Evaluation of YOLO Efficiency in Automatic Orange Detection in Multi-Exposure Images

Maurycio R. Oviedo Espinosa *, Letícia R. Porto, Vinicius S. W. Orlando, Antonio M. G. Tommaselli, Aluir P. Dal Poz, Nilton N. Imai

São Paulo State University (UNESP), Presidente Prudente, São Paulo, Brazil - (maurycio.oviedo, leticia.porto, vinicius.werneck, a.tommaselli, aluir.dal-poz, nilton.imai)@unesp.br

Key words: Fruit detection, Deep learning, Close-range, Agriculture, Computer Vision.

Abstract:

Brazil is the largest producer of oranges in the world and the automatic detection of fruits has been a challenging task in the context of remote sensing, due to variations in fruit appearance, changes in lighting and occlusions of foliage and neighboring fruits. In this sense, this paper focus on the detection of oranges in multispectral images, with different spectral bands and exposures, using a convolutional neural network (CNN) known as YOU ONLY LOOK ONCE (YOLO). The results indicate that, after 300 epochs, the model demonstrated an accuracy of 81.5% and an approximate recovery rate of 85%. Shutter speeds 1/640s and 1/250s are not suitable for detection due to low light and overexposure, respectively. Intermediate values may be more suitable for identifying a larger number of fruits.

1. Introduction

Brazil is the largest orange producer in the world and the leading exporter of concentrated orange juice. Most of the orange orchards are located in the São Paulo State and southwest of Minas Gerais State. In the 2022/2023 harvest, almost 12.8 million tons were produced in this region (Fundecitrus, 2023).

Automatic detection of fruits and vegetables in digital agriculture context is essential to estimate harvest and increasing productivity (Yamamoto et al., 2014; Bac et al., 2017). Automatic fruit detection in perennial crops such as apples and oranges is a challenging task because of variations in appearance due to illumination changes and occlusions from foliage and neighbouring fruits (Chen et al., 2017).

Terrestrial images of the citrus tree are affected by illumination problems caused by the complex environment of the tree canopy structure. The tree branches cause shadows in the tree canopy resulting in many dark areas in the terrestrial images. Variations in exposure, either by changing the shutter speed or the aperture, can improve the dynamic range and the image quality, enabling to find of a suitable illuminance scenario that allows detection of as many oranges as possible.

In this application field, object detection seeks to semantically locate and recognize objects in an image. Aiming at large datasets interpretation, convolutional neural networks (CNNs) emerged as attractive technology, which is notably known for their accuracy and speed (Vo, 2022). In order to address issues such as window overlap, the region proposal approach presents a promising solution by anticipating the potential location of the object (Zitnick, 2014). The success of AlexNet (Krizhevsky, 2012) in image classification highlights the strong ability of CNNs in feature extraction. Currently, the fastest R-CNN method leads in object detection, but its speed does not meet real-time requirements.

Among the most popular and well-known methods is YOU ONLY LOOK ONCE (YOLO) (Redmon, 2016). This approach adopts the concept of regression, where the input image is divided into multiple cells, and each cell predicts bounding boxes and class probabilities. YOLO transforms the detection challenge into a regression problem, resulting in extremely fast detection. This method can process an impressive 45 images per second. This work focuses on the detection of oranges in multispectral images, with different spectral bands and exposures.

2. Material and methods

The data to develop this study was acquired in September of 2022 in a citrus farm in the municipality of Matão, in the north of the São Paulo state. The farm produces citrus commercially and mostly sweet oranges. For the experiments, it was chosen an irrigated area with 5-years old orchards. The planted orange variety is Pera with rootstock Swingle.

An Agrowing model ALPHA 7RXXX Sextuple multispectral digital camera was employed for data acquisition. The camera’s sensor frame is divided into six parts (Figure 1) to acquire the same scene in 14 bands through six camera heads (lenses) capable of detecting radiation at specific wavelengths, as shown in Figure 1. Agrowing's multi-lens have a single mount and use a single CMOS sensor and a mechanical shutter (Tommaselli et al., 2020).

