IT in Industry, Vol. 9, No.1, 2021

Published Online 1-03-2021

COMPARATIVE ANALYSIS OF DEEP LEARNING ARCHITECTURES FOR GRAPE CLUSTER INSTANCE SEGMENTATION

Ms. Dhanashree Barbole[1], Dr. Parul Jadhav[2]

School of Electronics and Communication Engineering

Dr. Vishwanath Karad MIT World Peace University, Pune, India

barbole.dhanshree@gmail.com[1], parul.jadhav@mitwpu.edu.com[2]

Abstract

The grape cluster identification and its segmentation for the sake of total weight prediction task of wine yard shows the need of segmentation atomization with better accuracy. The challenge of grape cluster segmentation is considered to provide solution using deep neural network models such as YOLO v3, Mask RCNN, U-net. This paper contributes in terms of the modified U-net model for the segmentation of grape clusters using training and testing strategy for the validation of the results. The results are obtained for the accuracy of the classification of pixels as part of grape cluster or outside of clusters and comparative results show improvement in segmentation using modified U-net. The accuracy, precision and recall analysis is performed and comparatively proposed model shows better results

Keywords: YOLO v3, Mask RCNN, U-net, Modified U-net, Performance, Segmentation of grape cluster.

1. Introduction

Sparseness in plant structures and randomness in environmental conditions are the main challenges while atomizing the agriculture sector. The sector needs processing for monitoring plants and fruits for better productivity [1]. The detection and localization are the important challenges in image processing based methods [2]. Several applications require accurate fruit and plant detections. Accuracy in agribusiness applications, representing the executives of entomb and intra-field inconstancy can be inferred if identification information is appropriately restricted in special

domain. Natural product identification can likewise be a primer advance for sickness, supplement

insufficiency checking [3] and a significant part on activation is the one that is processed using computerized methods [4]. Past homesteads, organic product identification can be utilized in field phenotyping, supporting plant research and preparing programs [5, 6].Of the rack RGB cameras and computer vision (CV) can give moderate and adaptable answers for natural product discovery. Cutting edge CV vision frameworks dependent on profound convolution neural network [7] can manage varieties in shape, illumination and enormous between class changeability [8], fundamental highlights required for hearty acknowledgment of complex features in outside conditions. Ongoing investigates [9,10] have indicated that the Faster R-CNN (area based convolution neural organization) design [11] can deliver precise outcomes for a huge arrangement of natural products, including peppers, melons, oranges, apples, mangoes, avocados, strawberries and almonds. In paper [12], fruits and tree trunks are detected using CNNs. These detected fruits and trunks are then tracked across images and directly converted 2D tracks of fruits and trunks into 3D landmarks to avoid double counting. These past research work frameworks recognize singular object by rectangular bounding boxes, as found in Figure 1. Such boxes, if well fitted to the organic product's boundary, could give assessments of its natural shape and space inhabitance for oranges and apples (round shape). Notwithstanding, for

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Published Online 1-03-2021

grape bunches, rectangular boxes would not appropriately conform to the berries. Above and beyond past article discovery is example division [13]: the organic product/non-organic product pixel order joined with case task. Case division can appropriately distinguish berries pixels in the recognition box, giving better organic product portrayal. Likewise, impediments by leaves, branches, trunks and considerably different bunches can be reduced by mechanical control and other mechanization assignments.

The grape cluster instance segmentation is implemented using modified U-NET model. The addition of layers in U-NET at the input stage using convolution layer and addition of up sampling layer at the end have shown improvements over base U-NET architecture. Also, results are compared with Mask-RCNN, YOLOv3.

