Video-Based Fruit Detection and Tracking for Apple Counting and Mapping

Jordi Gené-Mola
Institute of AgriFood Research and Technology (IRTA), Fruitcentre, Parc Científic i Tecnològic Agroalimentari de Gardeny (PCiTAL).
Lleida, Catalonia, Spain.
jordi.gene@irta.cat

Marc Felip-Pomés
Research Group in AgroICT& Precision Agriculture – GRAP, Department of Agricultural and Forest Sciences and Engineering, Universitat de Lleida (UdL) – Agrotecnio-CERCA Center.
Lleida, Catalonia, Spain
marc.felip@udl.cat

Francesc Net-Barnés
Computer Vision Center, Campus UAB
Barcelona, Catalonia, Spain.
Fnet@cvc.uab.cat

Josep-Ramon Morros
Department of Signal Theory and Communications, Universitat Politècnica de Catalunya
Barcelona, Catalonia, Spain
ramon.morros@upc.edu

Juan C. Miranda
Research Group in AgroICT& Precision Agriculture – GRAP, Department of Agricultural and Forest Sciences and Engineering, Universitat de Lleida (UdL) – Agrotecnio-CERCA Center.
Lleida, Catalonia, Spain
juancarlos.miranda@udl.cat

Jaume Arnó
Research Group in AgroICT& Precision Agriculture – GRAP, Department of Agricultural and Forest Sciences and Engineering, Universitat de Lleida (UdL) – Agrotecnio-CERCA Center.
Lleida, Catalonia, Spain
jaume.arno@udl.cat

Luís Asín
Institute of AgriFood Research and Technology (IRTA), Fruitcentre, Parc Científic i Tecnològic Agroalimentari de Gardeny (PCiTAL).
Lleida, Catalonia, Spain.
luis.asin@irta.cat

Jaume Lordan
Institute of AgriFood Research and Technology (IRTA), Fruitcentre, Parc Científic i Tecnològic Agroalimentari de Gardeny (PCiTAL).
Lleida, Catalonia, Spain.
jaume.lordan@irta.cat

Javier Ruiz-Hidalgo
Department of Signal Theory and Communications, Universitat Politècnica de Catalunya
Barcelona, Catalonia, Spain
j.ruiz@upc.edu

Eduard Gregorio
Research Group in AgroICT& Precision Agriculture – GRAP, Department of Agricultural and Forest Sciences and Engineering, Universitat de Lleida (UdL).
Lleida, Catalonia, Spain
eduard.gregorio@udl.cat

***Abstract*— Automatic fruit counting systems have garnered interest from farmers and agronomists to monitor fruit production, predict yields in advance, and identify production variability across orchards. However, accurately counting fruits poses challenges, particularly due to occlusions. This study proposes a multi-view sensing approach using continuous motion videos captured by a camera moved along the row of trees, followed by fruit detection in all video frames and application of Multi-Object Tracking (MOT) algorithms to prevent double-counting. Three tracking methods, namely SORT, DeepSORT, and ByteTrack, are compared for fruit counting using the YOLOv5x object detector. The methodology is applied to map fruit production in an experimental apple orchard at two different dates: four weeks and one week before harvest. The results demonstrate that ByteTrack (MOTA = 0.682; IDF1 = 0.837; HOTA = 0.689) outperforms SORT and DeepSORT, indicating its superior tracking performance. Computational efficiency analysis reveals similar processing times between SORT and ByteTrack (about 15 ms), while DeepSORT requires significantly more processing time per image (128 ms). Fruit counting evaluation shows reasonably accurate yield predictions on both dates, with reduced errors and improved performance closer to the harvest date (MAPE = 7.47 %; R^2 = 0.70). The system proves effective in estimating orchard fruit production using computer vision technology, offering valuable insights for yield forecasting. These findings contribute to optimizing fruit production and supporting precision agriculture practices. The code and the dataset have been made publicly available and a video visualization of results is accessible at http://www.grap.udl.cat/en/publications/video_fruit_counting.**

Keywords—Yield prediction, Yield mapping, Fruit Tracking, Multi-Object-Tracking, Precision Agriculture

I. Introduction

With the growing global population, there is an increasing demand of fruits that requires to optimize the fruit production to obtain more from limited land resources. Achieving sustainable agriculture requires effective monitoring of orchards to extract valuable information for informed management decisions. To this end, farmers and agronomists are interested on the use of automatic fruit counting systems to monitor the fruit production, predict the yield in advance and identify the variability of the fruit production along the orchards [1].