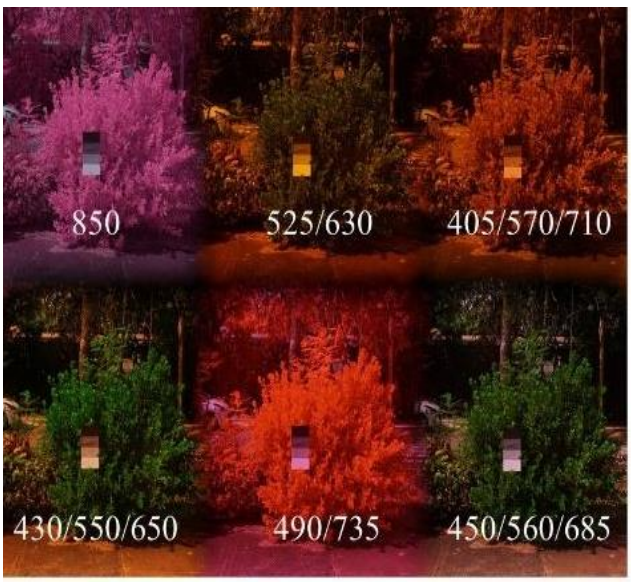

Figure 1. the wavelengths of each band in Agrowing camera.

For terrestrial acquisition, the multispectral camera was coupled to a tripod and positioned 1.70 m far from the citrus trees and at

a height of 1.49 m (Figure 2). The trees are approximately 3.5 m height; thus, the camera was positioned aiming at capturing the middle range of the canopy. Also, different shutter speeds were used in the camera configuration. The main aim was to obtain the best illumination conditions to ensure reliable orange detection. All images were processed by the manufactory software AgBasic and no radiometric calibrations were applied.

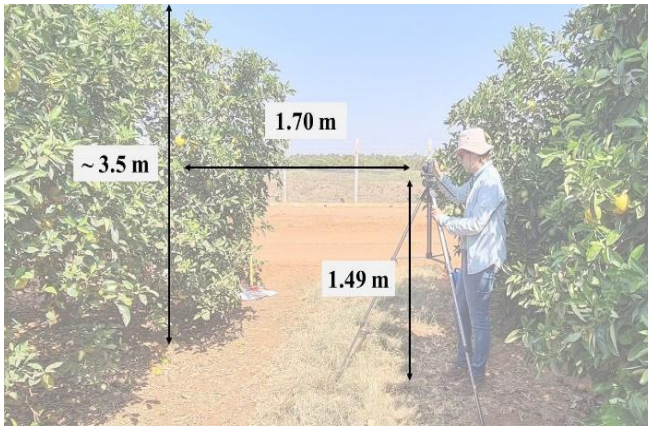

Figure 2. Distances used for image acquisition.

After the image processing, the data for training were annotated through bounding boxes with only an identification class “Orange”. The tool used for the annotation process was provided by Roboflow.IA (https://universe.roboflow.com/). This tool was used in the 60 images, of which 70% were selected for training, 20% for validation and 10% for testing. The model was prepared with Google Colab, which provides free high-performance GPUs, without any configuration. The YOLOv5 model was used for training the model for 300 epochs, spending **25 minutes and 12 seconds**. With this training step the weights of the neural network were trained, and after that, tested on new images.

3. Results and discussion

In machine learning and computer vision, metrics such as loss, play a crucial role in evaluating the performance of object detection algorithms. Both metrics were employed on the validation and training data, helping to quantify the quality of object detection in a model.
Analysing the loss function graph helps identify how well the model is fitting the data and making predictions. As shown in Figure 3.a and Figure 3.b, there is a sharp decrease in the loss function, indicating that the model is quickly learning to identify objects. As epochs progress, the decrease in loss becomes more gradual, suggesting that the model is approaching an optimal state.

