Fruit detection and 3D location using instance segmentation neural networks and structure-from-motion photogrammetry

Jordi Gené-Mola[a,*], Ricardo Sanz-Cortiella[a], Joan R. Rosell-Polo[a], Josep-Ramon Morros[b], Javier Ruiz-Hidalgo[b], Verónica Vilaplana[b], Eduard Gregorio[a]

[a]*Research Group in AgroICT & Precision Agriculture, Department of Agricultural and Forest Engineering, Universitat de Lleida (UdL) – Agrotecnio Center, Lleida, Catalonia, Spain.*

[b]*Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, Barcelona, Catalonia, Spain.*

Abstract

The development of remote fruit detection systems able to identify and 3D locate fruits provides opportunities to improve the efficiency of agriculture management. Most of the current fruit detection systems are based on 2D image analysis. Although the use of 3D sensors is emerging, precise 3D fruit location is still a pending issue. This work presents a new methodology for fruit detection and 3D location consisting of: (1) 2D fruit detection and segmentation using Mask R-CNN instance segmentation neural network; (2) 3D point cloud generation of detected apples using structure-from-motion (SfM) photogrammetry; (3) projection of 2D image detections onto 3D space; (4) false positives removal using a trained support vector machine. This methodology was tested on 11 Fuji apple trees containing a total of 1455 apples. Results showed that, by combining instance segmentation with SfM the system performance increased from an F1-score of 0.816 (2D fruit detection) to 0.881 (3D fruit detection and location) with respect to the total amount of fruits. The main advantages of this methodology are the reduced number of false positives and the higher detection rate, while the main disadvantage is the high processing time required for SfM, which makes it presently unsuitable for real-time work. From these results, it can be concluded that the combination of instance segmentation and SfM provides high performance fruit detection with high 3D data precision. The dataset has been made publicly available and an interactive visualization of fruit detection results is accessible at http://www.grap.udl.cat/documents/photogrammetry_fruit_detection.html

Keywords: Structure-from-motion; fruit detection; fruit location; Mask R-CNN; terrestrial remote sensing

1. Introduction

The need to provide food for an increasingly large population, while at the same time minimizing the agricultural impact on the environment, makes it essential to devote as much effort as possible to the development of techniques and methods that can ensure the increased efficiency, quality, and sustainability of agricultural activities. To achieve this goal, precision agriculture (PA) is establishing itself as a cornerstone approach which, based on crop information obtained with various techniques, provides tools for optimizing crop management and making appropriate decisions (ISPA, 2019). The

* *Corresponding author*.

E-mail address: j.gene@eagrof.udl.cat

monitoring of crops through the combination of sensors, processing systems, and mobile platforms –terrestrial, airborne or spaceborne– to carry this instrumentation, are key to providing precise and detailed crop information. Such questions are usually the starting point of optimization processes.

Knowledge of the spatial (3D) distribution of fruits through their detection and location, with different levels of resolution –within a specific tree and at plot level– is of enormous interest in agriculture. Having this information allows harvest and production estimates to be made, which leads to better planning of harvesting, storage and marketing tasks (Bargoti and Underwood, 2017; Nuske et al., 2014). With such information, it is also possible to know the spatial distribution of fruits and yield, and to relate it to the rest of the variables and factors that influence the management of plantations, such as the strategies of irrigation, fertilization and pruning, the characteristics and variability of the soil composition, the topographic characteristics of the plot, the size and structure of the trees, pest and disease impact, and so on. In addition, knowledge of the georeferenced distribution of fruits along the plot can be a starting point for robotized harvesting, as the harvester robot would have the coordinates of each fruit and could primarily focus on the collection process itself, with a resulting gain in speed and efficiency.

The characterization of the 3D spatial distribution of fruits, at both tree and plot scale, is a highly active research field. Commonly used sensors include RGB, multispectral, hyperspectral and thermal cameras, as well as 3D sensor technology such as LiDAR and depth cameras (RGB-D) (Li et al., 2014; Narvaez et al., 2017). Each of these sensors has its own strengths and weaknesses when used in real-field conditions, with the best choice depending on the specific application. Thus, while RGB cameras are economically affordable and user-friendly, they are severely affected by lighting conditions (Gongal et al., 2015). Both multi and hyperspectral cameras add spectral information beyond RGB bands, allowing the extraction of a rich set of parameters and vegetation indexes, but they are more expensive and time-consuming. In the case of thermal cameras, which capture the temperature information of objects, the different thermal inertia between fruits and background enables their differentiation. However, measurements are affected by the fruit size and the thermal evolution of the environment along the day, leading to a narrow temporal range of operations in field measurements (Bulanon et al., 2008; Gongal et al., 2015). Both LiDAR and RGB-D systems allow the 3D characteristics of fruits and plants to be directly obtained by determining the sensor-target distance, with time-of-flight and structured-light the most common measuring principles. Both systems allow the generation of high density 3D point clouds (coloured in the case of RGB-D sensors) of plants and fruits. While LiDAR sensors are usually quite expensive and not user-friendly, RGB-D are commonly low-cost plug-and-play sensors but they lose performance in high luminance environments, which is a drawback under real-field conditions (Rosell-Polo et al., 2015). Finally, through the post-processing of digital images, photogrammetry techniques are

being used to obtain 3D representations of different scenarios in many fields, including agriculture (Torres-Sánchez et al., 2018). One of the most successful and commonly used methods is called structure-from-motion (SfM), which identifies common characteristics in the collected images to infer the camera positions and then build the 3D representation of the scene (Westoby et al., 2012).

With respect to data processing, many state-of-the-art fruit detection systems use handcrafted features to encode the data acquired with different sensors and subsequently apply algorithms to obtain the fruit detection and location (Bargoti and Underwood, 2017; Gené-Mola et al., 2019c). More recently, remarkable progress has been achieved through the introduction of deep learning, which is based on multiple layer artificial neural networks (Koirala et al., 2019). Most approaches in fruit detection are based on the analysis of 2D images, although the processing of 3D images is quickly emerging (Nguyen et al., 2016; Tao and Zhou, 2017). Due to the unstructured environment of tree crops, occlusions of fruits with other vegetative organs and changing lighting conditions are the main problems that have to be dealt with (Gongal et al., 2015). To increase fruit visibility, some authors have proposed the use of multi-view imaging (Hemming et al., 2014), although it may lead to some fruits being counted twice if a proper image registration methodology is not used. To do so, Stein et al. (2016) proposed the use of epipolar geometry combined with the Hungarian algorithm (Kuhn, 2010). Similarly, Liu et al. (2018) used the Hungarian Algorithm refined with SfM to track fruits in video fruit counting. In contrast, Gongal et al. (2016) identified duplicate apples by projecting 2D image detections onto 3D models generated using RGB-D sensor data.

This work presents a new methodology for fruit detection and 3D location, combining the use of instance segmentation neural networks and SfM photogrammetry. The Mask R-CNN (He et al., 2017) deep neural network was used to detect and segment fruits in 2D RGB images. Then, SfM was used to generate an accurate 3D model and locate the detected fruits in the space. The main advantages of using SfM are that: (1) it is a multi-view approach and, in consequence, presents a reduced number of fruit occlusions; (2) the registration between images is automatically done, which ensures no double counting of apples appearing in different images. The remainder of this paper is structured as follows: **Section 2** presents the experimental setup, the acquired dataset, and the methodology pipeline, including a description of the deep neural network used for fruit detection, the SfM technique used to generate the 3D model, and the projection of 2D image detections onto the 3D generated model; **Section 3** evaluates the detections both in the 2D images and in the 3D model, while **Section 4** discusses the results; Finally, **Section 5** presents the conclusions obtained in this study and proposes future research directions.