Fruit detection is a particular case of object detection. In the last years, most of the approaches use end-to-end trained Convolutional Neural Networks (CNNs) to perform the detection. While there are many architectures, most of the methods can be classified into two families: single stage and two stage models. Generally speaking, two stage detectors [2] can be more accurate, while one stage detectors [3] tend to be faster. In the last years, the one stage YOLO family is a very popular choice in many applications due to its excellent accuracy, real-time performance and simplicity of use.

Counting fruits accurately presents a challenge, particularly due to fruit occlusions [4]. One approach to mitigate this challenge involves acquiring data from different perspectives, assuming that apples occluded from one point of view will be visible from another [5]. However, this approach may lead to overestimating the fruit count due to double-counting from different image positions. Some researchers have proposed using 3D fruit detection and location to avoid double-counting [6]. Nonetheless, this solution needs complex photogrammetry and camera calibration algorithms. As an

alternative, this work proposes employing multi-view sensing by capturing continuous motion videos using a camera moved along the row of trees. Subsequently, fruits are detected in all video frames and Multi-Object Tracking (MOT) algorithms are applied to prevent fruit double-counting.

MOT is the process of locating several objects in successive frames of a video sequence. For this, each object must be given a unique ID across all the video sequence. Because of the availability of excellent object detectors, in the last years, MOT is dominated by the tracking-by-detection approach. This paradigm consists in first detecting the objects and then, assigning the same ID to the detections of the same object in different frames. Finding the correspondence between detections of the same object in consecutive frames is usually performed using motion and appearance cues. For instance, SORT [7] and ByteTrack [8] uses a Kalman filter to predict the position of the object in the next frame. Similarly, DeepSORT [9] adds a trained appearance descriptor for each object, so that the association is a combination of motion information and the similarity between the object descriptors.

In this study, we compare SORT, DeepSORT and ByteTrack algorithms for fruit counting in recorded videos. Furthermore, we apply this methodology to map the fruit production of an experimental apple at two distinct dates: 4 weeks before harvest and 1 week before harvest. This work offers two primary contributions: firstly, a comparison between three state-of-the-art tracking methods for fruit counting; secondly, an evaluation of a video-based yield prediction method at different fruit growth stages. The subsequent sections of this manuscript are organized as follows: Section II details the dataset used for this research and provides an overview of the algorithms applied within the methodology pipeline. Section III presents and discusses the results obtained for fruit detection, tracking, and counting tasks. Finally, in Section IV, we present the conclusions derived from this study and propose future research directions.

II. Materials and Methods

A. Dataset

The data utilized in this study was obtained from an experimental apple orchard (Story ® Inored), located in the municipality of Mollerussa, Catalonia Spain. The trees in the orchard were trained in a tall spindle system, with a planting frame of $3.6\ m \times 1\ m$ and a maximum canopy height of 3.5 m.

For video recording, the experimental set-up consisted of a Kinect Azure DK (Microsoft Corporation, Redmond, WA, USA) synchronized with an Ardusimple simpleRTK2B (Ardusimple Co., Ltd., Lleida, Spain) GNSS sensor. Both sensors were controlled with a Windows operating system laptop running the AK_ACQS software [10]. This system was mounted on a vertical mast situated at the rear of an electric all-terrain vehicle (Fig. 1a).

Fig. 1b presents a top view of the study site, which includes four apple tree rows of 105 m length, containing a total of 420 apple trees. Each row was divided into 21 stretches of 5 m each, resulting in a total of 84 stretches. Apples harvest took place on October 5th, 2021, when the coloration pattern was matched to 8-9 with the corresponding degradation grade on a 1–10 point scale (EUROFRU, Ctifl—Centre Technique Interprofessionel des Fruits et Legumes, Paris, France) [11]. At that point apples had a soluble solids

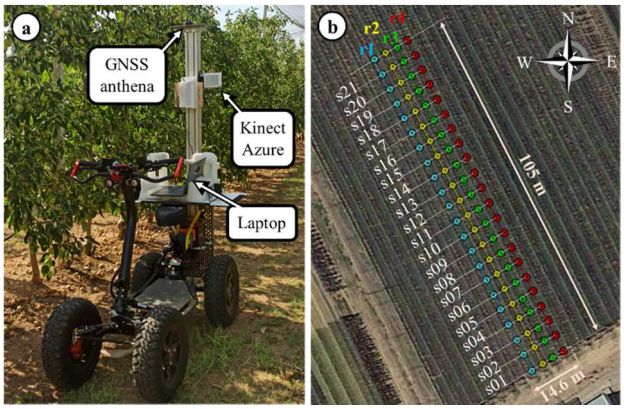

Fig. 1. (a) Orchard scanning equipment. (b) Top-view of the orchard showing the scanned rows and the row stretches used for evaluating yield mapping results.