During the analysis of the graph, unusual behavior can be observed, such as sudden spikes or oscillations. These fluctuations may indicate issues during training, such as excessively high learning rates or convergence problems. Other commonly used metrics in machine learning, such as recall, precision, and mAP at 0.5, were applied to this model. These metrics help measure the quality and effectiveness of object detection performed by a model.

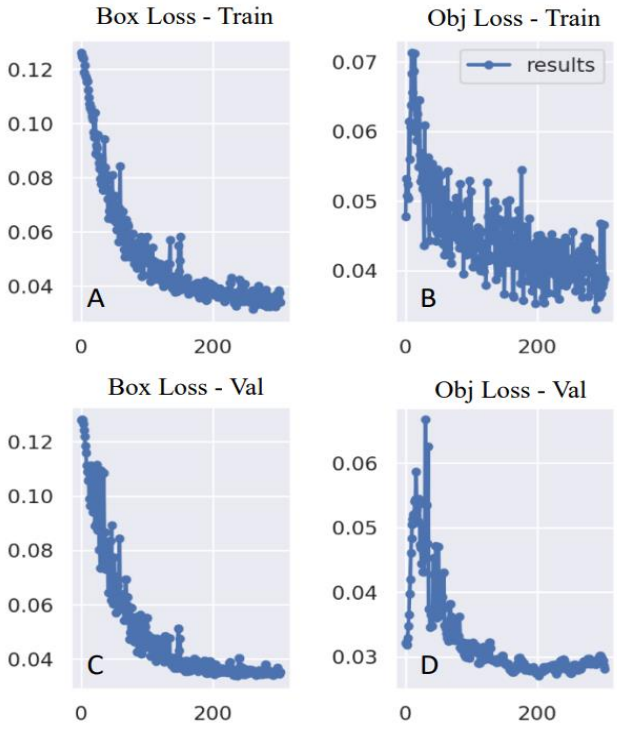

Figure 3. In this graph, the abscissa represents the epochs, and the ordinate represents the percentage ratio.The image presents four graphs, displaying performance statistics for the training and validation data.

Through the analysis and interpretation of a precision versus epochs graph, important information about the performance of the orange detection model during training can be gathered. In Figure 4.a, the variation of precision over the number of trained epochs is presented. On the y-axis, represents the percentage, which provides a direct measure of the precision rate. The x-axis represents the epochs, indicating the progression of the model's training. After 300 epochs, the model achieved a precision of 81.5%. The consistent growth and increasing trend in precision over epochs indicate that the model is progressively improving its performance, as shown in Figure 4.a.

Analysing and interpreting a recall versus epochs graph can provide valuable insights into the performance of the orange detection model. In Figure 4.b, the y-axis represents the percentage, providing a direct measure of the recall rate. A higher percentage indicates a higher recall, meaning that the model is effectively detecting objects correctly in the images. The x-axis represents the epochs, indicating the progression of the model's training. Each epoch represents a complete cycle of presenting the training data to the model. As epochs progress, the model can learn and adjust its parameters to improve its detection ability. A learning trend can be observed in the model as the epochs progress, with the recall reaching around 85% after 300 epochs (Figure 4.B).

After training, the mAP_0.5 is calculated to evaluate the YOLO model's ability to accurately detect oranges. The computation of the mean Average Precision (mAP) at the IoU threshold of 0.5 provides a comprehensive assessment of the model's effectiveness by considering both precision and recall, measuring the overlap between the predicted bounding boxes by the model and the manually annotated bounding boxes. An IoU (Intersection over Union) threshold of 0.5 is used to determine if

a detection is considered true or false. If the overlap between the predicted bounding box and the annotated bounding box is equal to or greater than 50%, the detection is considered correct.
The significant mAP_0.5 score achieved by the YOLO model trained for orange recognition, as depicted in Figure 4.C, substantiates its proficiency in making precise predictions. With a mAP_0.5 value of 91.6%, this model proves to be reliable in detecting oranges and holds promising prospects for practical implementations, including fruit classification tasks. It is worth noting, though, that the YOLO model's performance can fluctuate based on factors such as the dataset's size, quality, and diversity.