2. Materials and Methods

2.1. Data acquisition.

Tests were carried out in a commercial Fuji apple orchard (*Malus domestica* Borkh. cv. Fuji) located in the municipality of Agramunt, Catalonia, Spain (E: 336,297 m; N: 4,623,494 m; 312 m a.s.l., UTM 31T - ETRS89). Trees grown in the studied orchard were trained in a tall spindle system, with a plantation frame of 4 x 0.9 m and a maximum canopy height and width of approximately 3.5 m and 1.5 m, respectively. The studied section was formed by 11 consecutive trees from the same row of trees, containing a total of 1455 apples. Images were acquired at the end of September 2017, at BBCH phenological growth stage 85 –advanced ripening, increase in intensity of cultivar-specific color– (Meier, 2001).

In the choice of photographic equipment and its setup, the quality of the photographs was prioritized. An EOS 60D DSLR Canon camera, with an 18 MP (5184 x 3456 px) CMOS APS-C sensor (22.3 x 14.9mm) was used (Canon Inc. Tokyo, Japan). Regarding the optics, a Canon EF-S 24mm f/2.8 STM lens was chosen, with a 35 mm film equivalent focal length of 38 mm and with a field of view of [59° 10 ', 50° 35'] (horizontal, vertical).

A total of 582 photographs were taken, 291 images per row side. No artificial light was used. The photographs were taken freehand, which allowed an average shooting frequency of 8 photographs per minute. Thus, the lighting conditions between the first and last photograph were very similar. The east face was photographed in the morning (11:53 - 12:26h) and the west face in the afternoon (15:27 - 16:05h), with a similar illumination obtained in both faces.

Images were taken from 53 photographic positions (per side). In each position, a vertical sweep of 5-6 photographs was taken (**Fig. 1a**) from the lower part (soil-trunk) to the upper part of the trees. The separation between two consecutive positions was 22 cm (**Fig. 1b**). These photographic positions defined a line parallel with respect to the apple tree row. The distance between the camera and the middle plane of the row was around 3 m and the height of the camera above the ground was 1.7 m (**Fig. 1a**). With this configuration, the vertical and horizontal overlapping between neighbouring images was higher than 30% and 90%, respectively (**Fig. 2**). This dataset has been made publicly available at www.grap.udl.cat/en/publications/datasets.html (Fuji-SfM dataset).

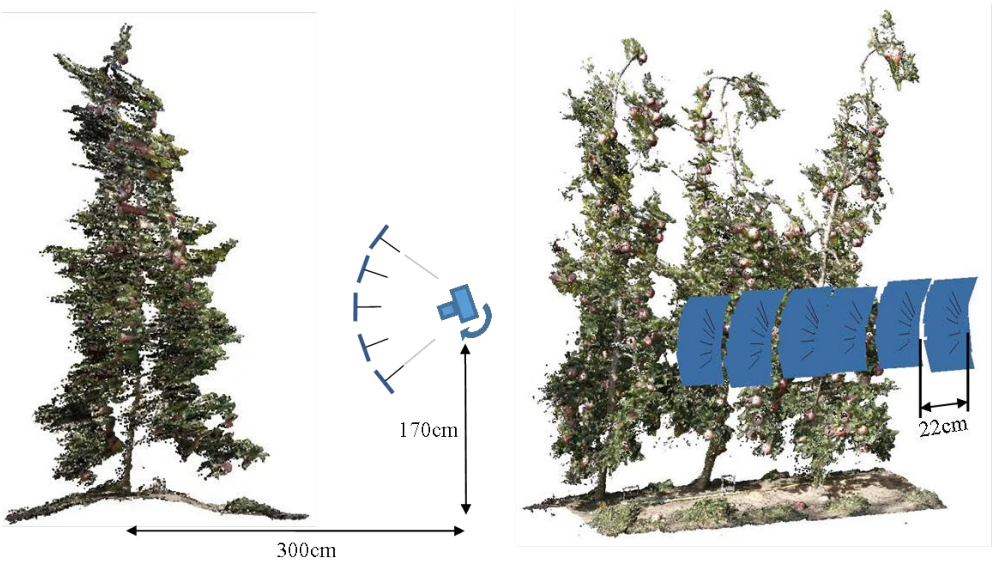

Fig. 1. a) Transversal scheme of the layout and distances of the photographic process. b) Isometric view of three scanned trees showing the separation between consecutive photographic positions.

Fig. 2. a) Vertical overlapping between two contiguous photographs. b) Horizontal displacement between two adjacent photographic positions.

2.2. Methodology pipeline

As shown in **Fig. 3**, the proposed fruit detection and location methodology includes the following processing steps: 1) 2D RGB image instance segmentation; 2) 3D point cloud generation using SfM photogrammetry; 3) Projection of 2D detections onto the 3D point cloud.

Due to the large amount of apples per image and the fact that convolutional neural networks performance decreases when detecting small objects, before applying the instance segmentation step the images were split into 24 sub-images of 1024x1024 pixels. Then, the convolutional neural network Mask R-CNN (He et al., 2017) was used to detect and segment the apples (**Section 2.2.1**). Apple detections and masks in the cropped images were translated to the original images. These

masked images were used to generate a 3D model by means of SfM photogrammetry, thus, only the 3D model of the objects of interest (apples) was generated (Section 2.2.2). To count the total number of fruits, and to know which 3D points belong to each apple, the last step used the camera matrices obtained from SfM camera alignment to project 2D detections onto 3D point clouds following the pinhole camera model (Section 2.2.3). Further details of the implementation of these steps are described in the following sub-sections.

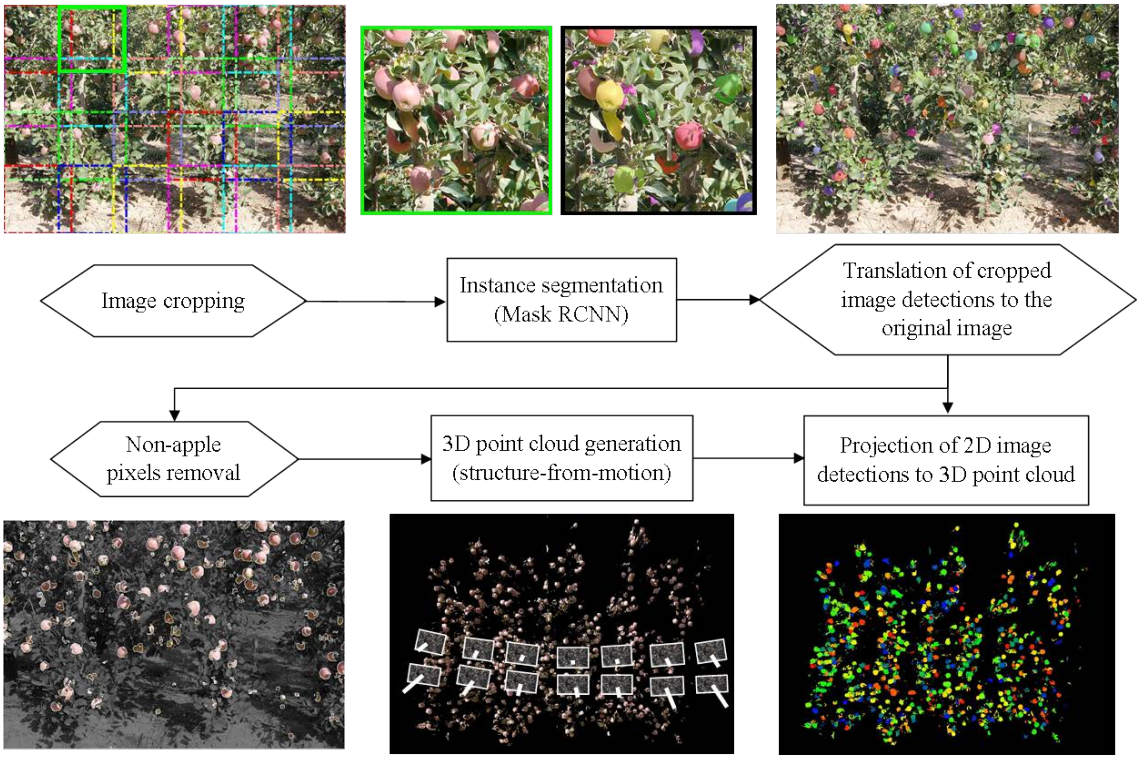

Fig. 3. Fruit detection and location methodology flowchart. Hexagons represent data preparation steps while rectangles define data processing steps.

