Universitat de Lleida

Fruit detection and 3D location using optical sensors and computer vision

Jordi Gené Mola

http://hdl.handle.net/10803/669110

Universitat de Lleida

TESI DOCTORAL

Fruit detection and 3D location using optical sensors and computer vision

Jordi Gené Mola

Memòria presentada per optar al grau de Doctor per la Universitat de Lleida
Programa de Doctorat en Ciència i Tecnologia Agrària i Alimentaria

Directors
Eduard Gregorio López
Joan Ramon Rosell Polo

2020

Agraïments / Acknowledgements

En primer lloc, m'agradaria agrair als meus tutors, l'Eduard Gregorio i el Joan Ramon Rosell, per la seva dedicació i esforç durant aquests quatre anys. Sou i sereu un exemple a seguir, tant com a investigadors, com a professors i com a persones. Gràcies per donar-me espai a les baixades i una empenta a les pujades, per tindre sempre les portes sempre obertes, per la vostra ajuda incondicional i pel tracte rebut.

També vull agrair a la resta de membres del GRAP, per tractar-me com un més des del primer dia, i per la vostra predisposició en ajudar quan fes falta. En especial, als membres amb qui he compartit més moments, al Ricardo, l'Àlex, el Jordi, el Jaume, el Jose Antonio, el Francesc, el Lluís, el Manel, l'Asier, la Carla i el Xavier.

A Fernando y Darío, compañeros de la Universidad Técnica Federico Santa María (Valparaíso, Chile). Ha sido un placer poder trabajar con vosotros. Gracias por vuestra ayuda y por vuestra rigurosidad en el trabajo.

Al Ramón, el Javier i la Verónica. Gràcies per la vostra acollida durant la meva estada a la UPC (Barcelona), pels vostres consells i tot el que m'heu ensenyat. Els vostres coneixements en visió per computador han estat claus pel desenvolupament d'aquesta tesi.

To Dr. Jochen Hemming, for welcoming me during the three month stay in Wageningen University and Research (WUR). Thank you very much for your kindness and to give me the opportunity to contribute in the Trimbot project. It was an absolutely pleasure to work with you. Also to Dirk de Hoog and Manya for letting me the opportunity to collaborate in the Fruit 4.0 project.

També voldria agrair a Vicents Maquinara Agricola i a la Nufri, especialment al Santiago Salamero i l'Oriol Morreres, pel seu suport durant l'adquisició de dades, así como a Ernesto Membrillo y Roberto Maturino por su ayuda en el etiquetaje de datos.

Per últim, vull donar un agraïment especial als de casa: als meus pares, Mario i Maite, pel seu suport, temps, consells i valors que m'han transmès; a la meva germana, l'Anna, el meu nord, un exemple a seguir des de que era ben petit; i a la meva parella, la Júlia, pel seu amor, la seva energia i vitalitat, per fer i deixar fer, per creure i fer-me creure.

Moltes gràcies a tots i totes!

Institutional acknowledgements

The Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya and the Spanish Ministry of Education are thanked for the pre-doctoral fellowships 2016FI_B 00669 and FPU15/03355.

The research 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).

Resum

Per tal de satisfer les necessitats alimentàries d'una població mundial creixent, és necessari optimitzar la producció agrícola, incrementant la productivitat i la sostenibilitat de les explotacions. Per aconseguir-ho, es preveu que els sistemes automàtics de detecció i localització de fruits seran una eina essencial en la gestió de les plantacions fructícoles, amb aplicacions directes a la predicció de la collita, el mapat de la producció i la recol·lecció automatitzada. Malgrat els avenços aconseguits en àmbits com la robòtica o la visió per computador, la localització 3D de fruits continua essent un repte que ha de fer front a problemes com la identificació de fruits oclosos per altres òrgans vegetatius, o la possibilitat de treballar en diferents condicions d'il·luminació. La present tesi pretén contribuir en el desenvolupament de noves metodologies de detecció i localització de fruits mitjançant la combinació de sensors de base fotònica i d'algoritmes de visió artificial. Per tal de minimitzar els efectes produïts per unes condicions d'il·luminació variable, es proposa l'ús de sensors actius que treballen en l'espectre de llum infraroja. En concret, s'han testejat sensors LiDAR (*light detection and ranging*) i càmeres de profunditat (RGB-D) basades en el principi de temps de vol (*time-of-flight*), els quals proporcionen els valors d'intensitat de llum reflectida pels diferents elements mesurats. D'altra banda, per minimitzar el número d'oclusions s'han estudiat dues estratègies: (1) l'aplicació forçada d'aire; (2) la utilització de tècniques d'escaneig des de diferents punts de vista, com ara *Structure-from-Motion* (SfM). Els resultats obtinguts demostren que les dades d'intensitat proporcionades pels sensors actius LiDAR i RGB-D són de gran utilitat per la detecció de fruits, el que suposa un avanç en l'estat de l'art, ja que aquesta capacitat radiomètrica no havia estat estudiada anteriorment. D'altra banda, les dues estratègies testejades per minimitzar el número de fruits oclosos han demostrat incrementar el percentatge de fruits detectats. De totes les metodologies estudiades, la combinació de xarxes neuronals profundes amb tècniques de SfM és la que presenta més bons resultats, amb percentatges de detecció superiors al 90% i menys d'un 4% de falsos positius.