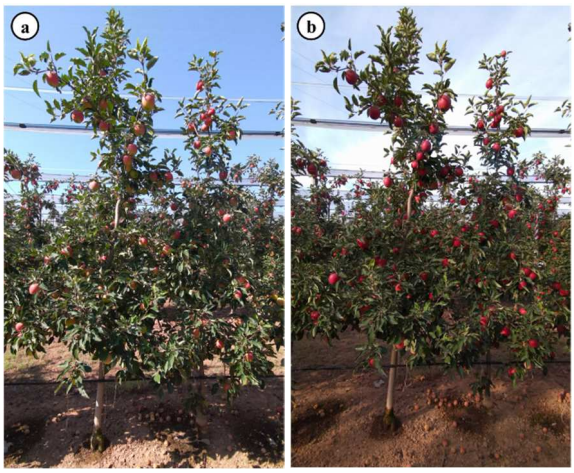

Fig. 2. Apple tree images acquired 4 weeks (a) and 1 week (b) before harvest.

content of 8.2º Brix, and a firmness of 14.6 kg/cm^2. Fruit load map ground truth was obtained by individually harvesting each stretch and counting the number of fruits per stretch with a commercial apple sorting line machine (Maf Roda Agrobotic, Montauban, France).

The study site was scanned two times: four weeks before harvest, specifically on September 6th, 2021, when apples had a bicoloured appearance with pink and green tones (Fig. 2a); and one week before harvest, on September 28th, 2021, when apples had attained a fully red coloration (Fig. 2b). To capture the data, the scanning system was driven at 0.556 m/s following the orchard alleyways. The distance between the sensor and the longitudinal axis of the imaged row of trees was maintained at 2.25 m to obtain a comprehensive view of the entire tree, from the bottom to the upper part. Each tree was scanned twice, once from the east and once from the west side. Videos were recorded at a frame rate of 30 frames per second.

A total of 1723 video frames from stretches 5 and 6 of row number 2 were annotated (1295 for training, 203 for validation and 225 for test). Bounding boxes with apple number IDs were manually assigned to all the apples visible in these frames using the Supervisly platform (https://supervisely.com/). To assess the accuracy of apple counting, we compared the number of fruits counted across all rows from 1 to 4 with the total amount of fruits harvested. Stretches 5 and 6 from row 2 were excluded from this evaluation to ensure that this test did not overlap with the training and validations sets.

B. Methodology pipeline

1) Fruit detection & tracking

In this work, we employ the YoloV5x [12] object detector CNN to detect fruits in each frame of the RGB video sequence recorded with Kinect Azure DK. The YoloV5x architecture was pretrained with the COCO [13] dataset and fine-tuned with our generated dataset (Section II.A). We employed the SGD with a learning rate of 0.01, weight decay of 0.0001 and a batch size of 4 for the fine-tuning process.

For the fruit tracking task, we compared and analyszed three different trackers: SORT, DeepSORT, and ByteTrack. SORT incorporates various association technics, including the Kalman Filter and Hungarian Algorithm, which rely on a motion model of detected objects [7]. DeepSORT extends SORT by incorporating appearance information obtained from a re-identification (Re-ID) network. In this study, we employed a cosine metric learning architecture [14] as a Re-ID network. In both SORT and DeepSORT, tracked identities are assigned by associating detections with scores higher than a threshold (set to 0.7 in this study). Alternatively, ByteTrack algorithm incorporates a second association step to associate lower scored detections to tracks that were not associated in the first step. This feature enhances the algorithm's robustness in the presence of objects partially occluded [8].