2.2.1. Instance segmentation

The Mask R-CNN (He et al., 2017) deep neural network was used for apple detection and segmentation (instance segmentation) in acquired 2D RGB images. For an input image, this model provides 2D bounding boxes and semantic masks for the objects in the scene. It is an extension of the Faster R-CNN (Ren et al., 2017) network that adds a branch for predicting segmentation masks on each region of interest (RoI).

The operation is depicted in Fig. 4. Two parts can be differentiated in the architecture: the backbone, used for feature extraction, and the network head for bounding-box recognition (classification and regression) and mask prediction, that is applied separately to each RoI.

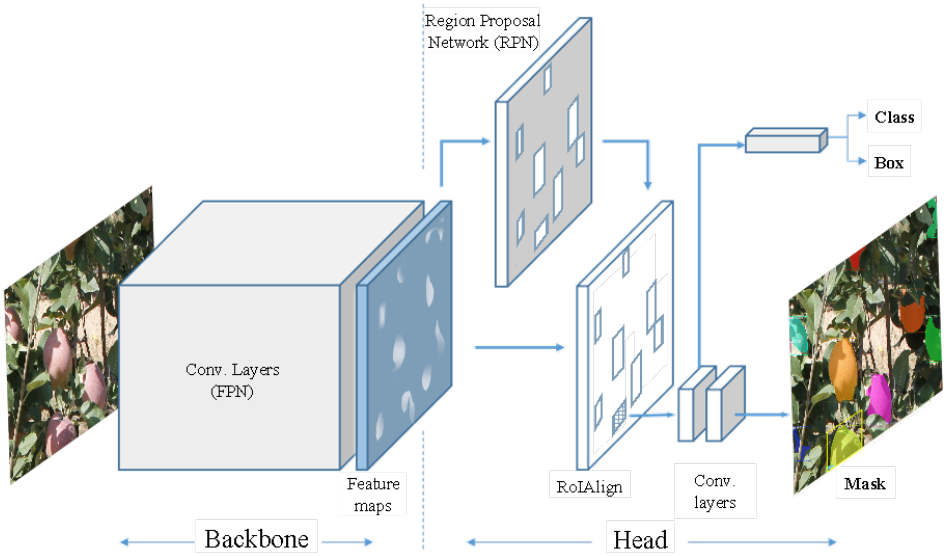

Fig. 4. Diagram of Mask R-CNN architecture.

The backbone is a feature pyramid network (FPN) (Lin et al., 2017), a type of fully convolutional network that exploits the inherent multi-scale, pyramidal hierarchy of deep convolutional networks to construct a feature pyramid map that provides RoI features from different levels of the feature pyramid according to their scale.

The Mask R-CNN network head is a small network that is slid over the feature map. Each sliding window is mapped to a lower-dimensional feature. At each sliding-window location, multiple region proposals are simultaneously predicted. The proposals are parameterized relative to a set of reference boxes, called anchors. An anchor is centred at the sliding window in question, and is associated with a scale and aspect ratio. This anchor-based design improves computational efficiency allowing features to be shared without an extra cost for addressing scales.

The obtained features are fed into two sibling fully connected layers—a box-regression layer and a box-classification layer. The process can be described in two stages. The first stage employs a region proposal network (RPN) to scan the feature pyramid map provided by the backbone and outputs a set of regions (region proposals) that are candidates to contain objects. The RoIAlign layer shares the forward pass of a CNN for an image across its subregions. Then, the features in each region are pooled using bilinear interpolation to maintain a precise alignment. The second stage classifies the object inside each one of the proposed regions into a set of predetermined classes, refines the bounding box and provides a pixel level mask for the object. The predictions of the class, bounding box and binary mask for each RoI are performed in parallel.

We used an existing implementation of the Mask RCNN obtained from Abdulla (2017) with a ResNet-101-FPN backbone. A model pre-trained in the COCO dataset (Lin et al., 2014) was adapted for Fuji apple detection by restricting

the number of classes to one and by fine-tuning the model using 12 images containing a total of 1749 manually annotated apples. This small dataset used to train and validate the Mask RCNN did not include images from trees assessed in the 3D location approach, ensuring that the data used to test the system was not used for training. In order to have a better relation between image size and fruit size, and due to the large number of fruits per image, each image was split into 24 sub-images of 1024x1024 pixels (6 horizontal and 4 vertical divisions, with an overlapping of 213 px in vertical and 192 px in horizontal). Thus, the dataset used to train and validate the Mask R-CNN consists of 288 sub-images, split into training and validation as shown in **Table 1**. Horizontal flipping data augmentation was used to increase the number of training images. The learning rate was set to 0.001, with a learning momentum of 0.9 and a weight decay of 0.0001. This dataset and the corresponding annotations have been made publicly available at www.grap.udl.cat/en/publications/datasets.html (Fuji-SfM dataset).

Table 1. Dataset configuration.

Mask R-CNN training - validation		
Raw image size	Sub-image size	
5184 x 3456 px	1024x1024 px	
Training	Validation	No. of fruits (annotated)
231 sub-images	57 sub-images	1749

Data for 3D point cloud generation	
Raw image size	No. of images
5184 x 3456 px	582 (291 per row side)

3D data			
No. of trees	No. of fruits	Training	Test
11	1455	3 trees	8 trees

2.2.2. 3D point cloud generation

To reconstruct the 3D information from the multiple 2D images, a classical multi-view SfM technique based on bundle adjustment (Triggs et al., 2000) was employed in each row side. This approach aims to simultaneously determine the structure (3D coordinates of scene points) and the calibration parameters of each of the cameras that minimize the total reprojection error.

In particular, Agisoft Professional Photoscan software was employed to perform the 3D reconstruction (v1.4, Agisoft LLC, St. Petersburg, Russia). The specific software configuration parameters set are detailed in Appendix **A**, Table **A1**. The three main steps followed to generate the 3D point cloud are:

a. Feature matching: where correspondences between points across different images are computed.
b. Camera estimation: using the previous correspondences, camera parameters and locations are estimated for each image.
c. Dense reconstruction: camera parameters are used to project 2D image points into their corresponding 3D locations.

The relationship between 2D image points and 3D locations is described following a pinhole camera model. Let x be a representation of a 3D point in homogeneous coordinates (a 4-dimensional vector), and let p be a representation of the 2D image of this point in the pinhole camera (a 3-dimensional vector in homogenous coordinates). Then, the relation between them can be expressed as:

$$p = C_i \cdot x \,, \tag{1}$$

where C_i is the 3x4 camera matrix that represents the intrinsic (matrix K) and extrinsic (matrix $[R_i\ T_i]$) camera parameters for camera i:

$$C_i = K\,[R_i\ T_i]\,, \tag{2}$$

In our case, as all images were taken with the same camera, intrinsic camera parameters are shared between all images (no i subindex in matrix K). Extrinsic parameters, on the other hand, are different for each image. Thus, rotation matrices R_i and translational vectors T_i are defined for each image and related to the first image of the dataset (camera $i = 0$ uses $R_0 = I$ and $T_0 = [0\ 0\ 0]$).