Resumen

Para satisfacer las necesidades alimentarias de una población mundial creciente, es necesario optimizar la producción agrícola, incrementando la productividad y la sostenibilidad de las explotaciones. Para conseguirlo, se prevé que los sistemas automáticos de detección y localización de frutos serán una herramienta esencial en la gestión de las plantaciones frutícolas, con aplicaciones directas a la predicción de cosecha, al mapeado de la producción y a la recolección automatizada. A pesar de los avances conseguidos en ámbitos como la robótica o la visión artificial, la localización 3D de frutos continua siendo un reto que debe de hacer frente a problemas como la identificación de frutos ocluidos por otros órganos vegetativos, o la posibilidad de trabajar en distintas condiciones de iluminación. La presente tesis pretende contribuir en el desarrollo de nuevas metodologías de detección y localizacion de frutos mediante la combinación de sensores de base fotónica y de algoritmos de visión artificial. A fin de minimizar los efectos producidos por unas condiciones de iluminación variable, se propone el uso de sensores activos que trabajan en espectros de luz infrarroja. En concreto, se han testeado sensores LiDAR (*light detection and ranging*) y cámaras de profundidad (RGB-D) basadas en el principio de tiempo de vuelo (*time-of-flight*), los cuales proporcionan valores de intensidad de la luz reflejada por los objetos escaneados. Por otra parte, para minimizar el número de oclusiones se han estudiado dos estrategias: (1) la aplicación forzada de aire; (2) la utilización de técnicas de escaneo desde distintas perspectivas, tales como *Structure-from-Motion* (SfM). Los resultados obtenidos demuestran que los datos de intensidad proporcionados por los sensores LiDAR y RGB-D son de gran utilidad para la detección de frutos, lo que supone un avance en el estado del arte, ya que esta capacidad radiométrica no había estado estudiada anteriormente. Por otra parte, las dos estrategias testeadas para minimizar el número de oclusiones han demostrado incrementar el porcentaje de detección. De todas las metodologías estudiadas, la combinación de redes neuronales profundas con técnicas de SfM es la que presenta mejores resultados, con porcentajes de detección superiores al 90% y con menos de un 4% de falsos positivos.

Summary

To meet the food demands of an increasing world population, farmers are required to optimize agriculture production by increasing crop productivity and sustainability. To do so, fruit detection and 3D location systems are expected to be an essential tool in the agricultural management of fruit orchards, with applications in fruit prediction, yield mapping, and automated harvesting. Despite the latest advances in robotics and computer vision, the development of a reliable system for 3D fruit location remains a pending issue to deal with problems such as the identification of occluded fruits and the variable lighting conditions of agricultural environments. The present thesis aims to contribute to the development of new methodologies for fruit detection and location by combining optical sensors and artificial intelligence algorithms. In order to minimize variable lighting effects, it is proposed the use of active sensors that work in the infrared light spectrum. In particular, light detection and ranging sensors (LiDAR) and depth cameras (RGB-D) based on the time-of-flight principle were evaluated. These sensors provide the amount of backscattered infrared light reflected by the measured objects. With respect to minimizing the number of fruit occlusions, two different approaches were tested: (1) the application of forced air flow; and (2) the use of multi-view scanning techniques, such as structure-from-motion (SfM) photogrammetry. The results have demonstrated the usefulness of the backscattered intensity provided by LiDAR and RGB-D sensors for fruit detection. This supposes an advance in the state-of-the-art, since this feature has not previously been exploited. Both of the strategies tested to minimize fruit occlusions showed an increase in the fruit detection rate. Of all the tested methodologies, the combination of instance segmentation neural networks and SfM photogrammetry gave the best results, reporting detection rates higher than 90% and false positive rates under 4%.

Nomenclature

NOTE: Abreviations used in Chapters III to VII (data and research articles) are defined in the corresponding chapters section and are not included in the following list.