Along with introduction in section I, Section II provides brief literature survey related to the topic. Section III focuses on proposed work which explains the proposed architecture and overview of comparative models. Results and analysis is provided in section four followed by conclusion

2. Review of literature

Prior works present issues that anticipate the points of interest and intensity of convolution neural network (CNNs). These organizations learn powerful portrayals for a given artificial intelligence (AI) task, supplanting highlight designing [14]. Efficiently, profound learning approaches are being embraced in fields introducing picture based perceptual issues, and rural applications are no exemption [15]. CNN's invariance to nearby interpretation gives vision frameworks vigor in circumstances where a component's essence is a higher priority than its careful area [16]. For instance, researcher revealed that a variety in the berry up-and-comer area influences their berry grouping [17]. CNNs are likewise ready to encode fluctuation with respect to posture, shading and light, if the data set information presents adequate instances of such variety, which eases the requirement for controlled imaging, luminance and camera settings. The primary endeavors utilized CNNs to perform pixel characterization, trailed by extra strides to section the singular natural products [18]. Further, these prior methodologies were replaced by start to finish object location dependent on the mainstream Faster R-CNN technique. Few researchers [10] utilized VGG16 network pre-prepared by Image-Net (VGG16) is the perceptual spine in the Faster R-CNN design). In [9], creators likewise utilized the Faster R-CNN engineering for natural product location. The work focuses on creating the datasets from pictures caught in plantations by an automated ground vehicle for apples and mangoes, and a dataset for almonds, additionally in plantations, yet utilizing a hand-held DSLR camera (advanced single-focal point reflex).

Faster R-CNN is genuinely perceived as an effective design for object location, yet it isn't the main meta-engineering [19] ready to arrive at cutting edge results. Another gathering of structures is the single shot finder (SSD) meta-design, single feed-forward convolution networks ready to anticipate classes and bouncing boxes in a solitary stage. The YOLO (You Only Look Once) networks, proposed by [20], are instances of the SSD family. Grape group present bigger fluctuation on size, shape and smallness contrasted with different organic products like peppers, apples or mangoes. An attention on berry identification, for example, can be viewed as an approach to dodge grape group fluctuation, performing yield forecast over berry checking, thus bypassing the grape bunch division issue [17]. CNNs can learn portrayals of complex visual examples, so are an intriguing option for grape bunch discovery. In any case, object recognition utilizing jumping boxes could be deficient for yield forecast applications, considering the tremendous inconstancy in grape groups' shapes and conservativeness. Then again, semantic division (the grouping of pixels as natural product or foundation) could likewise be lacking, thinking about the serious impediment between organic products seen in plantations. The joined undertaking of item discovery (where are the grape bunches?) and pixel grouping (this pixel has a place with which bunch?), is an elective AI task plan for yield forecast and computerized gathering applications. Cover R-CNN [6] is a deduction of Faster R-CNN which is ready to perform occasion division, mutually upgrading district

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Published Online 1-03-2021

proposition, bouncing box relapse and semantic pixel division.

In light of the possibility of fully convolution network (FCN) de-convolution to reestablish picture size and highlight, U-Net develops the encoder-decoder structure in the field of semantic division. The encoder steadily lessens the spatial measurement by ceaselessly blending the layers to remove included data, and the decoder parcel progressively reestablishes the objective detail and the spatial measurement as indicated by the component data. Among them, the progression of the encoder bit by bit diminishing the picture size is called down inspecting, and the progression of the decoder bit by bit lessening the picture subtleties and size is called up testing. Combination activity of the immediate expansion highlight when the FCN is up tested, the U-Net up-examining measure firstly uses the link activity to joining the element maps before the up-inspecting of the encoder and the down testing of the decoder. After linking, the element map is de-convolved. Not like the traditional convolution, pooling, and different activities, this system of straightforwardly using shallow highlights is called skip association. U-Net receives the skip association system of grafting to utilize the highlights of the down examining part of the encoder which is to be utilized for up testing. To accomplish a more refined result, this methodology is applied to shallow element data.

With the study of various methods of literature it can be understood that, grape cluster instance segmentation is the challenging task and hence there is need of accurate instance segmentation model

3. Proposed Work

Scenario of proposed work is shown in figure 1. The proposed methodology introduces a new public dataset for image-based grape detection, including a novel method for interactive mask annotation for instance segmentation.