Fig. 5a represents the 3D point cloud generated using original RGB images. This point cloud was manually annotated, placing rectangular bounding boxes around each apple (**Fig. 5b)**. A total of 1455 apples were annotated in the point cloud, which is similar to the total number of apples manually counted in the orchard (1444 apples). The small difference between the number of annotations and the number of apples counted in the orchard can be attributed to human error during fruit counting. Annotated 3D bounding boxes were used as ground truth to evaluate the performance of the system in **Section 3.2**.

By using a mask in the original images –obtained with the trained Mask R-CNN described in **Section 2.2.1**– only the apples (not the entire trees) are reconstructed in **Fig. 5c**. Using masked images was desirable to only reconstruct the 3D model of the objects of interest (apples) and to reduce the computational time. As the 3D reconstruction stage is scale invariant, a set of known markers (depicted in **Fig. 5d**) separated by 85 cm were used to scale the resulting 3D point cloud to a real-world scale.

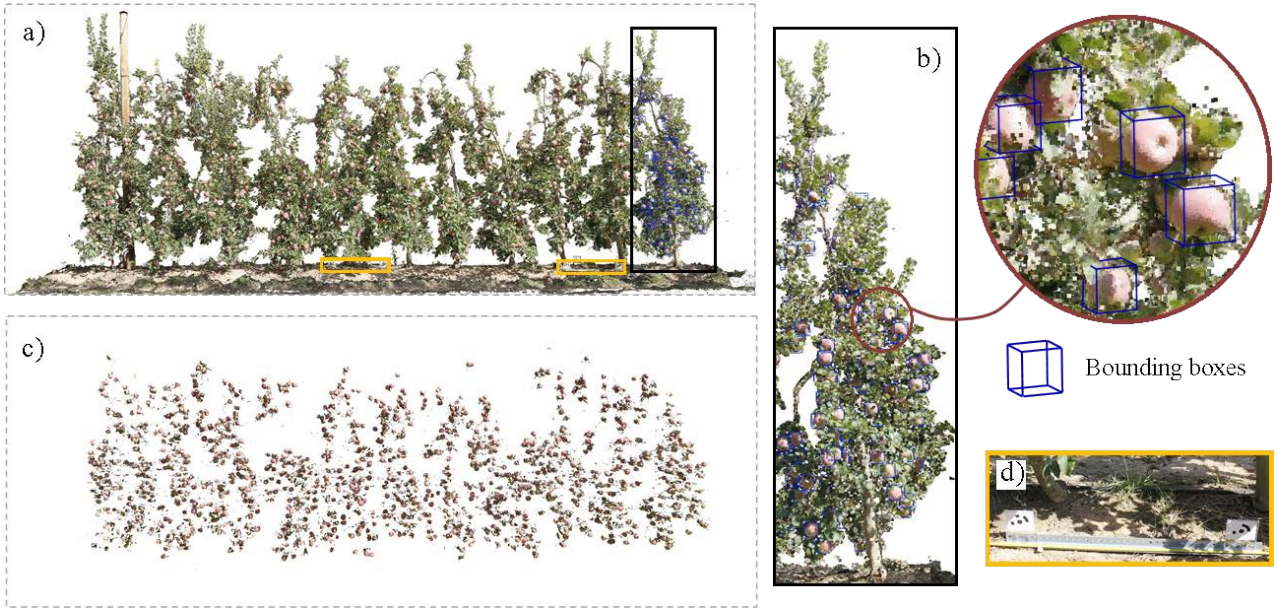

Fig. 5. a) Illustration of the 3D point cloud obtained using original RGB images. Yellow rectangles show the positions where reference markers were placed. b) Annotated point cloud with 3D rectangular bounding boxes placed around each apple. c) Apples 3D point cloud obtained using masked images. d) Illustration of reference markers used to scale the resulting 3D point cloud.

2.2.3. Projection of 2D detections onto 3D point cloud

Although SfM photogrammetry with masked images allows generation of the 3D model of only the objects of interest (apples), the resulting point cloud should be clustered in groups of 3D points per apple (3D apple detections) to count and locate detected fruits.

Knowing the intrinsic camera parameters (matrix *K*), as well as the pose and orientation of all images (matrix [*Ri Ti*]), 2D image detections were projected onto the 3D point cloud using the pinhole camera model (Eqs. (1) and (2)). The main issues to deal with during these projections were: (1) identification of objects (apples) behind detections; (2) unification of detections of an object detected from different photos.

Fig. 6 illustrates the steps carried out to perform the 2D to 3D projection, showing an example with two images taken from different positions. To assist visualization, Fig. 6a shows a small region of the scanned scene and Fig. 6b shows the 3D model obtained applying SfM photogrammetry with masked images. In Fig. 6c, detections from image 1 (img1) were projected onto the 3D point cloud. Due to the position of the camera with respect to the scene, an apple was occluded behind the green detection. In consequence, after projecting the 2D green detection, the detected and the occluded apples were clustered within the same group of 3D points (plotted in green in the 3D model of Fig. 6c). To identify objects behind a detection, a connected components labelling was applied to each 3D projection using the density-based scan algorithm DBSCAN (Ester et al., 1996). The minimum distance between connected points was set to 3 cm. If more than one group of connected points were found in a 3D detection, only the nearest (to the camera) was selected. Comparing Fig. 6c and Fig.

6d, it can be observed how the apple behind the green detection was released after applying DBSCAN. Having the detections of img1 in the 3D point cloud, the next image (img2) was processed. Detections from img2 that presented an overlap higher than 50% (IoU > 0.5) with previously detected apples were identified and unified (Fig. 6e), and new detections with no overlap with previous detections or with IoU < 0.5 were projected onto the 3D point cloud (Fig. 6f). The process was repeated for all the images used to generate the 3D point cloud.

In order to reduce the number of false positives, a linear support-vector-machine (SVM) was trained to identify and remove false positive detections. This SVM was fed using 4 features per detection:

- Number of points P that contain a 3D detection.
- Detection volume V.
- Detection density $\delta = \frac{V}{P}$.
- Geometric feature $\Psi = 27 \cdot \lambda_{1n} \cdot \lambda_{2n} \cdot \lambda_{3n}$, where $[\lambda_{1n}, \lambda_{2n}, \lambda_{3n}]$ are the normalized eigenvalues (so that $\lambda_{1n} + \lambda_{2n} + \lambda_{3n} = 1$), obtained applying singular value decomposition (SVD) on the 3D points of a detection. The applied coefficient of 27 allows Ψ to be bounded between 0 and 1, with 1 being for spherical detections.

The graphical representation of these features is shown in Appendix B, Fig. B 1. In order to train this SVM, 3 trees (out of 11) containing a total of 434 apples were used as the training dataset. The result of identifying and removing false positive detections can be observed in Fig. 6g, where the blue detection has been removed.