2) Fruit Counting

To count the total number of fruits per stretch, the number of fruits tracked from both the east (x_e) and west (x_w) sides of the row of trees were summed:

$$x_{ew,i,j} = x_{e,i,j} + x_{w,i,j} \quad (1)$$

Where $i \in \{1, 2, 3, 4\}$ refers to the row of trees number, and $j \in \{1,\ 2,\ 3, \dots,\ 21\}$ the stretch number. Then a linear model was used to estimate the actual number of fruits per stretch:

$$y_{i,j} = f(x_{ew,i,j},\ \beta_k, \varepsilon) = \beta_{k0} + \beta_{k1} x_{ew,i,j} + \varepsilon,\ k = 1,2,3,4 \quad (2)$$

To evaluate the fruit counting performance on each row, cross-validation was applied. To do so, the β_k values used to predict the number of fruits in row $k \in \{1, 2, 3, 4\}$, were estimated by $\hat{\beta}_k$ and obtained by minimizing the residual sum of squares (RSS_k) using the other rows (i , *where* $\forall\ i \neq k$):

$$RSS_k = \sum_i \sum_j (y_{i,j} - f(x_{ew,i,j},\ \hat{\beta}_k))^2,\ k = 1,2,3,4 \quad (3)$$

Where $y_{i,j}$ is the actual value of counted fruits in row *i* and stretch *j*. Uncertainty of the estimators $\hat{\beta}_k$ was evaluated using the corresponding standard errors, $se(\hat{\beta}_k)$. The coefficient of determination R^2, and the residual standard error, $\hat{\sigma}$, were used to check the goodness of fit.

3) Evaluation metrics

The evaluation of fruit detection results involves assessing the Precision (P), Recall (R), F1-score ($F1$) and mean average precision (mAP). P is calculated as the ratio of True Positives to the total number of detections. R represents the percentage of correctly detected apples compared to the total number of annotated apples. $F1$ is the armonic mean between P and R. The mAP is determined by computing the area under the P-R curve obtained across different Intersection over Union (IoU) thresholds.

For fruit tracking evaluation, we employ Multiple Object Tracking Accuracy ($MOTA$), Identification F1 (IDF1) and Higher Order Tracking Accuracy (HOTA). $MOTA$ considers false positives, false negatives, and identity switches to measure overall accuracy. $IDF1$ calculates the ratio of correctly identified detections over the average number of ground truth and computed detections [15]. $HOTA$ is an extension of the MOTA that combines various metrics to evaluate different error types such as detection recall and precision, association recall and precision, and localisation [16].

Furthermore, the fruit counting task is evaluated using Mean Absolute Error (MAE), Mean Bias Error (MBE), Mean Absolute Percentage Error ($MAPE$), Root Mean Square Error ($RMSE$), and the coefficient of determination (R^2). MAE measures the average absolute differences between the harvested and the predicted number of fruits. MBE is similar to MAE but without using the absolute value operator. $MAPE$ calculates the average absolute percentage differences. $RMSE$ averages the squared errors, giving more weight to larger errors. R^2 describes the strength of the linear regression between ground truth and predictions. Detailed descriptions of these metrics can be found in [17], where the same metrics are used to evaluate the performance for fruit sizing.

III. Results

A. Fruit detection and tracking

The object detection architecture YOLOv5x was trained for 273 epochs. At this point, the validation loss reached its minimum value and the network achieved a performance of $F1 = 0.834$ and $mAP = 0.871$. These results are in line with state-of-the-art outcomes in fruit detection, where $F1$ rates ranging from 0.730 to 0.970 have been reported [18].

Table I presents a comprehensive comparison of the performance of the three tested trackers. These results were obtained using validation data from stretches 5 and 6 from row number 2. The best performance for each evaluated metric are indicated in bold. ByteTrack demonstrated a MOTA score of 0.6817, surpassing SORT's 0.6401 and DeepSORT's 0.5735. Similar trends were observed in the IDF1 metric, with ByteTrack outperforming both SORT and DeepSORT. The HOTA scores showed more comparable results, with ByteTrack scoring 0.6894, slightly edging out SORT's 0.6804 and DeepSORT's 0.6824. It should be noted that the performance of these trackers is highly affected by the accuracy of the object detector, which must precisely localize the tracked objects [19]. To ensure a reliable comparison between these trackers, the same object detections from YOLOv5 were used.