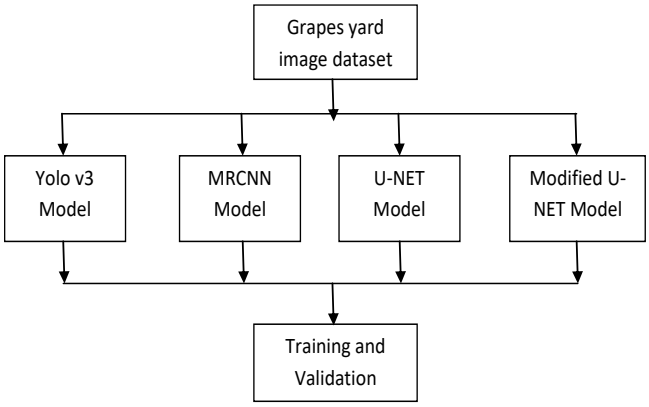


Figure 1: Proposed work scenario for segmentation using deep neural networks

Three neural networks are trained and evaluated for fruit detection: Mask R-CNN and YOLOv3, U-net and modified U-net evaluation measures the semantic segmentation, object detection, and instance segmentation. Instance segmentation requires,

- Object detection of all objects in an image: Here the goal is to classify individual objects and localize each object instance using a bounding box
- Segmenting each instance: Here the goal is to classify each pixel into a fixed set of categories without differentiating object instances

The segmentation task presented in this paper is to identify the grape cluster from the input image.

Dataset details:

The Embrapa Wine Grape Instance Segmentation Dataset (WGISD) [21] is composed by 300 RGB images showing 4,432 grape clusters from different grape varieties.

YOLO v3 model:

“You Only Look Once” (YOLO) model is single end to end model that have shown considerable performance for object detection in images. The model used in this paper is composed of multiple convolution layers. The approach involves a single deep convolution neural network (originally a version of Google-Net, later updated and called Dark-Net based on VGG) that splits the input into a grid of cells and each cell directly predicts a bounding box

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Published Online 1-03-2021

and object classification. The result is a large number of candidate bounding boxes that are consolidated into a final prediction by a post-processing step.

There are three main variations of the approach, at the time of writing; they are YOLOv1, YOLOv2, and YOLOv3. The first version proposed the general architecture, whereas the second version refined the design and made use of predefined anchor boxes to improve bounding box proposal, and version three further refined the model architecture and training process.

Mask R-CNN:

Mask R-CNN has an additional branch for predicting segmentation masks on each Region of Interest (ROI) (Grape cluster ROI) in a pixel-to pixel manner Mask R-CNN model is divided into two parts:

- Region proposal network (RPN) to propose candidate object bounding boxes.
- Binary mask classifier to generate mask for every class.

Processing of image using Mask R-CNN:

Image is run through the CNN to generate the feature maps.

Region Proposal Network (RPN) uses a CNN to generate the multiple region of interest (ROI) using a lightweight binary classifier. It does this task using 9 anchors boxes over the image. The classifier returns object/no-object scores. Non Max suppression is applied to Anchors with high object existence score.

The ROI Align network outputs multiple bounding boxes rather than a single definite one and warps them into a fixed dimension.

Warped features are then fed into fully connected layers to make classification using 'Soft-max' and boundary box prediction is further refined using the regression model.

Warped features are also fed into Mask classifier, which consists of two CNN's to output a binary mask for each ROI. Mask Classifier allows the network to generate masks for every class without competition among classes.

Mask R-CNN uses anchor boxes to detect multiple objects, objects of different scales, and overlapping objects in an image. This improves the speed and efficiency for object detection. Anchor boxes are a set of predefined bounding boxes of a certain height and width. These boxes are defined to capture the scale and aspect ratio of specific object classes you want to detect.