B/W	Black and white
CCD	Charged Coupled Device
CMOS	Complementary Metal-Oxid-Semiconductor
D.Tree	Decision tree
GT_{field}	Number of fruits manually counted in the field
GT_{MTLS}	Fruits manually annotated in the MTLS point cloud.
GT_{SfM}	Fruits manually annotated in the SfM point cloud
HD	High Definition
IR	Infra-red
MTLS	Mobile Terrestrial Laser Scanner
PA	Precision agriculture
RGB	Read-green-blue
RGB-D	Red-green-blue-depth sensor
RMSE	Root mean square error
SfM	Structure-from-motion
SRS	Simple random sampling
Th	Threshold
ToF	Time-of-flight

Contents

List of figures

List of tables

Chapter I. Introduction

1. Background

To meet the food demands of a growing world population, farmers are required to optimize agronomic management and increase fruit production (Siegel et al., 2014). Additionally, rising farming costs, the lack of skilled labour and the need to reduce the environmental impact make it essential to find new strategies to increase the efficiency, quality and sustainability of agricultural activities (Tilman et al., 2011). To confront these challenges, precision agriculture (PA) is establishing itself as a cornerstone approach due to its capacity for gathering, processing and analysing temporal, spatial and individual data and combining them with other information to support management decisions based on the estimated variability (ISPA, 2019).

Advances in technological fields such as robotics and computer science have provided an opportunity to better understand orchard health and variability. New technologies have been used in PA to obtain a precise characterization of trees at different growth stages by non-destructive methods. This characterization can include phenology monitoring, plant geometric characterization and yield monitoring, among others. Remote fruit detection and 3D location is an active research field that combines sensing technologies and computer vision to characterize the distribution of fruits –within a specific tree or/and at plot level–. This detection and quantification of fruits distribution provides a valuable information to the farmers for the optimization of agricultural processes such as water irrigation, agrochemical applications, fertilization, pruning and thinning (Auat Cheein and Carelli, 2013; Bargoti and Underwood, 2017b). Despite the advances achieved in sensing and computer vision fields during the last decades, the development of high performance fruit detection and accurate 3D location systems is still a pending issue to deal with problems such as occlusions with other vegetative organs and variable lighting conditions.

2. State of the art

2.1 Sensors for fruit detection

Over the years, different sensors and systems have been used for fruit detection (Table 1). The earliest studies used black and white (B/W) cameras (Whittaker et al., 1987), but, since the emergence of colour cameras, RGB sensors based on Charged Coupled Devices (CCD) or Complementary Metal-Oxide-Semiconductors (CMOS) have been the most commonly used sensors for fruit detection (Linker, 2017; Maldonado and Barbosa, 2016; Zhao et al., 2016). These are affordable sensors, which allow the detection of fruits by using colour (Linker et al., 2012; Liu et al., 2016), geometric (Barnea et al., 2016; Lak et al., 2010) and texture (Chaivivatrakul and Dailey, 2014; Qureshi et al., 2017) features. The main disadvantages of RGB cameras are their sensitivity to lighting conditions and the fact that they only provide 2D information.

Other 2D sensors, such as multispectral, hyperspectral and thermal cameras, have allowed the exploration of non-visible wavelengths. Multi and hyperspectral cameras add spectral information beyond the RGB bands, allowing the use of a rich set of features and vegetation indexes for fruit detection (Okamoto and Lee, 2009; Safren et al., 2007; Zhang et al., 2015). Thermal cameras have been used to distinguish fruits from background based on their different thermal inertia. Their performance depends on the thermal evolution of the environment along the day, resulting in a narrow temporal range of operation (Bulanon et al., 2008; Stajnko et al., 2004). The main drawbacks of these sensors are their higher cost, the lack of 3D information and the higher level of training required for their operation (Linker, 2018).

The evolution in photonics has led to the introduction of 3D sensors, allowing the detection and subsequent 3D location of the fruit. The most commonly used 3D sensors are RGB-D (depth) cameras, which combine a colour and a depth sensor. Depending on the principle on which they are based, RGB-D cameras can be classified as stereovision, structured light, or time-of-flight (ToF) (Gongal et al., 2015; Vázquez-Arellano et al., 2016). Stereovision sensors (RGB-D_{Stereo}) incorporate two calibrated cameras and derive the depth of each pixel by applying triangulation (Figure 1a) (Font et al., 2014; Wang et al., 2017). Most of the RGB-D_{Stereo} are passive sensors and, in consequence,

present some issues in terms of finding pixel correspondence between images with low illuminated and/or low texture objects. Structured-light based sensors (RGB-$D_{Struct.Light}$) work on similar principles, but are not influenced by low illumination levels because the triangulation is carried out with an infra-red (IR) light pattern projected onto the scene and the image is acquired with an IR camera (Figure 1b) (Nguyen et al., 2016). Finally, depth cameras based on the ToF principle (RGB-D_{ToF}) measure distances to the objects by computing the time required by an IR light pulse to complete the round trip between the sensor and the scene (Figure 1c) (Barnea et al., 2016; Gongal et al., 2018; Li, 2014). The main disadvantage of all these sensors is that their performance decreases in high illuminated environments, such as direct sunlight.