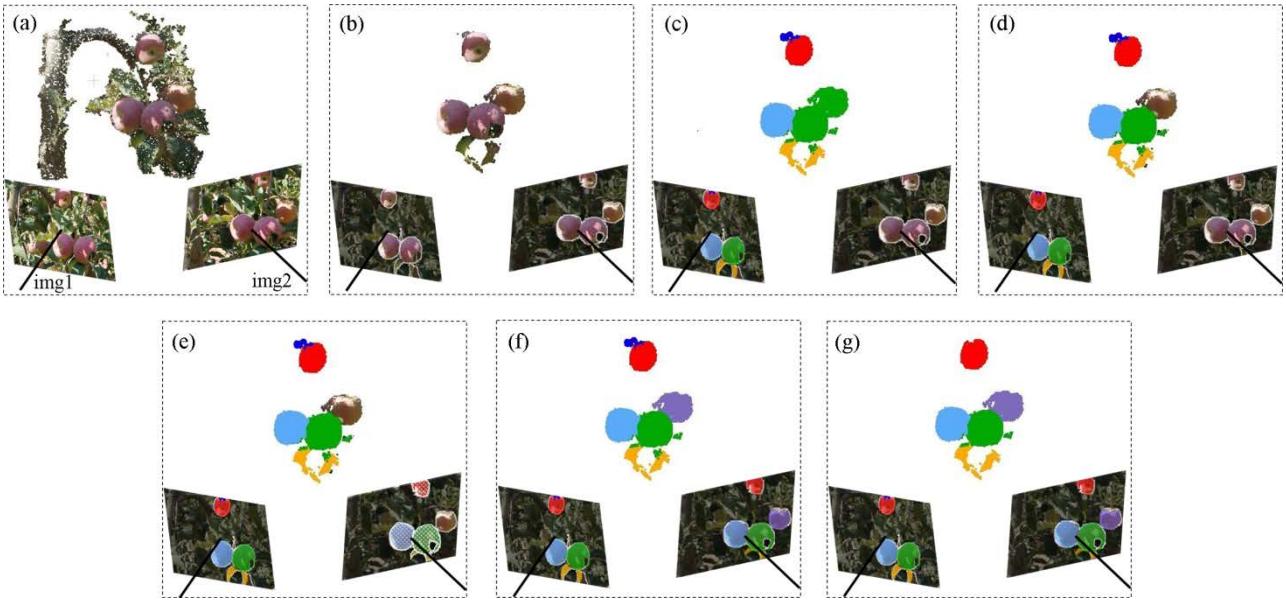

Fig. 6. Projection of 2D detections onto 3D point cloud. a) Data acquisition. b) 3D model obtained using structure-from-motion with segmented images. c) Projection of detections from image 1 (img1) onto the 3D point cloud. d) Identification of apples behind detections. e) Identification of apples appearing in a new image that were previously detected in other images. f) Projection of a new detection (coloured in purple) from image 2 (img2). g) False positive removal.

3. Results

This section evaluates the fruit detection performance both in the 2D images and in the 3D point cloud. Instance segmentation results are reported in terms of recall (R), precision (P), F1-score and average precision (AP) (Zhang and Zhang, 2009), considering as true positives detections with a ground truth mask overlap higher than 50% (IoU > 0.5). Similarly, the 3D fruit detection results are assessed in terms of detection rate (DR), recall (R), precision (P), false positive rate (FPR), muti-detection rate (MDR), and F1-score, as follows:

$$DR = \frac{LD}{T}, \quad (3)$$

$$R = \frac{TP}{T}, \quad (4)$$

$$P = \frac{TP}{D}, \quad (5)$$

$$FPR = \frac{FP}{D}, \quad (6)$$

$$MDR = \frac{MD}{D}, \quad (7)$$

$$F1 = 2\frac{R \cdot P}{R+P}, \quad (8)$$

where T is the total number of fruits in the dataset, D is the number of detections, LD is the number of labels detected (annotations bounding boxes detected), TP is the number of true positives (detection with a ground truth overlap higher than 50%), FP is the number of false positives (detection with a ground truth overlap lower than 50%), and MD is the number of multi-detections produced when a single apple is detected multiple times.

3.1. 2D detection results

Table 2 presents instance segmentation results after training Mask R-CNN during 18 epochs (number of epochs not presenting overfitting). Results show an *AP* of 0.8599, and an F1-score of 0.8573. Although the best balance between *P* and *R* was achieved with a confidence threshold of 0.9, all detections classified as “apple” (confidence level > 0.5) where used for the 3D point cloud generation. This is because an increase of false positives (lower precision) is not as critical as decreasing the recall, since to build the 3D model an object has to be seen in, at least, two different images. Then, false positive objects that are only detected in one image will be automatically removed when applying SfM photogrammetry.

Table 2. Instance segmentation results at different confidence levels. Best F1-score result is in bold type.

Confidence	R	P	F1
0.5	0.8779	0.7622	0.8160
0.55	0.8746	0.7737	0.8211
0.6	0.8746	0.7840	0.8268
0.65	0.8729	0.7991	0.8344
0.7	0.8680	0.8117	0.8389
0.75	0.8663	0.8242	0.8447
0.8	0.8663	0.8333	0.8495
0.85	0.8647	0.8465	0.8555
0.9	**0.8597**	**0.8569**	**0.8583**
0.95	0.8399	0.8761	0.8576
AP	0.8599		

Fig. 7 shows 6 selected images from the validation dataset and the corresponding fruit detections, allowing a qualitative evaluation of instance segmentation results. As can be observed, most of the apples were successfully detected, including highly occluded or shadowed ones. In addition, Mask-RCNN masked correctly the pixels belonging to an apple, even when apples were visually split by branch or leaves, which is of interest to generate the 3D model of only apples when applying SfM. It was also observed that some of the detections reported as false positive were actually apples miss-annotated due to human error when labeling (pink rectangles in Fig. 7 b-d,f). Other false positives were wrong detections at the image borders, in parts of the image presenting a similar pattern to apples (red rectangles in Fig. 7 b-d), or multi-detections (blue rectangles in Fig. 7 a,e-f). As for the apples not detected, it can be seen that false negatives (yellow rectangles in Fig. 7 a-b,e) were apples cut at the image borders, highly occluded and/or small apples. To overcome the increase of false positives and negatives at image borders, a certain overlap between sub-images was considered when splitting the original image into sub-images (Section 2.2.1). Thus, detection failures at image borders did not affect the performance of the 3D model.