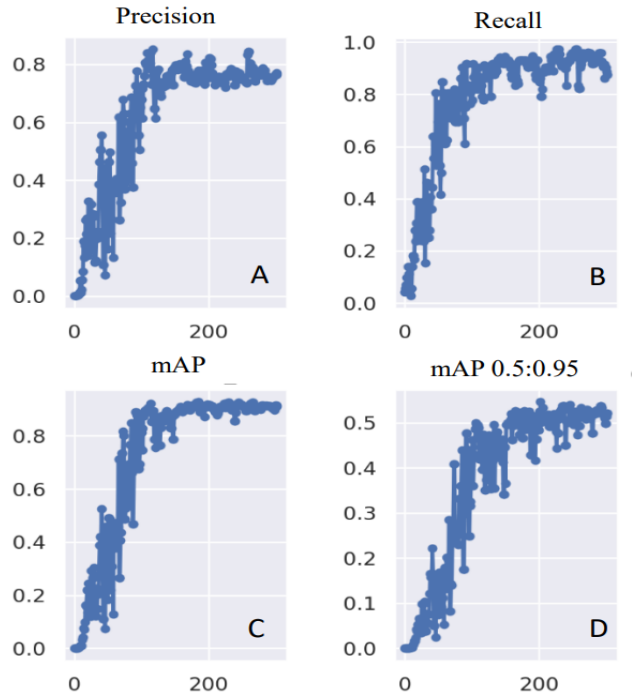

Figure 4. In this graph, the abscissa represents the epochs, and the ordinate represents the percentage ratio.

Other works have shown similar results regarding fruit detection using CNNs. An example is the work developed by Williams et al (2019), which presented a robot for collecting kiwis using multiple robotic arms.

In the study conducted by Yu et al., (2019). Convolutional Neural Networks (CNNs) were employed with the implementation of Mask R-CNN to enhance computer vision performance in strawberry harvesting robotics. The results obtained for fruit detection in more than 100 test images were notable, with an average detection precision rate of 95.78% and an equally remarkable recall rate of 95.41% (Yu, 2019).

To fulfil the scarcity of research on computer vision for date palm detection in orchard environments, Altheri et al. (2019) presented an innovative computer vision framework for robots involved in harvesting. The proposed approach involved the implementation of three classification models capable of real-time identification of different date characteristics, such as maturity, type, and determining the optimal harvest time. This methodology brought significant advancements in date detection, opening new possibilities for automation and optimization of the process (Altheri, 2019), achieving an impressive accuracy rate of 99.01%.

To individually identify fruits and obtain a pixel mask for each fruit in an image, Ganesh et al. (2019) developed a deep neural network approach named Deep Orange, based on a segmentation framework implemented with Mask R-CNN using ResNet-101. The initial results revealed that incorporating HSV data led to a significant improvement in precision, increasing from 0.8 to 0.9753.

Another widely used metric for performance evaluation is the F1-score, commonly employed in classification models, including convolutional neural networks. It considers both precision (the model's ability to correctly classify positive samples) and recall (the model's ability to identify all positive samples) to provide a unified performance measure.

Evaluating the ability to detect and classify objects in images is essential when training convolutional neural networks, with the F1-score being a crucial metric in this context. Achieving a high F1-score is crucial to ensure accurate and reliable object detection. After evaluating the model on the test images, an average precision of 81.5% and an average F1-score of 0.81 can be obtained, as shown in Figure 5.