Table 1. Photon-based sensors used for fruit detection.

	Sensors	Features	Advantages	Limitations
2D	B/W	-2D shape	-Low sensitivity to lighting conditions	-Lack of colour
	RGB	-Colour -2D shape	-Affordable	-High sensitivity to lighting conditions
	Thermal	-Temperature	-Independent of fruit colour	-Higher cost -Requires higher level of training
	Multispectral	-Colour -Spectral info. -2D shape	-Other spectral information beyond colour	-Higher cost -Requires higher level of training
	Hyperspectral	-Colour -Spectral info. -2D shape	-Spectral information in a wide range of bands	-Higher cost -Requires higher level of training
3D	RGB-D_{Stereo}	-Colour -3D shape	-3D + colour data -Affordable	-Dependence on scene illumination and texture -Computationally expensive
	RGB-$D_{Struct.Light}$	-Colour -3D shape	-3D + colour data -Fast data acquisition -Suitable for dark/low illumination conditions	-Susceptible to direct sunlight
	RGB-D_{ToF}	-Colour -3D shape -IR backscattering	-3D + colour + IR data -Suitable for dark/low illumination conditions	-Susceptible to direct sunlight
	LiDAR	-3D shape -IR backscattering	-3D + IR data -No dependence on illumination conditions	-High cost -Lack of colour -Requires higher level of training

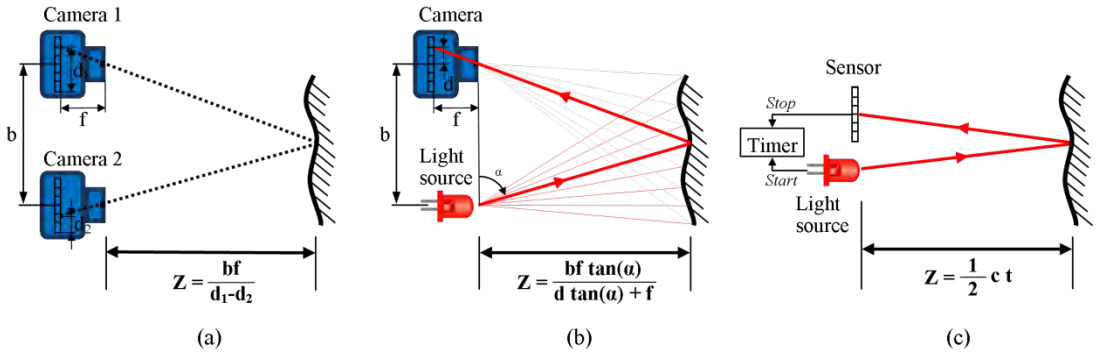


Figure 1. Schematic representation of the basic principles of 3D sensors: (a) stereo vision; (b) structured light; (c) time-of-flight.

Another technology based on the ToF principle is the one used in laser range finders and Light Detection and Ranging (LiDAR) systems. These are more expensive sensors which, since they operate with higher illumination energy sources (laser emitters), can measure longer distances than RGB-D sensors and are much less affected by sunlight. LiDAR sensors have been widely used for the geometric characterization of orchards (Rosell and Sanz, 2012; Vázquez-Arellano et al., 2016), but their use is marginal for fruit detection, probably because of the lack of colour data. Besides providing 3D data, one of the advantages of using ToF sensors is that they provide the amount of IR light backscattered by the scene, which is related to target reflectance after range correction and sensor calibration (Rodríguez-Gonzálvez et al., 2016). To the best of the author's knowledge, this capability of ToF sensors has not previously been exploited in fruit detection.

2.2 Algorithms and methods for fruit detection

Most of the data processing algorithms developed for fruit detection are based on handcrafted features (e.g. colour, texture, shape, intensity) that encode row data acquired with different sensors and use them to differentiate fruits from background by applying classification and clustering methods.