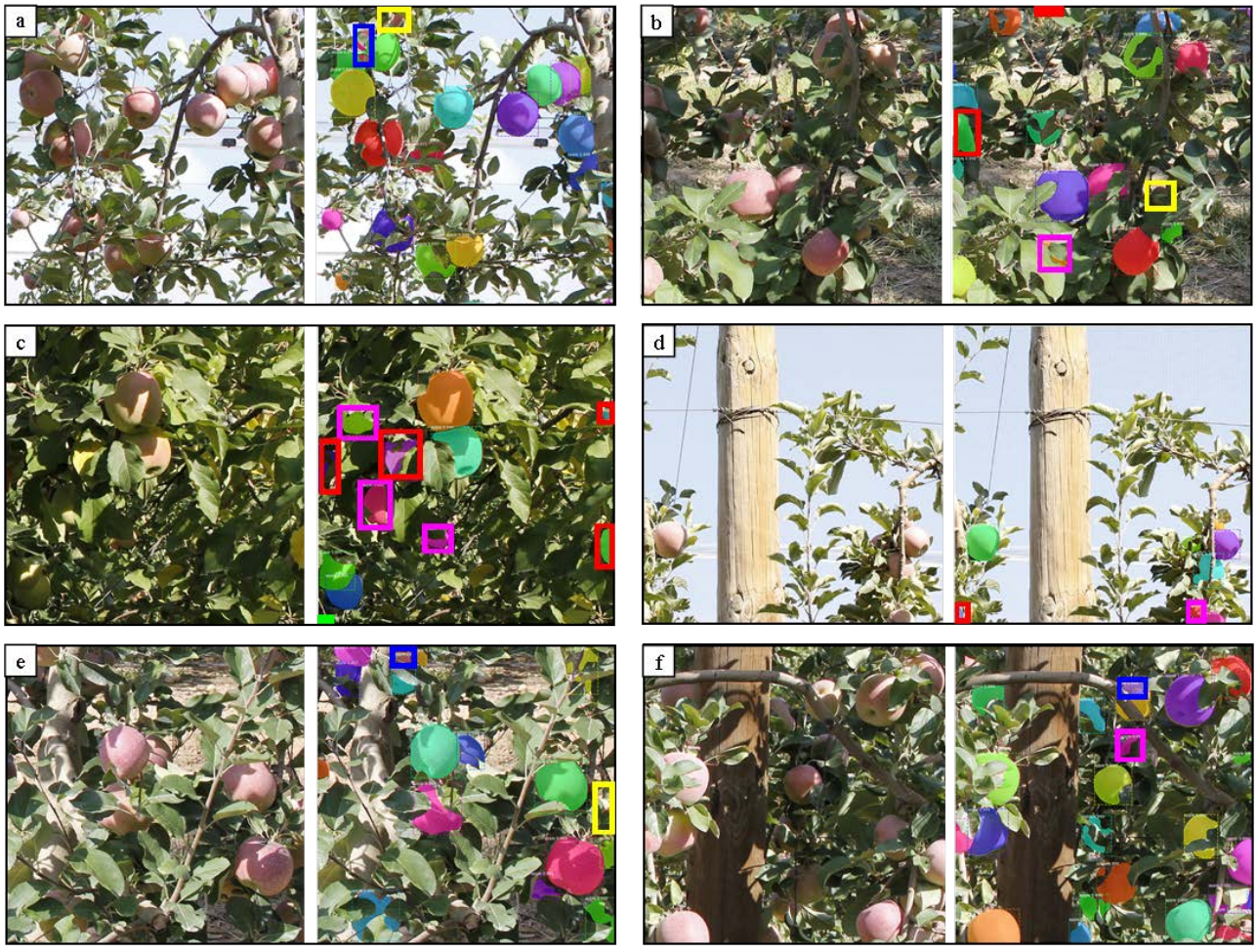

Fig. 7. Selected examples of instance segmentation results to show correct detections (colour masks), false positives due to network failures (red rectangles), false positives due to miss-annotated apples (pink rectangles), false positives due to multi-detections (blue rectangles), and false negatives (yellow rectangles). For each capture, the original sub-image (left) and the corresponding detections (right) are shown.

3.2. 3D location results

This section evaluates quantitatively and qualitatively the performance of the proposed methodology for 3D fruit detection and location. **Table 3** presents the detection rates achieved in the training (3 trees, 434 apples) and test (8 trees, 1021 apples) datasets. Results show a high detection rate (DR=0.991) with low false detections (FDR=0.037). However, because some apples were clustered in a unique detection (as shown in **Fig. 9**) and due to the presence of multi-detections (MDR=0.106), the recall and precision decreased to 0.906 and 0.857, respectively, which represents an F1-score of 0.881.

Table 3. 3D fruit detection and location results from training and test datasets.

	DR	R	P	FDR	MDR	F1-score
Training dataset	0.984	0.905	0.881	0.038	0.081	0.893
Test dataset	0.991	0.906	0.857	0.037	0.106	0.881

For yield prediction, the percentage of detected fruits and false positives is not as important as having a high correlation between the number of detections ($D = TP + FP + MD$) and the actual number of fruits in the trees (T) (Linker, 2017). Fig. 8 illustrates the correspondence between D and T in all trees of the dataset (11 trees). Results show the existence of a linear correlation between these variables, presenting a coefficient of determination of R^2=0.80 and a root mean square deviation of 6.42% of fruits.

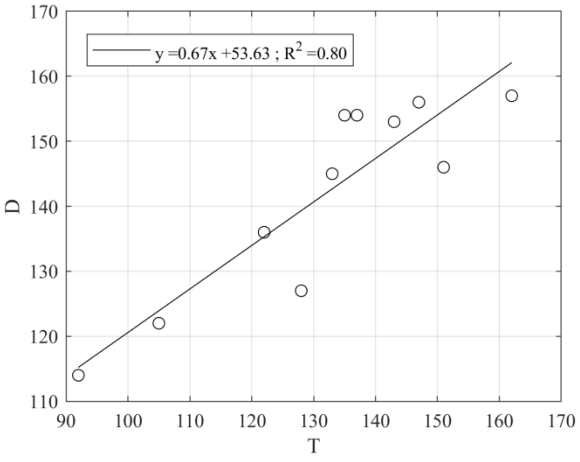

Fig. 8. Linear regression between the number of detections (D) and the actual number of fruits per tree (T).

For a qualitative evaluation, the reader is referred to inspect an interactive 3D visualization of the test scene and the corresponding fruit detections by opening the following link in a web-browser: http://www.grap.udl.cat/documents/photogrammetry_fruit_detection.html. Using the side menu, the reader can either visualize the scanned scene, the 3D point cloud of the apples obtained using SfM with masked images, or the apple detections obtained after 2D-3D projection and false positive removal steps.

The obtained point cloud showed higher 3D data precision compared with data provided by other sensors used for fruit detection, such as LiDAR or depth-cameras (Gené-Mola et al., 2019a, 2019b; Gongal et al., 2016; Nguyen et al., 2016; Tao and Zhou, 2017; Williams et al., 2019). Moreover, most of the apples were correctly detected, identifying the 3D points that belong to each apple. The presence of false positives is almost non-existent (FDR=0.037), while most of the multi-detections appeared in apples seen from both sides of the row of trees, when the detection from one side did not overlap sufficiently (they were not unified) with the detection from the other tree side. In contrast, as shown in Fig. 9, some groups of apples were unified in a single detection, which explains the difference between the detection rate and the recall values reported in Table 3. This is because when two apples were detected in a single detection, only one true positive is counted to compute the recall metric.

Fig. 9. Illustration of 3D fruit detection and location results from the test dataset: a) 3D visualisation of the scanned scene. b) Test scene with coloured fruit detections. A zoom view is shown to assist the visualization of the detections in the first tree of the dataset. Black circles show two examples where two apples were unified in a single detection. The reader is referred to the following link for an interactive 3D visualization of test fruit detection results: http://www.grap.udl.cat/documents/photogrammetry_fruit_detection.html

4. Discussion

This paper proposes a combination of instance segmentation neural networks and SfM for fruit detection and 3D location. By projecting 2D segmentation masks onto the 3D point cloud, results showed an increase of 2.8% in recall (from 0.878 to 0.906), 9.5% in precision (from 0.762 to 0.857) and 6.5% in F1-score (from 0.816 to 0.881). This difference could be even larger because 2D instance segmentation results were evaluated with respect to the number of visible fruits in the images –since it was not possible to estimate the number of occluded fruits in the 2D images–, while the 3D fruit detection was evaluated with respect to the total number of fruits in the tree. The use of SfM helped to increase the detection rate because of the multi-view approach of this technique. As stated by Hemming et al. (2014), due to the unstructured environment of orchards most fruits are partially/fully occluded from a single viewpoint, and thus multi-view imaging increases fruit detectability. When using multi-view imaging, an image registration is necessary to not double-count apples appearing in different images. In this work, this registration was automatically done by projecting 2D detections onto the 3D point cloud; even so, results showed a 10.6% multi-detection rate. Other authors have proposed similar approaches: Gongal et al. (2016) reported an error of 21.1% when identifying duplicate apples by projecting 2D image detections onto 3D models from RGB-D sensors, while Stein et al. (2016) used the 3D point cloud acquired from LiDAR-based sensors to identify multi-detections, although they did not assess the performance of this multi-detection identification. Using SfM not

only helped to increase the detection rate, but also decreased the number of false detections, because, to build the 3D point cloud, an object has to be detected in at least two different images, but the same false positive is not likely to be detected in two different images. Then, false positives only detected in one image were automatically removed. This fact, combined with the use of an SVM to identify false positives, explains the increase of 11.9% in precision, from 0.762 (2D image detections) to 0.881 (3D detections).