These results reveal that ByteTrack consistently achieved slightly higher scores in terms of MOTA, IDF1, and HOTA compared to SORT, while DeepSORT exhibited comparatively lower scores. The authors attribute this finding to the inherent similarity in appearance among apples, which may limit the effectiveness of the Re-ID network. These outcomes suggest that the appearance features employed in the DeepSORT algorithm did not significantly contribute to improved fruit tracking performance. Villacres et. al (2023) [19] also compared different trackers including SORT and DeepSORT. In their work, SORT presented higher MOTA

TABLE I. VIDEO FRUIT TRACKING RESULTS

Tracker	MOTA	IDF1	HOTA	Time [ms/img]
SORT	0.6401	0.8091	0.6804	**15.3**
DeepSORT	0.5735	0.7646	0.6824	128.0
ByteTrack	**0.6817**	**0.8369**	**0.6894**	15.4

results than DeepSORT when considering a high probability of detection, while DeepSORT performed better as this probability decreases.

Considering the computational efficiency, both SORT and ByteTrack exhibited similar processing times per image, with SORT at 15.3 milliseconds and ByteTrack at 15.4 milliseconds. In contrast, DeepSORT required significantly more processing time per image, with an average of 128.0 milliseconds.

Based on these results, ByteTrack emerges as the best-performing method among the three trackers. It achieves higher accuracy metrics (MOTA, IDF1, and HOTA) compared to SORT and DeepSORT while maintaining similar computational efficiency to SORT. Therefore, ByteTrack offers a desirable combination of superior tracking performance and reasonable computational cost, making it a recommended choice for fruit tracking applications. For a video visualization of fruit tracking results the reader is referred to http://www.grap.udl.cat/en/publications/video_fruit_counting.

B. Fruit count prediction models

Table II shows the estimated parameters of the linear models for fruit counting introduced in Section II.B.2.. Regression coefficients ($\hat{\beta}_{k1}$) were homogeneous but greater than 1 for the detection carried out four weeks before the harvest, and less than 1 just one week before. This is possibly explained by the better detection at dates close to harvest, with fruits that are fully colored and somewhat larger. Regarding the uncertainty of predictions, the residual standard error ($\hat{\sigma}$) also reached similar values between models within the same date, with error averages of 49 and 45 apples depending on whether the detection was applied respectively four weeks or one week before harvesting (Table II).

C. Yield mapping

Table III presents the fruit counting results obtained at two different dates: four weeks before harvest (06/09/2021) and one week before harvest (28/09/2021). For this evaluation ByteTrack method was selected as it demonstrated superior performance in the previous section.

On 06/09/2021, the system achieved an MAE of 38.43, which represents an average percentage difference of 8.45% (±0.8 % SE) between the predicted and actual yield. The R^2 value of 0.66 indicates a moderate level of goodness of fit between the predicted and actual yield (Fig. 3a). On 28/09/2021, the system's performance improved slightly, evident in the reduced MAE of 34.98 and MAPE of 7.47% (±0.6 % SE). The R^2 value of 0.70 further demonstrates an enhanced correlation between the predicted and actual yield (Fig. 3b).

TABLE III. ESTIMATION OF PARAMETERS AND UNCERTAINTIES OF LINEAR FRUIT COUNT PREDICTION MODELS

	$\hat{\beta}_{k0}$	$se(\hat{\beta}_{k0})$	$\hat{\beta}_{k1}$	$se(\hat{\beta}_{k1})$	R^2	$\hat{\sigma}$
Scan performed four weeks before harvest - September 6th, 2021						
Row 1 (k=1)	62.13	43.39	1.07	0.11	0.62	49.36
Row 2 (k=2)	13.51	40.36	1.18	0.10	0.67	44.74
Row 3 (k=3)	29.88	41.05	1.14	0.10	0.67	50.61
Row 4 (k=4)	36.69	38.61	1.12	0.10	0.69	51.03
Scan performed four weeks before harvest - September 28th, 2021						
Row 1 (k=1)	-5.94	44.87	0.94	0.09	0.67	45.76
Row 2 (k=2)	-13.12	38.53	0.93	0.08	0.71	41.61
Row 3 (k=3)	-10.29	39.47	0.95	0.08	0.72	46.35
Row 4 (k=4)	-26.24	36.90	0.97	0.07	0.76	44.82