U-net Model:

The goal of semantic image segmentation is to label each pixel of an image with a corresponding class of what is being represented. Because of prediction for every pixel in the image, this task is commonly referred to as dense prediction. Unlike the previous tasks, the expected output in semantic segmentation is not just labels and bounding box parameters. The output itself is a high resolution image (typically of the same size as input image) in which each pixel is classified to a particular class. Thus it is a pixel level image classification. The architecture of original U-net is shown in figure 2:

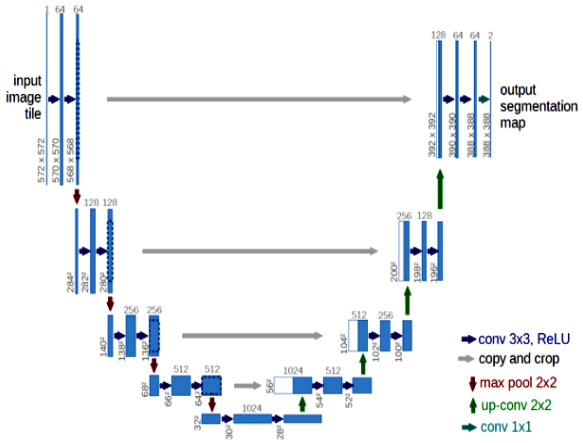


Figure 2: Original U-net Architecture

The U-NET model consists of 9 layers with combination of Conv2D, Max-pool and Up-Sampling layers. These layers are responsible to provide segmentation of ROI objects with respect to training dataset.

Modified U-net:

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Published Online 1-03-2021

Modified U-net architecture is constructed using base U-net Architecture in which 2 more layers are added. The architecture is shown in figure 3. First layer of up sampling is used to increase the image dimensionality at the input stage itself. The output is better resolution instance segmented image from grape cluster input image in which grape clusters are segmented.

The important aspect of designing the model to be considered here is that, the grape cluster has polygonal and nonlinear variations in shape and hence instance segmentation is the challenging task which requires more training in terms of epochs and at the same time the predefined accurate labeled region.

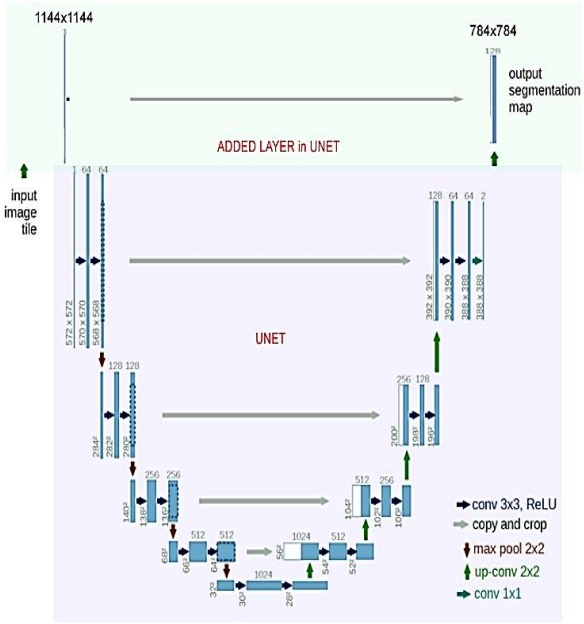


Figure 3: Modified U-net Architecture

4. Results and analysis

The proposed model is used to obtain instance segmentation of grape clusters with experimentation using python based scripting. Table 1 provides brief overview of the platforms used for the experimentation.

Table 1: Experimentation configuration and platforms

Scripting	**Python**
Neural network platforms	Tensor-flow, Keras
Supporting tools	Scikit-learn, pandas, numpy
Machine configuration	Intel 5th gen, 64-bit CPU, 4GB RAM
CNN models	YOLOv3, Mask R-CNN, U-net, Modified U-net
Dataset Images	1000+ images with labeled ground truth
Experimentation parameters	80% images from dataset with labeled data, Number of epochs=100 Steps per epoch = 10 Predefined weights=MSCOCO Validation= 20% images from dataset Optimizer= 'ADAM' Performance metric= 'Loss rate', 'accuracy'

The experimentation is performed extensively for each type of model from which average results are obtained for loss rate analysis. Figure 4 shows the comparison of loss rate analysis for all the four models. Figure 5 show the test input image and its ground truth mask. Figure 6 shows the segmentation region predicted from trained model for the test image input for different models.