Although it is difficult to compare results from different datasets, our implementation of Mask R-CNN (F1-score=0.8583) performed similarly to other state-of-the-art fruit detection works based on deep convolutional neural networks, which reported F1-score values between 0.73 and 0.97 (Koirala et al., 2019). Mask R-CNN is not as fast as other object detection networks used for fruit detection – such as YOLO (Redmon and Farhadi, 2018; Tian et al., 2019) –, but it has the advantage of providing segmentation masks for each detection, which is necessary in our application to obtain the proper 3D location when projecting 2D detections onto the 3D point cloud. As for the 3D apple location performance, few works have provided 3D detection rates with respect to the total amount of fruits in trees. For instance, Stein et al. (2016) reported a good correlation (R^2=0.9) between the number of fruits detected and the actual number of fruits in the trees, but the methodology was not assessed in terms of precision, recall and F1-score (or similar metrics). Tao and Zhou (2017) reported a similar 3D detection performance to that of our methodology (F1-score = 0.921), but they tested the system on a smaller dataset of 59 apples. Finally, comparing the presented methodology with respect to other computer vision systems used in fruit harvesting robots, our system performed well compared to most of those presented in Bac et al. (2014) and Williams et al. (2019), which reported detection rates below 85%. However, the presented methodology is not suitable for harvesting robots because it cannot work at real-time due to the computationally-intensive processing of SfM (Wang et al., 2019). Nevertheless, the evolution of computing hardware and the development of efficient algorithms could overcome this limitation in the future.

Finally, from a qualitative/visual analysis of the 3D data, the point cloud obtained using SfM presented a higher precision compared with other sensors used for 3D fruit location, such as LiDAR-based and depth cameras (Gené-Mola et al., 2019a; Nguyen et al., 2016; Tao and Zhou, 2017). This suggests that the methodology could potentially be used to measure fruit size, which, combined with the good correlation between the number of fruit detections and the number of total fruits in the tree, would allow computation of fruit load in weight (yield estimation).

5. Conclusions

This work proposes the combination of instant segmentation neural networks and structure-from-motion (SfM) for apple detection and 3D location. Due to the multi-view approach on which SfM is based, results showed a small number of fruit occlusions compared with other fruit detection systems, reporting a detection rate of 99.1%. However, 8.5% of the apples were grouped in detections with more than one apple, with the result that the recall rate decreased to 0.906. Another advantage of using SfM was the reduction of false positives. Since SfM only generates the 3D model of those objects appearing in, at least, two different images, false positives only detected in one image were automatically discarded. This false positive reduction from SfM, combined with the use of a support vector machine to identify false positive detections, produced an increase in the precision metric from 0.762 (2D image detections) to 0.857 (3D detections). 3D location results reported an F1-score of 0.881 with respect to the total amount of fruit on the trees, with the conclusion that the proposed methodology performs well compared to other state-of-the-art 3D fruit location systems. The main disadvantage of this methodology is that, due to the computationally-intensive operations of SfM, it cannot process the data in real-time, which is an important limitation for its application in harvesting robots. However, the evolution of computing hardware and the development of efficient algorithms could overcome this issue in the future. The dataset and the corresponding annotations have been made publicly available, being the first dataset for 3D photogrammetric fruit detection and location. Due to the high spatial precision obtained with SfM and the good correlation between the number of detections and the actual number of fruits in the tree (R^2=0.8), future works should extend the methodology to measure fruit size and, consequently, perform fruit yield estimations.

Acknowledgements

This work was partly funded by the Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya (grant 2017 SGR 646), the Spanish Ministry of Economy and Competitiveness (project AGL2013-48297-C2-2-R) and the Spanish Ministry of Science, Innovation and Universities (project RTI2018-094222-B-I00). Part of the work was also developed within the framework of the project TEC2016-75976-R, financed by the Spanish Ministry of Economy, Industry and Competitiveness and the European Regional Development Fund (ERDF). The Spanish Ministry of Education is thanked for Mr. J. Gené's pre-doctoral fellowships (FPU15/03355). We would also like to thank Nufri (especially Santiago Salamero and Oriol Morreres) and Vicens Maquinària Agrícola S.A. for their support during data acquisition, and Ernesto Membrillo and Roberto Maturino for their support in dataset labelling.

Appendix A. Parameter values used for 3D point cloud generation

Table A1. Configuration set to perform the 3D reconstruction using Agisoft Professional Photoscan (v1.4, Agisoft LLC, St. Petersburg, Russia).

Step	Parameter	Configuration set	Description
Camera alignment	*Accuracy*	High	Images used in original size
	Key point limit	100000	Upper limit of feature points per image
	Tie point limit	10000	Upper limit of matching points per image
Dense cloud	*Quality*	Medium	Images downscaled by factor of 16 (4 times per side)
	Depth filtering	Mild	Filter used to sort out outliers

Appendix B. False positive feature analysis

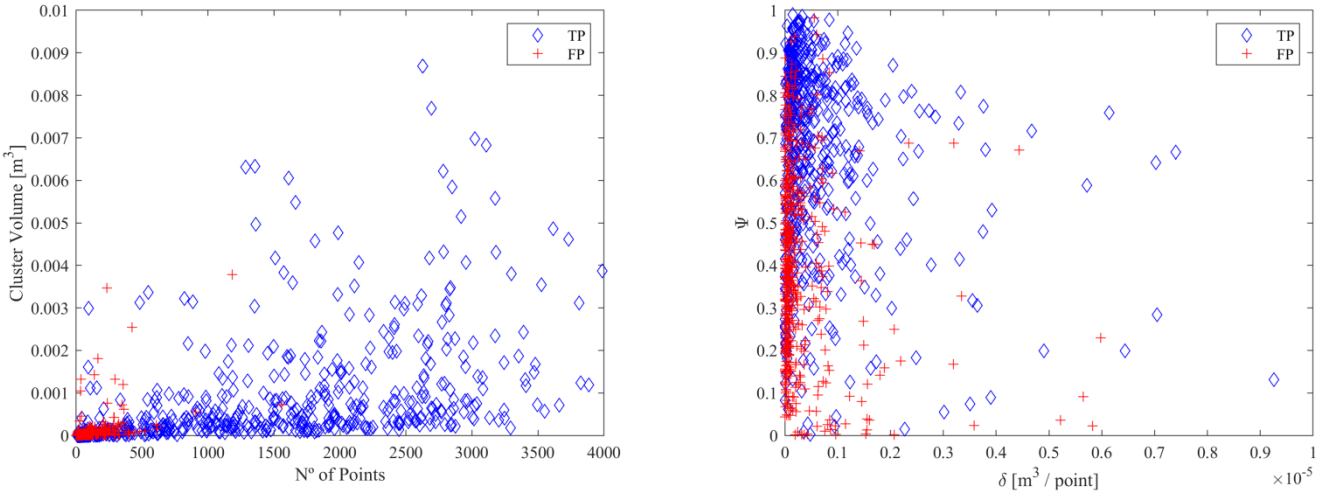

Fig. B 1 Graphical representation of apple detection features. The features analysed are the volume, number of points, the geometric parameter Ψ, and the detection point density δ. False positives are represented in red crosses; true positives are represented in blue diamonds. This analysis was performed on the training data set and was used to train the SVM for false positives identification (explained in **Section 2.2.3**).

References

Abdulla, W., 2017. Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow. GitHub Repos.

