations level (19.3% test size).

3.4 Comparison to other Datasets

TomatOD dataset differentiates over common generic object detection datasets in many ways. At first, due to its task specific nature the total number of classes contained in our dataset is significantly smaller. However, our set of classes pose a much harder problem due to the high correlation between our categories, i.e. all classes refer to the same tomato fruit, in different ripening stages. Another differentiating factor suggest the constrained imaging conditions of our data; in contrast to generic OD, our goal is to simulate a specific scenario. So, our data derive from the same machine vision camera system, maintain a similar camera pose while external conditions such as lighting or the greenhouse context also remain constant. Such a setup allows OD algorithms to be able to train with less data compared to generic OD datasets, like COCO, due to the constrained nature of our task. Last, while most generic OD datasets rely on rough annotations, or even crowdsourced ones, in combination with massive data to provide a useful dataset, we adopt a radically different approach; we invest in rich high-quality annotations to overcome the need for more data. In

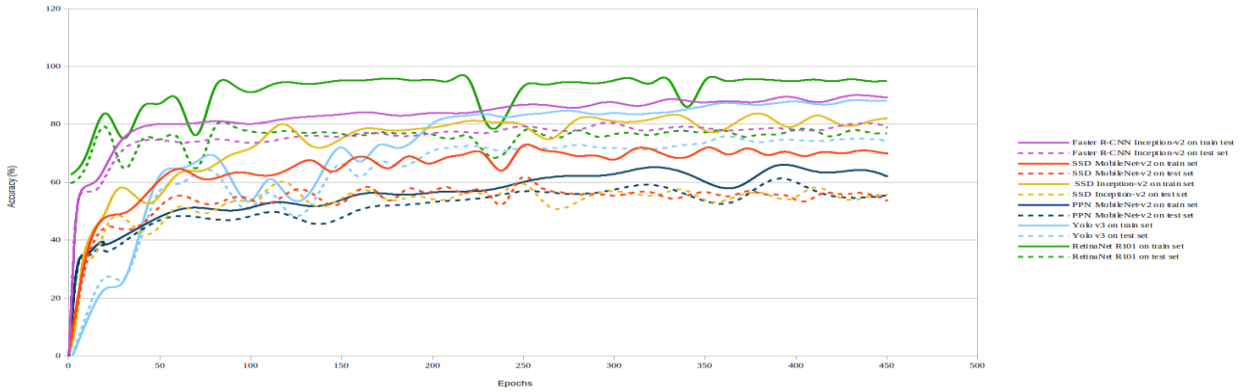

Figure 6. Accuracy vs Epochs diagram for the OD algorithms on train and test sets.

addition, the high specificity and correlation of our classes set requires expert knowledge from a group of agriculturists to provide annotations of the highest quality.

Compared to other specialized datasets, designed for classification and detection tasks for Precision Agriculture similar to ours, the *TomatOD* dataset contains less images and annotations. However, the data annotations were done by expert agriculturists; thus, the annotations are more precise. Additionally, the *TomatOD* data describes a real-world use case scenario involving tomato plants in a greenhouse, while the images of the dataset resemble those captured from a camera mounted on a robotic arm. This kind of specialized set-up is useful in a wide range of application in a greenhouse. The size distribution of annotations in our dataset significantly differs to that of other specialized datasets; in fact, the *TomatOD* dataset contains much smaller objects, so the detection of such small bounding boxes poses a challenging task. Moreover, the *TomatOD* is a freely publicly available dataset for the detection and classification of tomato fruits in unripe, semi-ripe and fully ripe classes from images.

4. OBJECT DETECTION METHODS

Over the last years, more and more powerful deep learning models for object detection are built. Those models vary on their architecture, but also differ on their training procedure, the optimization loss function, etc. Roughly, modern object detection networks can be divided in two main categories, the one-stage and the two-stage detectors. The one-stage detectors can achieve high inference speed, while the two-stage detectors can detect objects with high localization and object recognition accuracy.

The difference between those two categories of networks is mostly in their architectures and the way they predict the bounding boxes. Specifically, an architecture of a two-stage detector typically contains a region proposal step, which proposes candidate object bounding boxes. Afterwards, features are extracted from each candidate bounding box and classification and bounding-box regression tasks are performed. Such approaches include R-CNN (Girshick et al., 2014), Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2015) and Mask R-CNN (He et al., 2017). Next, the most common of them, the Faster RCNN detector, is presented.