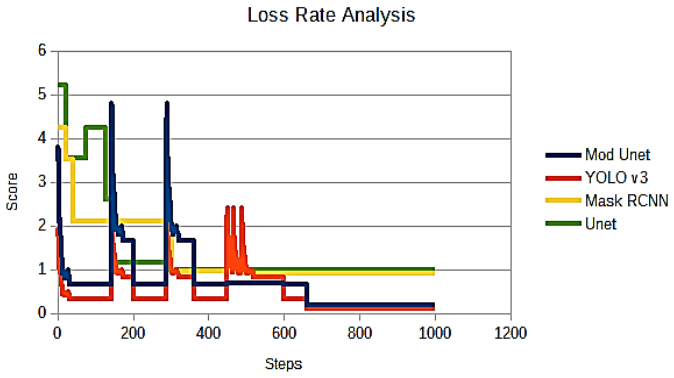


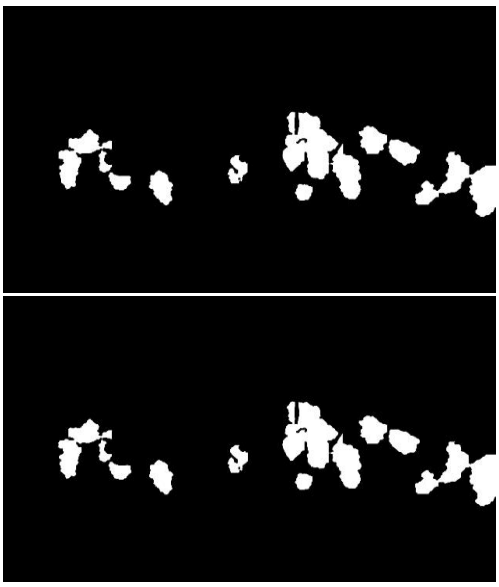
Figure 4: Loss rate analysis

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

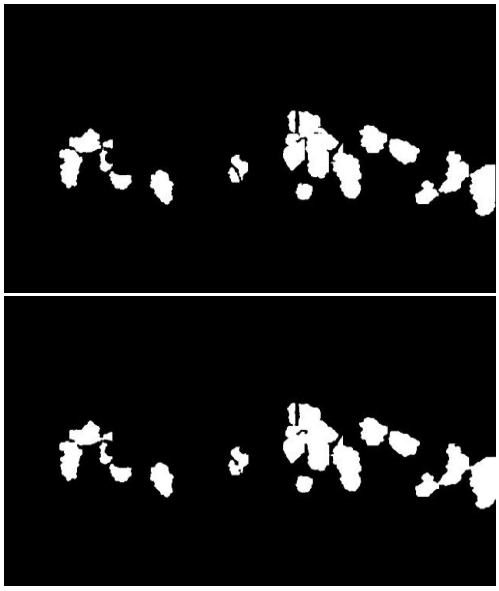
Published Online 1-03-2021

(a) **(b)**

Figure 5: (a) Test input Image (b) Ground truth region of grape clusters



(a) **(b)**



(c)

(d)

Figure 6: Results of grape clusters predicted using (a) Yolov3
(b) Mask RCNN (c) U-net (d) Modified U-net

Performance evaluation:

The performance is evaluated for the segmented true regions using region properties for the total true area pixels from entire image for the regions of grape clusters. The obtained average results of the input image from figure 5 are shown in table 1.

Table 1: Details of true region area pixels and error

Model	Area of true regions predicted	Error
YOLOv3	323850	-0.3006
Mask R-CNN	415650	-0.1024
U-net	415650	-0.1022
Modified U-net	515355	0.825
Ground Truth	463080	-

The accuracy analysis for each pixel in terms of confusion matrix analysis is performed from which every pixel is nominated as shown in table 2.