Colour features have been widely used, either in the RGB colour space or in other colour spaces less affected by the illuminance such as YCbCr or HSV (Maldonado and Barbosa, 2016; Teixidó et al., 2012). The main disadvantage of using colour features is

that they present a high sensitivity to the illumination conditions and are not effective in the detection of green fruit varieties, where the colour of the fruit is similar to the background. Texture features have also been applied for fruit detection. For instance, Rakun et al. (2011) used a spatial-frequency representation of images to differentiate fruits based on their texture in the image. Finally, shape features include edges, corners and blobs. The well-known Canny edge detector (Heath et al., 1998) has been used for edge detection and to subsequently find circles by applying the circular Hough Transform (Gongal et al., 2016). Harris, SIFT and SURF algorithms have also been used to extract corner and blob features and subsequently classify image regions as fruit or background (Chaivivatrakul and Dailey, 2014).

Clustering and classification algorithms allow the identification of regions of interest and the determination of their class. The simplest classification/segmentation method is thresholding (Zhou et al., 2012). However, this method presents some weakness when using sensors that are influenced by the acquisition conditions because the threshold values cannot be generalized with datasets acquired under different conditions. Other more robust and efficient algorithms used for fruit detection are K-means clustering, KNN clustering, Bayesian classifiers and Support Vector Machines (SVM) (Gongal et al., 2015).

More recently, remarkable progress has been achieved through the introduction of deep learning (Koirala et al., 2019). The deep neural networks used for fruit detection are Convolutional Neural Networks (CNN), which consist of a neural network where the neurons of each layer are organized in 3D matrices and the operation that connects two consecutive layers is based on convolutions. The use of CNNs has meant a breakthrough in computer vision, reporting similar performances to that of the human eye in tasks such as image classification, object detection and segmentation (Voulodimos et al., 2018). When using CNNs, there is no need to extract handcrafted features since they are automatically selected and extracted in the first convolutional layers. The main disadvantages of deep learning are the high amount of annotated data required and the computational intensive operations required to train training the networks, while the advantages are the high performance, the high inference speed and the fact that the features are automatically learned. The commonly used CNN

architectures for fruit detection are Faster R-CNN (Bargoti and Underwood, 2017a; Ren et al., 2017; Sa et al., 2016) and YOLO (Redmon and Farhadi, 2018; Tian et al., 2019).

2.3 Applications of fruit detection

Fruit detection systems have been applied in agriculture for yield prediction, yield mapping and automated harvesting. Yield prediction provides valuable information which enables the farmer to better plan the harvest campaign, fruit storage and marketing strategies (Bargoti and Underwood, 2017b; Nuske et al., 2014). Often, yield is predicted by manual counting of a few randomly selected samples. Although simple random sampling (SRS) is widely used for yield estimation, it may lead to inaccurate predictions if the number of selected samples is not large enough, which may be unfeasible by manual counting. This limitation can be overcome by using a fruit detection system that automatically counts the number of fruits in large sample sets. Yield mapping also provides valuable information to the farmer (Kurtulmus et al., 2014). The in-field spatial variability can be influenced by the agricultural management strategies, such as irrigation, fertilization and pruning, as well as the soil composition, the topographic characteristics or the impact of pests and diseases. The analysis of yield maps can help to find the less productive areas, figure out the reasons for this variability, and propose solutions. Finally, fruit detection and 3D location systems are important elements of harvesting robots. Hand harvesting is a hard and human-resource-intensive task that exposes farmers to dangerous conditions, working on ladders and platforms with heavy loads and under high temperatures (De-An et al., 2011; Gongal et al., 2015). The detection and location of fruits are key aspects in the development of efficient automated harvesting robots (Bac *et al*., 2014).

3. Objectives and hypothesis

The main objective of the present thesis is to contribute to the development of new methodologies for fruit detection and 3D location based on optical sensors and computer vision. To do so, this thesis aims to explore the capabilities of RGB, RGB-D and LiDAR sensors that have not been exploited before for fruit detection, and to develop new methodologies to take advantage of these capabilities, enhancing the potential of these sensors in fruit detection.

The research was based on the following hypotheses:

H1. Fruits have higher reflectance than background elements such as leaves and trunks.

H2. The number of fruits occluded by leaves can be reduced by applying forced air flow.

H3. The combination of multimodal data from RGB-D (colour, backscattered intensity, and depth) enhances the fruit detection rates.

H4. Fruits detected in 2D RGB images can be located in 3D by using structure-from-motion (SfM) photogrammetry.

These hypotheses were contrasted in order to fulfil the following specific objectives:

O1. Develop and test a methodology to detect and 3D locate fruits using LiDAR sensors.

O2. Develop and test a methodology to mitigate fruit occlusions.

O3. Study and analyse the potential of combining colour, IR backscattered and depth images provided by RGB-D_{ToF} sensors.

O4. Develop and test a methodology to detect and 3D locate fruits using RGB cameras and SfM photogrammetry.