Faster-RCNN (Ren et al., 2015) is a region-based CNN detector. It is an enhanced version of the former Fast-RCNN (Girshick, 2015) and the original R-CNN (Girshick, 2015). This detector consists of two networks, the first one is a region proposal network (RPN), which is a fully convolutional network, for the generation of region proposals in a wide range of scales and aspect ratios (anchors) and the second one is a detection CNN, which uses those proposals for classification and localization. Additionally, anchors of various scales and ratios are used, making the network able to detect objects of different sizes. Due to the fact that RPN shares some convolutional layers with the detection network, the overall detection procedure is further accelerated.

On the other hand, one-stage detectors predict boxes and class probabilities directly from image pixels, without any region proposal step, for this reason they can achieve much faster inference time and therefore are able to be used for real-time applications and on mobile devices. Such one-stage detectors include MultiBox (Erhan et al., 2014), SSD (Liu et al., 2016), YOLO (Redmon et al., 2016), YOLOv2 (Redmon, Farhadi, 2017), YOLOv3 (Redmon, Farhadi, 2018), RetinaNet (Lin et al., 2017) and Pooling Pyramid Network (Jin et al., 2018). A presentation of the most common and novel ones follows.

SSD (Single Shot MultiBox Detector) (Liu et al., 2016) is a one-stage detector that predicts multiple category scores and box locations per inference step. The predictions are done for a fixed set of default bounding boxes, which have different aspect ratios and scales at each location for several feature maps. The SSD network predicts a set of offsets for those default boxes as well as their confidence scores using an aggregation step of some specific feature maps at the end of the network. The VGG16 is used as a common backbone architecture for SSD network, however variants with other backbones such as the Inception v2 or some MobileNets do exist. SSD is able to handle objects with various sizes, by combining predictions from feature maps with different resolutions. Under typical configurations, the SSD detector cannot perform equally well in detecting objects in smaller scales.

YOLOv3 (Redmon, Farhadi, 2018) is an one-stage object detector and an improved version of YOLO (Redmon et al., 2016) and YOLOv2 (Redmon, Farhadi, 2017). The contribution of this algorithm is the real-time detection of objects on images. The YOLO series of algorithms have a unified architecture which can determine the location and the class of objects with a single neural network. As an improvement to the previous networks, YOLOv3 performs multilabel classification for object detections in images. Moreover, the newer version of YOLO makes detections at three different scales. Due to the benefits of multi-scale predictions, YOLOv3 has better performance for small objects but worse for medium and large objects.

	unripe AP	*semi-ripe AP*	*fully-ripe AP*	*mAP*
Faster-RCNN	89.29%	43.8%	66.9%	66.66%
SSD MNetv2	62.07%	44.68%	49.26%	52%
SSD Inc.v2	67.99%	30.39%	45.69%	47.02%
PPN	66.69%	48.04%	68.57%	61.1%
Yolo v3	85.09%	49.11%	57.57%	63.92%
RetinaNet	91.47%	55.28%	76.77%	74.51%

Table 3. Average precision of each class and mean Average Precision over all classes

RetinaNet (Lin et al., 2017) is another one-stage object detector. During the training of one-stage detectors, there is an extreme foreground-background class imbalance problem, which is not happening in the case of two-stage detectors. In order to overcome this issue, RetinaNet uses the focal loss. This loss function is formed to down-weight the easy examples and thus focus on the hard training examples, which improve the overall prediction accuracy. As a result, RetinaNet achieves high inference speed, as the other one-stage detectors, but also more accurate results. As an extra advantage, RetinaNet improves the precision for detecting small and medium objects over a typical SSD.

PPN (Pooling Pyramid Network) (Jin et al., 2018) is also a one-stage object detector, similar to SSD (Liu et al., 2016) with two simple changes. Firstly, the box predictor of PPN is shared for feature maps at different scales, instead of using independent box predictors of SSD. Secondly, the convolutions between feature maps of SSD are replaced by max pooling operations. The PPN model manages to reduce the model size but also to maintain accuracy similar to that of an SSD.

5. EXPERIMENTS

In order to set benchmark results, six state-of-the-art detectors are evaluated at the proposed *TomatOD* dataset. In detail, Faster RCNN with Inception v2, SSD with both Inception v2 and Mobilenet v2, PPN with Inception v2. RetinaNet (ResNet 101) and Yolo v3 are chosen, all of them pretrained on COCO dataset. In this section, the experimental set-up is described and the results of different detectors at *TomatOD* dataset, after hyperparameter fine-tuning, are listed.