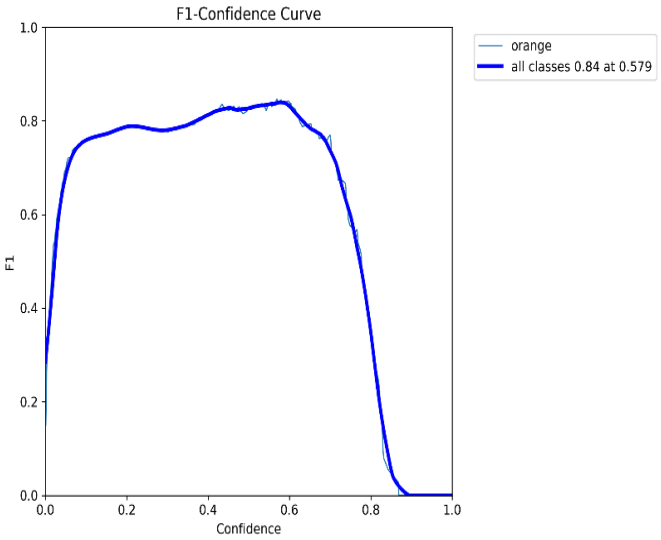

Figure 5. Graph showing the relationship between confidence and F1-Score.

Finally, the confusion matrix is widely used to assess the performance of a model and evaluate the performance of the convolutional neural network. However, in certain scenarios, it is possible that no values are identified as true negatives in the confusion matrix. This occurs when a model fails to correctly classify any examples as negative, resulting in the absence of values in the true negative cell. This situation can arise when there is an imbalance in the data, where the negative class is rare or underrepresented in the test set, as depicted in Figure 6.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-3-2024-303-2024 | © Author(s) 2024. CC BY 4.0 License.

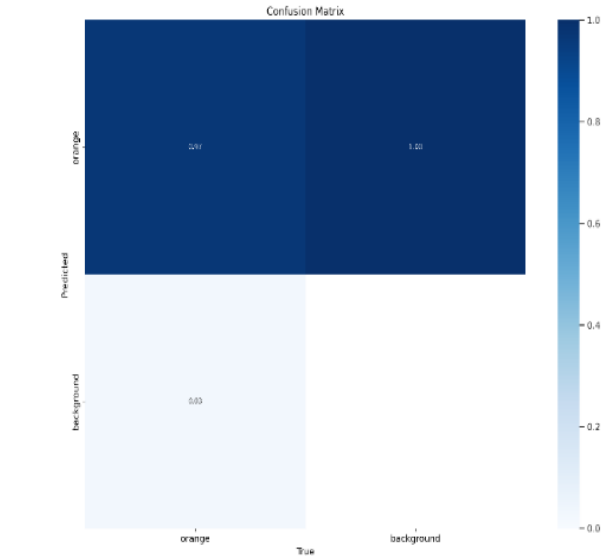

Figure 6. Confusion matrix for orange detection results.

With the model trained, a new test set was chosen, containing 5 images, each of them acquired with shutter speeds of 1/640 s, 1/500 s, 1/400 s, 1/320s, 1/250 s, in the 710 nm band. The results are presented from the lowest to the highest illumination conditions, with a confidence interval of 80% (Figure 7 to 11).

Figure 7. Image from the 710nm band, captured with a shutter speed of 1/640s.

Figure 8. Image from the 710nm band, captured with a shutter speed of 1/500s.

Figure 9. Image from the 710nm band, captured with a shutter speed of 1/400s.

Figure 10. Image from the 710nm band, captured with a shutter speed of 1/320s.

Figure 11. Image from the 710nm band, captured with a shutter speed of 1/250s.

The variations in the object's illumination affected the neural network identification results. The images acquired with a shutter speed of 1/640 s, enabled the identification of the shadowed fruits, while with a shutter speed of 1/250 s the images did not enable the detection of the same object. In contrast, the images with a larger shutter speed enabled the identification of fruits that are more evident. This can be attributed to the fact that the faster the capture of the image, the lower its brightness, making it difficult to identify fruits further in the background. On the other hand, when the capture speed is slower, more light enters the image, causing saturation in some areas, making it more challenging to identify objects that are closer.

It should be noted that during data acquisition, we aimed to alter only one variable involved in image acquisition, without changing the camera aperture and ISO of the image, so that only one variable would be studied and learned in the process.