4. Thesis structure

This PhD thesis includes seven papers: three data papers (Chapter III, papers P1, P2 and P3) and four research papers (Chapters IV-VII, papers P4, P5, P6, P7). Papers P2, P4, P5, P6, and P7 have already been published in SCI journals, while the other two have been submitted.

After this Introduction chapter, the rest of the thesis is structured as follows (Figure 2):

- **Chapter II** briefly sets out the methodology, explaining where the experimental tests were conducted, the sensors that were used, and how the data was processed.
- **Chapter III** includes the three data articles: P1 presents a dataset acquired with a Mobile Terrestrial Laser Scanner, which was used to test the methodologies described in the P4 and P5 research papers; P2 presents a dataset acquired with an RGB-D sensor and includes the annotated multi-modal images used in P6 for

fruit detection; finally, P3 presents a dataset acquired with RGB cameras, containing two sets of images used in P7, one for instance segmentation training and the other for 3D reconstruction using SfM photogrammetry.

- **Chapter IV** includes the research paper P4, which consists of a proof of concept of using LiDAR in detecting Fuji apples. This research was based on hypothesis H1 and fulfils the specific objective O1.
- **Chapter V** presents the research paper P5, which aims to tackle the problem of fruit occlusions by applying forced air flow. This research contrasts hypotheses H1 and H2, and meets the specific objectives O1 and O2.
- **Chapter VI** includes the research paper P6, which studies the usefulness of fusing RGB-D and range-corrected backscattered IR intensity for fruit detection. This research is based on hypotheses H1 and H3, and satisfies the objective O3.
- **Chapter VII** presents the research paper P7, which presents a new methodology for fruit detection and 3D location based on the combination of instance segmentation neural networks and SfM photogrammetry.
- **Chapter VIII** includes a general discussion of all the tested methods.
- **Chapter IX** presents the conclusions obtained from this research.
- Finally, **Chapter X** lists the PhD author contributions, including the refereed scientific papers, as well as other journal and conference contributions.