5.1 Experimental setup

The training of the Faster RCNN, SSD and PPN was performed using the Tensorflow Object Detection API framework (Hung et al., 2017), the Detectron2 API was used for the training of RetinaNet while the Darknet framework (Redmon, Farhadi, 2018) was used for the training of YoloV3. All models were trained and evaluated on a system with a modern Intel Core i7 CPU paired with a GTX 1080ti GPU and 64 GB RAM.
The detectors were trained on the train set for 450 epochs and were evaluated on the test set; the train/test split is described in Section 3.3. Furthermore, hyperparameter fine-tuning was performed for all of the networks in order to perform optimally on the *TomatOD* dataset. The Faster RCNN, SSD and PPN were trained on *TomatOD* images with fixed image size of 1000x1000 pixels, while the Yolo v3 on resized images of size 608x608 pixels and RetinaNet on the original images.
All methods were trained using an initial learning rate of 2e-4 with decay using the RMS Prop with momentum optimizer. Data augmentations techniques were applied to obtain better results. Such techniques included random horizontal flip, random adjust of brightness or contrast and random crop. For the SSD and PPN we adjusted their target scale range to [8%, 60%] to better fit the size distribution of the objects of our training data.

5.2 Results

In Figure 6 accuracy over epochs is illustrated for both the train and the test set for every model trained. Both train and test curves were used to select the optimal epoch for each model. RetinaNet yielded the best overall results with a mere 74.51% for the mAP metric, as shown in table 3. Regarding per class performance, the RetinaNet achieved the best performance for every single class. Among the rest detectors, the Faster RCNN yielded good results, as well as YoLo v3 and PPN. The SSD detectors, albeit very fast, performed the worst with a mAP of 47% and 52% for the Inception v2 and the MobileNet v2 backbone selection respectively.

Figure 7. Detections of RetinaNet on two test images. The green outline indicates predictions and the yellow outline the ground truth. TP: green outline & fill; FP: green outline – red fill; FN: yellow ouline – blue fill

The precision-recall curve for each model in figure 8 confirm the results stated above. Given that in a pr-curve the bigger the area under curve the better, we observe that RetinaNet, Faster R-CNN and Yolo v3 yielded the best overall results. Furthermore, these models maintained good performance for various values of the IoU threshold. On the other hand, the two SSDs and the PPN models didn't perform well, especially for class "semi-ripe" and "fully-ripe". We observe that there is a significant tradeoff

between inference time per image and accuracy; as shown in table TT2, the SSD-based detectors need about 0.02 seconds per image, or about 50fps. Such performance suggests that these models can be deployed in real time systems. The other models however work on about 10fps for the Faster R-CNN and the RetinaNet. In our tests, YoLo v3 was the best tradeoff between accuracy and speed, however its overall accuracy might not be enough for deployment in production.

The behaviour of these OD algorithms on our dataset can also give us some valuable insight about the challenges of the specific use-case scenario supported by our dataset. The most serious challenge is undoubtedly the small-favouring size distribution of the objects in our dataset. Furthermore, we observe many mis-classification cases were the two "extreme" classes (fully ripe and unripe) get confused with the "middle one" (semi- ripe) and vice versa. This difficulty in class discrimination confirms our estimate that such kind of information needs expert knowledge to ensure correctness in our ground truth data.

6. CONCLUSION

This work represents the first version of our dataset, namely *TomatOD*. We plan to actively enrich our dataset in the following months to contain even more data from a diversity of soilless cultivation greenhouses. Incorporating more greenhouse environments would allow models trained in our dataset to perform adequately well in a plug and play fashion facilitating their integration to operational systems. We also plan to augment our annotations by including even more experts, in order to be able to perform a statistically-sound comparison of their annotations.