Bac, C.W., Van Henten, E.J., Hemming, J., Edan, Y., 2014. Harvesting Robots for High-value Crops: State-of-the-art Review and Challenges Ahead. J. F. Robot. 31. doi:10.1002/rob.21525

Bargoti, S., Underwood, J.P., 2017. Image Segmentation for Fruit Detection and Yield Estimation in Apple Orchards. J. F. Robot. 00, 1–22. doi:10.1002/rob.21699

Bulanon, D.M., Burks, T.F., Alchanatis, V., 2008. Study on temporal variation in citrus canopy using thermal imaging for citrus fruit detection. Biosyst. Eng. 101, 161–171. doi:10.1016/j.biosystemseng.2008.08.002

Ester, M., Kriegel, H.P., Sander, J., Xu, X., 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proc. 2nd Int. Conf. Knowl. Discov. Data Min. 96, 226–231. doi:10.1.1.71.1980

Gené-Mola, J., Gregorio, E., Guevara, J., Auat, F., Sanz-cortiella, R., Escolà, A., Llorens, J., Morros, J.-R., Ruiz-Hidalgo, J., Vilaplana, V., Rosell-Polo,

J.R., 2019a. Fruit detection in an apple orchard using a mobile terrestrial laser scanner. Biosyst. Eng. 187, 171–184. doi:10.1016/j.biosystemseng.2019.08.017

Gené-Mola, J., Vilaplana, V., Rosell-Polo, J.R., Morros, J.-R., Ruiz-Hidalgo, J., Gregorio, E., 2019b. KFuji RGB-DS database: Fuji apple multi-modal images for fruit detection with color, depth and range-corrected IR data. Data Br. doi:10.1016/j.dib.2019.104289

Gené-Mola, J., Vilaplana, V., Rosell-Polo, J.R., Morros, J.R., Ruiz-Hidalgo, J., Gregorio, E., 2019c. Multi-modal deep learning for Fuji apple detection using RGB-D cameras and their radiometric capabilities. Comput. Electron. Agric. 162, 689–698. doi:10.1016/j.compag.2019.05.016

Gongal, A., Amatya, S., Karkee, M., Zhang, Q., Lewis, K., 2015. Sensors and systems for fruit detection and localization: A review. Comput. Electron. Agric. 116, 8–19. doi:10.1016/j.compag.2015.05.021

Gongal, A., Silwal, A., Amatya, S., Karkee, M., Zhang, Q., Lewis, K., 2016. Apple crop-load estimation with over-the-row machine vision system. Comput. Electron. Agric. 120, 26–35. doi:10.1016/j.compag.2015.10.022

He, K., Gkioxari, G., Dollar, P., Girshick, R., 2017. Mask RCNN. Proc. IEEE Int. Conf. Comput. Vis. doi:10.1109/ICCV.2017.322

Hemming, J., Ruizendaal, J., Willem Hofstee, J., van Henten, E.J., 2014. Fruit detectability analysis for different camera positions in sweet-pepper. Sensors (Switzerland) 14, 6032–6044. doi:10.3390/s140406032

ISPA, (International Society of PrecisionAgriculture), 2019. ISPA Official Definition of Precision Agriculture. ISPA Newsl. 7 (7) July.

Koirala, A., Walsh, K.B., Wang, Z., McCarthy, C., 2019. Deep learning – Method overview and review of use for fruit detection and yield estimation. Comput. Electron. Agric. 162, 219–234. doi:10.1016/j.compag.2019.04.017

Kuhn, H.W., 2010. The Hungarian method for the assignment problem, in: 50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-Art. doi:10.1007/978-3-540-68279-0_2

Li, L., Zhang, Q., Huang, D., 2014. A review of imaging techniques for plant phenotyping. Sensors (Switzerland). doi:10.3390/s141120078

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. doi:10.1109/CVPR.2017.106

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common objects in context, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). doi:10.1007/978-3-319-10602-1_48

Linker, R., 2017. A procedure for estimating the number of green mature apples in night-time orchard images using light distribution and its application to yield estimation. Precis. Agric. 18, 59–75. doi:10.1007/s11119-016-9467-4

Liu, X., Chen, S.W., Aditya, S., Sivakumar, N., Dcunha, S., Qu, C., Taylor, C.J., Das, J., Kumar, V., 2018. Robust Fruit Counting: Combining Deep Learning, Tracking, and Structure from Motion. IEEE Int. Conf. Intell. Robot. Syst. 1045–1052. doi:10.1109/IROS.2018.8594239

Meier, U., 2001. Growth stages of mono- and dicotyledonous plants, BBCH Monograph. doi:10.5073/bbch0515

Narvaez, F.Y., Reina, G., Torres-Torriti, M., Kantor, G., Cheein, F.A., 2017. A survey of ranging and imaging techniques for precision agriculture phenotyping. IEEE/ASME Trans. Mechatronics 22, 2428–2439. doi:10.1109/TMECH.2017.2760866

Nguyen, T.T., Vandevoorde, K., Wouters, N., Kayacan, E., De Baerdemaeker, J.G., Saeys, W., 2016. Detection of red and bicoloured apples on tree with an RGB-D camera. Biosyst. Eng. 146, 33–44. doi:10.1016/j.biosystemseng.2016.01.007

Nuske, S., Wilshusen, K., Achar, S., Yoder, L., Singh, S., 2014. Automated visual yield estimation in vineyards, in: Journal of Field Robotics. doi:10.1002/rob.21541

Redmon, J., Farhadi, A., 2018. YOLOv3: An Incremental Improvement.

Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans. Pattern Anal. Mach. Intell. 39, 1137–1149. doi:10.1109/TPAMI.2016.2577031

Rosell-Polo, J.R., Cheein, F.A., Gregorio, E., Andújar, D., Puigdomènech, L., Masip, J., Escolà, A., 2015. Advances in Structured Light Sensors Applications in Precision Agriculture and Livestock Farming, in: Advances in Agronomy. doi:10.1016/bs.agron.2015.05.002

Stein, M., Bargoti, S., Underwood, J., 2016. Image Based Mango Fruit Detection, Localisation and Yield Estimation Using Multiple View Geometry. Sensors 16, 1915. doi:10.3390/s16111915

Tao, Y., Zhou, J., 2017. Automatic apple recognition based on the fusion of color and 3D feature for robotic fruit picking. Comput. Electron. Agric. 142, 388–396. doi:10.1016/j.compag.2017.09.019

Tian, Y., Yang, G., Wang, Z., Wang, H., Li, E., Liang, Z., 2019. Apple detection during different growth stages in orchards using the improved YOLO-V3 model. Comput. Electron. Agric. 157, 417–426. doi:10.1016/j.compag.2019.01.012

Torres-Sánchez, J., de Castro, A.I., Peña, J.M., Jiménez-Brenes, F.M., Arquero, O., Lovera, M., López-Granados, F., 2018. Mapping the 3D structure of almond trees using UAV acquired photogrammetric point clouds and object-based image analysis. Biosyst. Eng. doi:10.1016/j.biosystemseng.2018.10.018

Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W., 2000. Bundle Adjustment — A Modern Synthesis Vision Algorithms: Theory and Practice. Vis. Algorithms Theory Pract. 298–375. doi:10.1007/3-540-44480-7_21

Wang, X., Rottensteiner, F., Heipke, C., 2019. Structure from motion for ordered and unordered image sets based on random k-d forests and global pose estimation. ISPRS J. Photogramm. Remote Sens. 147, 19–41. doi:10.1016/j.isprsjprs.2018.11.009

Westoby, M.J., Brasington, J., Glasser, N.F., Hambrey, M.J., Reynolds, J.M., 2012. “Structure-from-Motion” photogrammetry: A low-cost, effective tool for geoscience applications. Geomorphology 179, 300–314. doi:10.1016/j.geomorph.2012.08.021

Williams, H.A.M., Jones, M.H., Nejati, M., Seabright, M.J., Bell, J., Penhall, N.D., Barnett, J.J., Duke, M.D., Scarfe, A.J., Seok, H., Lim, J., Macdonald, B.A., 2019. Robotic kiwifruit harvesting using machine vision , convolutional neural networks , and robotic arms. Biosyst. Eng. 181, 140–156. doi:10.1016/j.biosystemseng.2019.03.007

Zhang, E., Zhang, Y., 2009. Average Precision, in: LIU, L., ÖZSU, M.T. (Eds.), Encyclopedia of Database Systems. Springer US, Boston, MA, pp. 192–193. doi:10.1007/978-0-387-39940-9_482