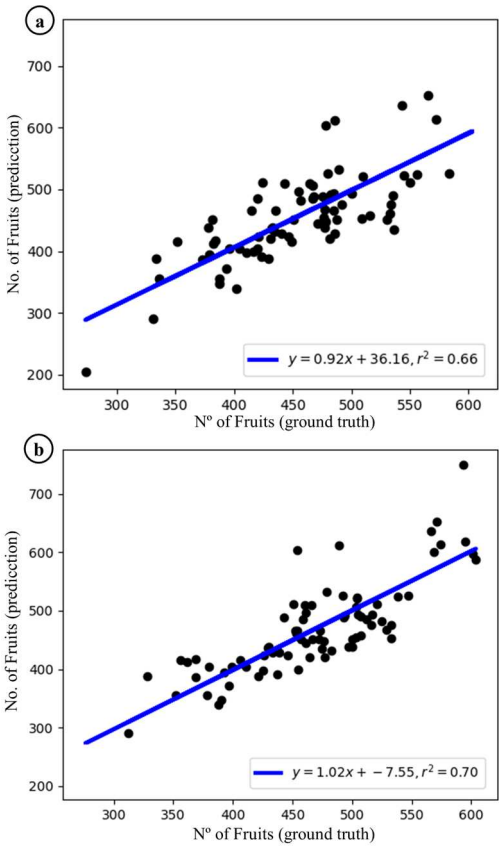

Fig. 3. Scatter plot of the total number of fruits harvested (x-axis) in each evaluated apple tree stretch versus the number of fruits counted with the designed system (y-axis) 4 weeks before harvest (a) and 1 week before harvest (b).

TABLE II. FRUIT COUNTING RESULTS

Date	MAE	MBE	MAPE	RMSE	R^2
06/09/2021	38.43	-0.20	8.45 %	49.04	0.66
28/09/2021	34.98	0.15	7.47 %	45.64	0.70

Fig. 4. shows a qualitative visualization of the ground truth (harvested) and predicted fruit production maps. Due to an error during ground truth data acquisition, the harvested fruit production of stretches 4, 5 and 6 of row 2, as well as stretch 4 of row 4, are missing.

While the quantitative evaluation (Table III) indicated a better prediction performance when scanning closer to the harvest date, the qualitative evaluation show that both predicted fruit production maps successfully identify the zones of the orchard with higher (southern part of the orchard) and lower fruit production (northern part of the orchard) (Fig. 4. *b* and *c*). The prediction error maps show that most of the stretches presented an error of about 50 apples although three outliers are identified in the southern part, where prediction was underestimated by more than -100 apples (Fig. 4 d and e).

Overall, the fruit counting system showed reasonably accurate yield predictions on both dates, with a reduction in errors and improved performance closer to the harvest date. The results indicate the system's effectiveness in estimating orchard yield using computer vision technology, providing valuable insights for yield forecasting and management decisions.