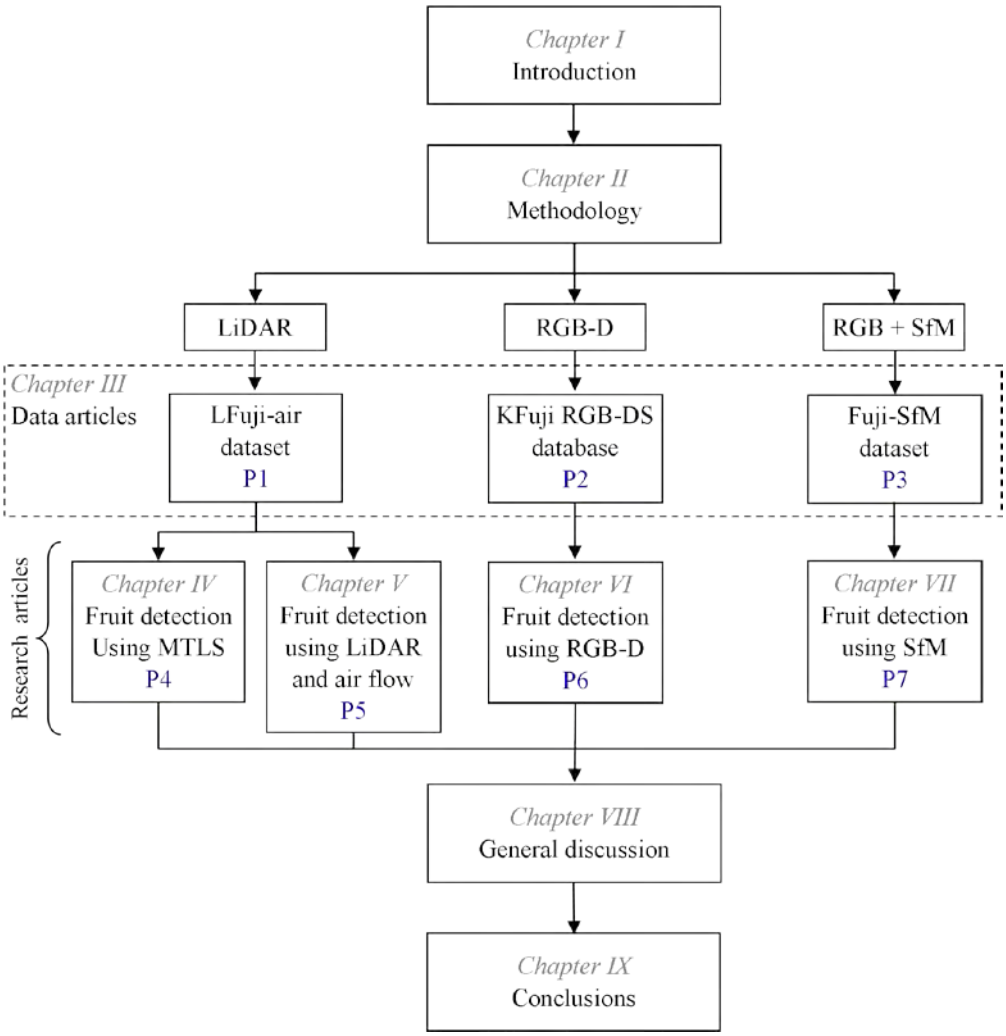


Figure 2. Thesis structure

References

Auat Cheein, F.A., Carelli, R., 2013. Agricultural robotics: Unmanned robotic service units in agricultural tasks. *IEEE Ind. Electron. Mag.* 7, 48–58. doi:10.1109/MIE.2013.2252957

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, 888–911. doi:10.1002/rob.21525

Bargoti, S., Underwood, J., 2017a. Deep Fruit Detection in Orchards. *2017 IEEE Int. Conf. Robot. Autom.* 3626–3633.. doi:10.1109/ICRA.2017.7989417

Bargoti, S., Underwood, J.P., 2017b. Image Segmentation for Fruit Detection and Yield Estimation in Apple Orchards. *J. F. Robot.* 34(6), 1039–1060. doi:10.1002/rob.21699

Barnea, E., Mairon, R., Ben-Shahar, O., 2016. Colour-agnostic shape-based 3D fruit detection for crop harvesting robots. *Biosyst. Eng.* 146, 57–70. doi:10.1016/j.biosystemseng.2016.01.013

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

Chaivivatrakul, S., Dailey, M.N., 2014. Texture-based fruit detection. *Precis. Agric.* 15, 662–683. doi:10.1007/s11119-014-9361-x

De-An, Z., Jidong, L., Wei, J., Ying, Z., Yu, C., 2011. Design and control of an apple harvesting robot. *Biosyst. Eng.* 110, 112–122. doi:10.1016/j.biosystemseng.2011.07.005

Font, D., Pallejà, T., Tresanchez, M., Runcan, D., Moreno, J., Martínez, D., Teixidó, M., Palacín, J., 2014. A proposal for automatic fruit harvesting by combining a low cost stereovision camera and a robotic arm. *Sensors (Switzerland)* 14, 11557–11579. doi:10.3390/s140711557

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., Karkee, M., Amatya, S., 2018. Apple fruit size estimation using a 3D machine vision system. *Inf. Process. Agric.* 5, 498–503. doi:10.1016/j.inpa.2018.06.002

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

Heath, M., Sarkar, S., Sanocki, T., Bowyer, K., 1998. Comparison of Edge Detectors: A Methodology and Initial Study. *Comput. Vis. Image Underst.* 69(1), 38-54. doi:10.1006/cviu.1997.0587

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

Kurtulmus, F., Lee, W.S., Vardar, A., 2014. Immature peach detection in colour images acquired in natural illumination conditions using statistical classifiers and neural network. *Precis. Agric.* 15, 57–79. doi:10.1007/s11119-013-9323-8

Lak, M.B., Minaei, S., Amiriparian, J., Beheshti, B., 2010. Apple fruits recognition under natural luminance using machine vision. *Adv. J. Food Sci. Technol.* 2, 325–327.

Li, L., 2014. Time-of-Flight Camera–An Introduction. *Texas Instruments* - Tech. White Pap.

Linker, R., 2018. Machine learning based analysis of night-time images for yield prediction in apple orchard. *Biosyst. Eng.* 167, 114–125. doi:10.1016/j.biosystemseng.2018.01.003

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

Linker, R., Cohen, O., Naor, A., 2012. Determination of the number of green apples in RGB images recorded in orchards. *Comput. Electron. Agric.* 81, 45–57.

doi:10.1016/j.compag.2011.11.007

Liu, X., Zhao, D., Jia, W., Ruan, C., Tang, S., Shen, T., 2016. A method of segmenting apples at night based on color and position information. *Comput. Electron. Agric.* 122, 118–123. doi:10.1016/j.compag.2016.01.023

Maldonado, W., Barbosa, J.C., 2016. Automatic green fruit counting in orange trees using digital images. *Comput. Electron. Agric.* 127, 572–581. doi:10.1016/j.compag.2016.07.023

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. *J. F. Robot.* 31(5), 837–860. doi:10.1002/rob.21541