Table 2: Nominations of pixel for confusion matrix

True Positive (TP)	if pixel is actually from grape cluster region and detected as part of true region
True Negative (TN)	if pixel is actually from grape cluster region and detected as outside of true region
False Positive (FP)	if pixel is actually Not from grape cluster region and detected as part of true region
False Negative (FN)	if pixel is actually Not from grape cluster region and detected as outside of true region

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Published Online 1-03-2021

Based on these values of count of TP, TN, FP and FN the accuracy, precision and recall can be estimated as,

$$Accuracy = \frac{(TP+FN)}{(TP+TN+FP+FN)} \quad \ldots\ldots (1)$$

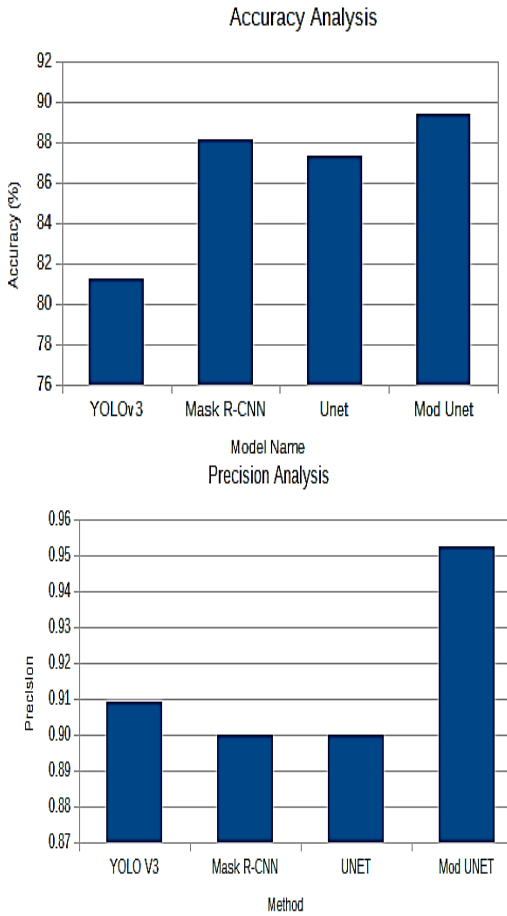
$$Precision = \frac{TP}{(TP+FP)} \quad \ldots\ldots (2)$$

$$Recall = \frac{TP}{(TP+FN)} \quad \ldots\ldots (3)$$

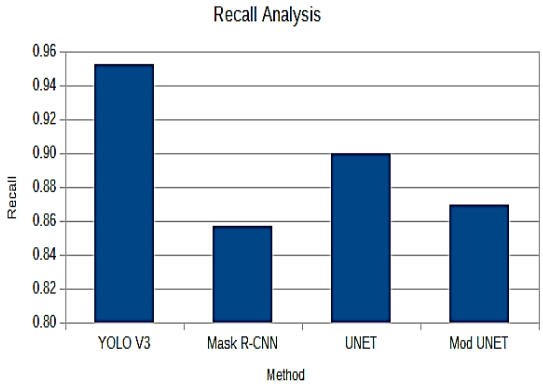
The experimentation performed on dataset obtained from [21] dataset is compared and shown the graphically in figure 7.

Table 3: Confusion Matrix

Method	TP	TN	FP	FN
YOLO V3	20	3	2	1
Mask R-CNN	18	1	2	3
U-net	18	1	2	2
Mod U-net	20	2	1	3



(a) **(b)**



(c)

Figure 7: Comparison charts of (a) Accuracy analysis (b) Precision analysis (c) Recall Analysis

The modified U-net shows better performance over other methods, which is almost 89% in accuracy, 95% of precision and 86% of recall.

5. Conclusion

The grape cluster segmentation is the first stage task while predicting the total weight of the clusters. The deep neural network based grape cluster identification is the instance segmentation task using neural network model. The YOLO v3, Mask RCNN, U-net models are considered for the segmentation suing training along with ground truth labeled masks. The trained model is then used to predict the result of cluster identification of the input image with the help of trained model. The performance in terms of the accuracy of proposed modified U-net model shows better results compared to other models and is almost 89% accurate for the prediction of the pixels belonging to the grape cluster region of interest.