Detector	***Inference time per image***
Faster-RCNN	0.103secs
SSD MNetv2	0.021 secs
SSD Inc.v2	0.018 secs
PPN	0.017 secs
Yolo v3	0.048 secs
RetinaNet	0.105 secs

Table 4. Inference time of each OD algorithms

Current state of the art detectors fail to be characterized as production-ready in a use-case scenario similar to that of our dataset. While some two-stage detectors approach near-acceptable detection accuracy level, they severely lack in real-time inference capacity. On the other hand, "fast" single shot detectors don't cope well with the singular object size distribution of *TomatOD* resulting in poor detection accuracy. Given the evaluation results of common state of the art OD algorithms in our dataset, it is clear that the main challenge is no other than the small size in which objects mostly appear in our

Figure 8. Precision-Recall curves for each OD algorithm and class: fully ripe (red), semi-ripe (orange), unripe (green)

data. To overcome this issue, OD research should focus on aligning the target scales of the models to better fit the overall object size distribution of our dataset in order to develop models that operate on an acceptable accuracy level while maintaining a near real-time inference performance.

ACKNOWLEDGEMENTS

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code T1EDK-04171).

REFERENCES

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009, June. Imagenet: A large-scale hierarchical image database, in: Conference on computer vision and pattern recognition, IEEE, 248-255.

Erhan, D., Szegedy, C., Toshev, A. and Anguelov, D., 2014. Scalable object detection using deep neural networks, in:Conference on computer vision and pattern recognition,IEEE, 2147-2154.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J. and Zisserman, A., 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2), 303-338.

Ferentinos, K.P., 2018. Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, *145*,311-318.

Geiger, A., Lenz, P. and Urtasun, R., 2012, June. Are we ready for autonomous driving? the kitti vision benchmark suite, in: Conference on Computer Vision and Pattern Recognition, IEEE, 33354-3361.

Girshick, R., 2015. Fast r-cnn., in: Conference on computer vision, IEEE,1440-1448.

Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation, in Conference on computer vision and pattern recognition, *IEEE*, 580-587.

Grinblat, G.L., Uzal, L.C., Larese, M.G. and Granitto, P.M., 2016. Deep learning for plant identification using vein morphological patterns. *Computers and Electronics in Agriculture*, *127*,418-424.

Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R. and Schmid, C., 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions, in: Conference on Computer Vision and Pattern Recognition, IEEE, 6047-6056.

He, K., Gkioxari, G., Dollár, P. and Girshick, R., 2017. Mask r-cnn, in: Conference on computer vision, IEEE,pp. 2961-2969.

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S. and Murphy, K., 2017. Speed/accuracy trade-offs for modern convolutional object detectors, in: Conference on computer vision and pattern recognition, IEEE, 7310-7311.

Jin, P., Rathod, V. and Zhu, X., 2018. Pooling pyramid network for object detection. *arXiv preprint arXiv:1807.03284*.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T. and Ferrari, V., 2018. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*.

Lin, T.Y., Goyal, P., Girshick, R., He, K. and Dollár, P., 2017. Focal loss for dense object detection, in: International conference on computer vision, IEEE, 2980-2988.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., 2014, September. Microsoft coco: Common objects in context, in: *European conference on computer vision*,pp. 740-755 Springer, Cham.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y. and Berg, A.C., 2016, October. Ssd: Single shot multibox detector, in: European conference on computer vision, 21-37 Springer, Cham.

Pantazi, X.E., Moshou, D., Alexandridis, T., Whetton, R.L. and Mouazen, A.M., 2016. Wheat yield prediction using machine learning and advanced sensing techniques. Computers and Electronics in Agriculture, 121,57-65.

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A., 2016. You only look once: Unified, real-time object detection, in: Conference on computer vision and pattern recognition, 779-788.

Redmon, J. and Farhadi, A., 2017. YOLO9000: better, faster, stronger, in: Conference on computer vision and pattern recognition, 7263-7271.

Redmon, J. and Farhadi, A., 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks, in Advances in neural information processing systems, 91-99.

Sun, J., He, X., Ge, X., Wu, X., Shen, J. and Song, Y., 2018. Detection of key organs in tomato based on deep migration learning in a complex background. Agriculture, 8(12),196.

Santos, T.T., de Souza, L.L., dos Santos, A.A. and Avila, S., 2020. Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association. Computers and Electronics in Agriculture, 170, 105247.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P. and Belongie, S., 2018. The inaturalist species classification and detection dataset, in: Conference on computer vision and pattern recognition, IEEE, 8769-8778.

Zheng, Y.Y., Kong, J.L., Jin, X.B., Wang, X.Y., Su, T.L. and Zuo, M., 2019. CropDeep: the crop vision dataset for deep-learning-based classification and detection in precision agriculture. Sensors, 19(5), 1058.