4. Conclusion

In summary, analyzing the precision and recall graphs provides a comprehensive understanding of the orange detection model's capability. After 300 epochs, the model demonstrated an impressive precision of 81.5% and an approximate recall rate of 85%. These results clearly indicate that the model is well-tuned to its parameters, ensuring reliable performance in accurately detecting oranges.

Moreover, the computation of the mean Average Precision (mAP) at the IoU threshold of 0.5 provides a comprehensive assessment of the model's effectiveness by considering both precision and recall. The achieved mAP value of 91.6% demonstrates the YOLO model's ability to accurately detect oranges, exhibiting a substantial alignment between the predicted bounding boxes and the ground truth annotations.

Another noteworthy point is that the experiments revealed that the choice of shutter speed significantly influences the ability of oranges to be detected by the YOLO convolutional neural network. It was observed that different shutter speeds, such as 1/640s and 1/250s, which represent faster and slower speeds respectively, presented challenges in detection due to low illumination and overexposure. Intermediate values are thus more suitable for identifying a larger number of fruits. Alternatively, generating High Dynamic Range Images from multiple shots is another option.

For future work, there is a focus on improving evaluation by expanding the dataset with additional training images, incorporating new spectral bands, and exploring different confidence intervals.

Acknowledgments

This study was funded by: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES Grants: 88887.840159/2023-00, 88887.817757/2023-00 and 88887.817757/2023-00); São Paulo Research Foundation (FAPESP)- Thematic Project grant n. 2021/06029-7, and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), GRANT nº 308747/2021-6.

References

Altaheri, H., Alsulaiman, M., & Muhammad, G. (2019). Date fruit classification for robotic harvesting in a natural environment using deep learning. IEEE Access, 7, 117115-117133.

Bac, C. W., Hemming, J., van Tuijl, B., Barth, R., Wais, E., and van Henten, E. J. (2017). Performance evaluation of a harvesting robot for sweet pepper. J. Field Robot. 34, 1123–1139. doi: 10.1002/rob.21709

Chen, S.W., Shivakumar, S.S., Dcunha, S., Das, J., Okon, E., Qu, C., Taylor, C.J. Kumar, V. Counting apples and oranges with deep learning: A data-driven approach. IEEE Robot. Autom. Lett. 2017, 2, 781–788.

Fundecitrus (Fund for Citrus Protection). Final Orange Crop Update for the São Paulo and West-Southwest Minas Gerais Citrus Belt. Available: https://www.fundecitrus.com.br/pdf/pes_relatorios/0423_Final_Orange_Crop_Update.pdf, 2023.

Ganesh, P., Volle, K., Burks, T. F., & Mehta, S. S. (2019). Deep orange: Mask R-CNN based orange detection and segmentation. IFAC-PapersOnLine, 52(30), 70-75.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

Redmon, J., Santosh Divvala, Ross Girshick, and Ali Fa hadi. You only look once: Unified, real-time object detection. In CVPR, 2016.

Tommaselli, A. M. G. et al. Geometric performance of a camera with single sensor and multiple heads. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 43, 389-396, 2020.

Vo, T., 2022: Synseq4ed: A novel event-aware text representation learning for event detection. Neural Process. Lett. 54(1), 227– 249.

Williams, H. A., Jones, M. H., Nejati, M., Seabright, M. J., Bell, J., Penhall, N. D., ... & MacDonald, B. A. (2019). Robotic kiwifruit harvesting using machine vision, convolutional neural networks, and robotic arms. biosystems engineering, 181, 140-156.

Yamamoto, K., Guo, W., Yoshioka, Y., and Ninomiya, S. 2014. On plant detection of intact tomato fruits using image analysis and machine learning methods. Sensors 14, 12191–12206. doi: 10.3390/s140712191.

Yu, Y., Zhang, K., Yang, L., & Zhang, D. (2019). Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. Computers and Electronics in Agriculture, 163, 104846.

Zitnick, C. L., & Dollár, P. (2014). Edge boxes: Locating object proposals from edges. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13 (pp. 391-405). Springer International Publishing.