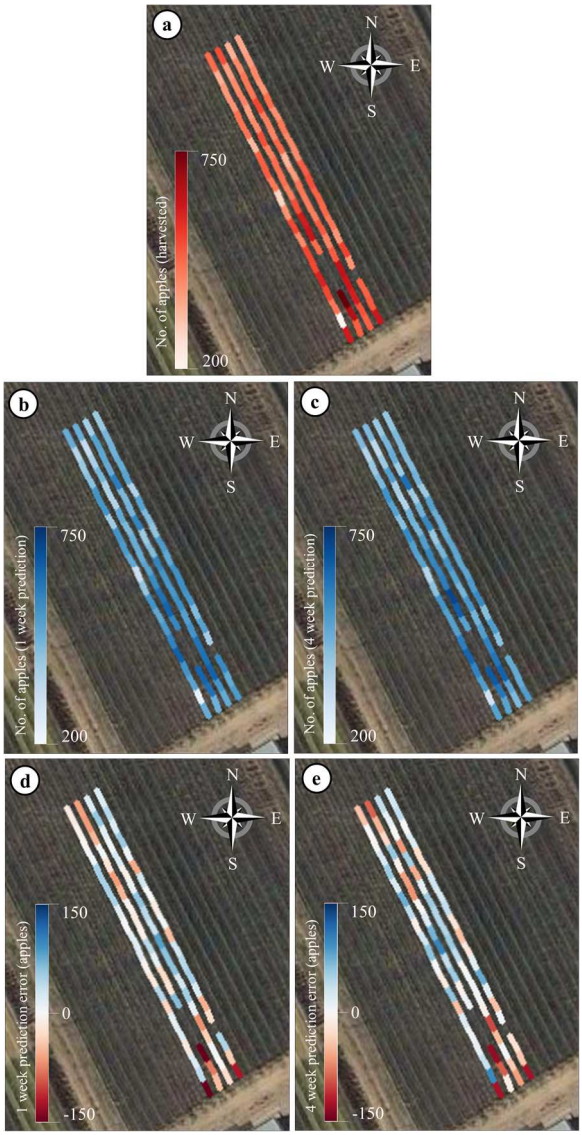

Fig. 4. (a) Ground truth yield map obtained at harvest (05/10/2021). (b) Predicted yield map obtained 1 week before harvest (28/09/2021). (c) Predicted yield map obtained 4 weeks before harvest (06/09/2021). (d) Prediction error obtained 1 week before harvest (28/09/2021). (e) Prediction error obtained 4 week before harvest (06/09/2021).

IV. Conclusions

In this study, we developed and evaluated a fruit counting system based on computer vision to predict orchard yield. The system was tested on an experimental apple orchard, capturing data four weeks and one week before harvest.

Firstly, the performance of three state-of-the-art tracking methods, namely SORT, DeepSORT, and ByteTrack, was compared for fruit counting in recorded videos. Among these methods, ByteTrack consistently demonstrated slightly higher scores in terms of MOTA, IDF1, and HOTA compared to SORT, while DeepSORT exhibited lower scores. These findings suggest that the appearance features utilized in the DeepSORT algorithm do not contribute significantly to improved fruit tracking. This can be attributed to the inherent similarity in appearance among apples, which limits the effectiveness of feature-based approaches.

Computational efficiency was also considered, and it was found that both SORT and ByteTrack exhibited similar processing times per image, with SORT at 15.3 milliseconds and ByteTrack at 15.4 milliseconds. In contrast, DeepSORT required significantly more processing time per image, averaging 128.0 milliseconds.

Furthermore, the fruit counting system was evaluated using metrics such as MAE, MBE, MAPE, RMSE, and R^2. The results showed reasonably accurate yield predictions on both dates, with reduced errors and improved performance closer to the harvest date. The system achieved an MAE of 38.43, corresponding to percentage difference of MAPE = 8.45% between the predicted and actual yield. On the other hand, one week before harvest, the system's performance improved slightly, with an MAPE of 7.47%. The R^2 values of 0.65 and 0.70 demonstrated moderate levels of goodness of fit between the predicted and actual yield for the respective dates. Besides these quantitative results, qualitative evaluation of the predicted fruit production maps showed the system's ability to identify zones of higher and lower fruit production in the orchard.

In conclusion, the fruit counting system based on computer vision, utilizing YOLOv5x and ByteTrack tracking methods, proved effective in estimating orchard yield. The system's performance improved as the harvest date approached, enabling accurate predictions and providing valuable insights for yield forecasting. The results obtained will contribute to the optimization of fruit production and support precision agriculture practices. Future research directions may focus on further refining the tracking methods and exploring additional factors that affect fruit counting accuracy. This includes investigating the influence of scanning conditions, such as lighting, vehicle and data acquisition speeds, and camera orientation and field-of-view, on the system's performance. Additionally, extending the applicability of the method to other fruit varieties with different appearances would provide valuable insights into the global use and robustness of the system.

Acknowledgment

This work was partly funded by the Departament de Recerca i Universitats de la Generalitat de Catalunya (grant 2021 LLAV 00088), the Spanish Ministry of Science, Innovation and Universities (grants RTI2018-094222-B-I00 [PAgFRUIT project], PID2021-126648OB-I00 [PAgPROTECT project] and PID2020-117142GB-I00 [DeeLight project] by MCIN/AEI/10.13039/501100011033 and by "ERDF, a way of making Europe", by the European Union). The work of Jordi Gené Mola was supported by the Spanish Ministry of Universities through a Margarita Salas postdoctoral grant funded by the European Union - NextGenerationEU. The Secretariat of Universities and Research of the Department of Business and Knowledge of the Generalitat de Catalunya and European Social Fund (ESF) are also thanked for financing Juan Carlos Miranda's pre-doctoral fellowship (2020 FI_B 00586).