Acknowledgement

This work was supported by department of science and technology (DST), India under grant given by central government of India under woman scientist (WOS) scheme B.

References

[1] Jayme Garcia Arnal Barbedo, Detection of nutrition deficiencies in plants using proximal images and machine learning: A review, Computers and Electronics in Agriculture,

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Volume 162, 2019, Pages 482-492, ISSN 0168-1699, https://doi.org/10.1016/j.compag.2019.04.035.

[2] Roser, M. (2019). Employment in agriculture. Our World in Data, Https://ourworldindata.org/employment-in-agriculture.

[3] Kicherer, A., Herzog, K., Bendel, N., Kluck, H.-C., Backhaus, A., Wieland, M., Rose, J. C., Klingbeil, L., L• abe, T., Hohl, C., Petry, W., Kuhlmann, H., Sei ert, U., & Topfer, R. (2017). Phenoliner: A new field phenotyping platform for grapevine research. Sensors, 17. Doi: 10.3390/s17071625.

[4] Rose, J., Kicherer, A., Wieland, M., Klingbeil, L., Topfer, R., & Kuhlmann, H. (2016). Towards Automated Large-Scale 3D Phenotyping of Vineyards under Field Conditions. Sensors, 16, 2136. Doi: 10.3390/s16122136.

[5] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, and 436{444. doi: 10.1038/nature14539.

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770{778). IEEE. doi:10.1109/CVPR.2016.90.

[7] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 25 (pp. 1097{1105).

[8] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In ICLR. URL: http://arxiv.org/abs/1409.1556.

[9] Bargoti, S., & Underwood, J. (2017a). Deep fruit detection in orchards. In 2017 IEEE International Conference on Robotics and Automation (ICRA) (pp. 3626{3633). IEEE. Doi:10.1109/ICRA.2017.7989417.

[10] SA, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T., & McCool, C. (2016). Deep-Fruits: A Fruit Detection System Using Deep Neural Networks. Sensors, 16, 1222. Doi: 10.3390/s16081222.

[11] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real Time Object Detection with Region Proposal Networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), Advances in Neural Information Processing Systems 28 (pp. 91{99). Curran Associates, Inc.

[12] Liu, X., Chen, S. W., Liu, C., Shivakumar, S. S., Das, J., Taylor, C. J., Underwood, J., & Kumar, V. (2019). Monocular Camera Based Fruit Counting and Mapping With Semantic Data Association. IEEE Robotics and Automation Letters, 4, 2296{2303. doi:10.1109/LRA.2019.2901987.

[13] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll_ar, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), Computer Vision {ECCV 2014 (pp. 740{755). Springer International Publishing.

[14] Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35, 1798{1828. doi:10.1109/TPAMI.2013.50.

[15] Kamilaris, A., & Prenafeta-Boldffu, F. X. (2018). Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 147, 70{90. doi:10.1016/J.COMPAG.2018.02.016.

[16] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. http://www.deeplearningbook.org.

[17] Nuske, S., Achar, S., Bates, T., Narasimhan, S., & Singh, S. (2011). Yield estimation in vineyards by visual grape detection. In IEEE International Conference on Intelligent Robots and Systems (pp. 2352{2358). doi:10.1109/IROS.2011.6048830.

[18] Chen, S. W., Shivakumar, S. S., Dcunha, S., Das, J., Okon, E., Qu, C., Taylor, C. J., & Kumar, V. (2017). Counting Apples and Oranges With Deep Learning: A Data-Driven Approach. IEEE Robotics and Automation Letters, 2, 781{788. doi:10.1109/LRA.2017.2651944.

[19] Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., & Murphy, K. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. In The IEEE

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Published Online 1-03-2021

Conference on Computer Vision and Pattern Recognition (CVPR).

[20] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788.

[21] Embrapa Wine Grape Instance Segmentation Dataset Link: https://github.com/thsant/wgisd

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731