References

[1] L. He *et al.*, "Fruit yield prediction and estimation in orchards: A state-of-the-art comprehensive review for both direct and indirect methods," *Comput Electron Agric*, vol. 195, no. September 2021, p. 106812, 2022, doi: 10.1016/j.compag.2022.106812.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Trans Pattern Anal Mach Intell*, vol. 39, no. 6, pp. 1137–1149, 2017, doi: 10.1109/TPAMI.2016.2577031.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE*

Computer Society Conference on Computer Vision and Pattern Recognition, 2016. doi: 10.1109/CVPR.2016.91.

[4] J. Gené-Mola *et al.*, "Looking behind occlusions: A study on amodal segmentation for robust on-tree apple fruit size estimation," *Comput Electron Agric*, vol. 209, Jun. 2023, doi: 10.1016/j.compag.2023.107854.

[5] D. Rapado-Rincón, E. J. van Henten, and G. Kootstra, "Development and evaluation of automated localisation and reconstruction of all fruits on tomato plants in a greenhouse based on multi-view perception and 3D multi-object tracking," *Biosyst Eng*, vol. 231, pp. 78–91, Jul. 2023, doi: 10.1016/j.biosystemseng.2023.06.003.

[6] J. Gené-Mola *et al.*, "Fruit detection and 3D location using instance segmentation neural networks and structure-from-motion photogrammetry," *Comput Electron Agric*, vol. 169, no. 105165, 2020, doi: https://doi.org/10.1016/j.compag.2019.105165.

[7] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2016, pp. 3464–3468. doi: 10.1109/ICIP.2016.7533003.

[8] Y. Zhang *et al.*, "ByteTrack: Multi-Object Tracking by Associating Every Detection Box," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. [Online]. Available: https://github.com/ifzhang/ByteTrack.

[9] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," *Proceedings - International Conference on Image Processing, ICIP*, vol. 2017-Septe, pp. 3645–3649, 2018, doi: 10.1109/ICIP.2017.8296962.

[10] J. C. Miranda, J. Gené-Mola, J. Arnó, and E. Gregorio, "AKFruitData: A dual software application for Azure Kinect cameras to acquire and extract informative data in yield tests performed in fruit orchard environments," *SoftwareX*, vol. 20, Dec. 2022, doi: 10.1016/j.softx.2022.101231.

[11] G. Planton, "Apple maturity: CTIFL-Eurofru starch code and the starchmeter. In The Apple Starch Test as a Decision Help for Harvest," *Infos CTIFL: Saint-Épain, France.*, 1995.

[12] Ultralytics, "YoloV5." Zenodo, 2022.

[13] T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *European conference on computer vision*, 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.

[14] N. Wojke and A. Bewley, "Deep Cosine Metric Learning for Person Re-identification," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 748–756. doi: 10.1109/WACV.2018.00087.

[15] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance Measures and a Data Set for Multi-target, Multi-camera Tracking," in *Computer Vision - ECCV 2016. Lecture Notes in Computer Science*, in Lecture Notes in Computer Science, vol. 9914. Cham: Springer, 2016. doi: 10.1007/978-3-319-48881-3.

[16] J. Luiten *et al.*, "HOTA: A Higher Order Metric for Evaluating Multi-object Tracking," *Int J Comput Vis*, vol. 129, no. 2, pp. 548–578, Feb. 2021, doi: 10.1007/s11263-020-01375-2.

[17] M. Ferrer-Ferrer, J. Ruiz-Hidalgo, E. Gregorio, V. Vilaplana, J. R. Morros, and J. Gené-Mola, "Simultaneous fruit detection and size estimation using multitask deep neural networks," *Biosyst Eng*, vol. 233, pp. 63–75, Sep. 2023, doi: 10.1016/j.biosystemseng.2023.07.010.

[18] A. Koirala, K. B. Walsh, Z. Wang, and C. McCarthy, "Deep learning – Method overview and review of use for fruit detection and yield estimation," *Computers and Electronics in Agriculture*, vol. 162. Elsevier B.V., pp. 219–234, Jul. 01, 2019. doi: 10.1016/j.compag.2019.04.017.

[19] J. Villacrés, M. Viscaino, J. Delpiano, S. Vougioukas, and F. Auat Cheein, "Apple orchard production estimation using deep learning strategies: A comparison of tracking-by-detection algorithms," *Comput Electron Agric*, vol. 204, Jan. 2023, doi: 10.1016/j.compag.2022.107513.