Okamoto, H., Lee, W.S., 2009. Green citrus detection using hyperspectral imaging. *Comput. Electron. Agric.* 66, 201–208. doi:10.1016/j.compag.2009.02.004

Qureshi, W.S., Payne, A., Walsh, K.B., Linker, R., Cohen, O., Dailey, M.N., 2017. Machine vision for counting fruit on mango tree canopies. *Precis. Agric.* 18, 224–244. doi:10.1007/s11119-016-9458-5

Rakun, J., Stajnko, D., Zazula, D., 2011. Detecting fruits in natural scenes by using spatial-frequency based texture analysis and multiview geometry. *Comput. Electron. Agric.* 76, 80–88. doi:10.1016/j.compag.2011.01.007

Redmon, J., Farhadi, A., 2018. YOLOv3: An Incremental Improvement. *Tech Report*, arXiv:1804.02767.

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

Rodríguez-Gonzálvez, P., Gonzalez-Aguilera, D., González-Jorge, H., Hernández-López, D., 2016. Low-Cost Reflectance-Based Method for the Radiometric Calibration of Kinect 2. *IEEE Sens. J.* 16, 1975–1985. doi:10.1109/JSEN.2015.2508802

Rosell, J.R., Sanz, R., 2012. A review of methods and applications of the geometric characterization of tree crops in agricultural activities. *Comput. Electron. Agric.* 81, 124–141. doi:10.1016/j.compag.2011.09.007

Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T., McCool, C., 2016. DeepFruits: A Fruit Detection System Using Deep Neural Networks. *Sensors* 16, 1222. doi:10.3390/s16081222

Safren, O., Alchanatis, V., Ostrovsky, V., Levi, O., 2007. Detection of Green Apples in Hyperspectral Images of Apple-Tree Foliage Using machine Vision. *Trans. ASABE,* 50, 2303–2313. doi:10.13031/2013.24083

Siegel, K.R., Ali, M.K., Srinivasiah, A., Nugent, R.A., Narayan, K.M.V., 2014. Do we produce enough fruits and vegetables to meet global health need? *PLoS One*, 9 (8), e104059. doi:10.1371/journal.pone.0104059

Stajnko, D., Lakota, M., Hocevar, M., 2004. Estimation of number and diameter of apple fruits in an orchard during the growing season by thermal imaging. *Comput. Electron. Agric.* 42, 31–42. doi:10.1016/S0168-1699(03)00086-3

Teixidó, M., Font, D., Pallejà, T., Tresanchez, M., Nogués, M., Palacín, J., 2012. Definition of linear color models in the RGB vector color space to detect red peaches in orchard images taken under natural illumination. *Sensors (Switzerland)* 12, 7701–7718. doi:10.3390/s120607701

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

Tilman, D., Balzer, C., Hill, J., Befort, B.L., 2011. Global food demand and the sustainable intensification of agriculture. *Proc. Natl. Acad. Sci.* 108, 20260–20264. doi:10.1073/pnas.1116437108

Vázquez-Arellano, M., Griepentrog, H.W., Reiser, D., Paraforos, D.S., 2016. 3-D Imaging Systems for Agricultural Applications — A Review. *Sensors (Basel).* 16(5), 618. doi:10.3390/s16050618

Voulodimos, A., Doulamis, N., Doulamis, A., Protopapadakis, E., 2018. Deep Learning for Computer Vision: A Brief Review. *Comput. Intell. Neurosci.* 2018, 7068349. doi:10.1155/2018/7068349

Wang, Z., Walsh, K.B., Verma, B., 2017. On-Tree Mango Fruit Size Estimation Using RGB-D Images. *Sensors (Basel).* 17 (12), 2738. doi:10.3390/s17122738

Whittaker, A.D., Miles, G.E., Mitchell, O.R., Gaultney, L.D., 1987. Fruit location in a partially occluded image. *Trans. Am. Soc. Agric. Eng.* 30(3), 591–0596. doi:10.13031/2013.30444

Zhang, B., Huang, W., Wang, C., Gong, L., Zhao, C., Liu, C., Huang, D., 2015. Computer vision recognition of stem and calyx in apples using near-infrared linear-array structured light and 3D reconstruction. *Biosyst. Eng.* 139, 25–34. doi:10.1016/j.biosystemseng.2015.07.011

Zhao, C., Lee, W.S., He, D., 2016. Immature green citrus detection based on colour feature and sum of absolute transformed difference (SATD) using colour images in the citrus grove. *Comput. Electron. Agric.* 124, 243–253. doi:10.1016/j.compag.2016.04.009

Zhou, R., Damerow, L., Sun, Y., Blanke, M.M., 2012. Using colour features of cv. “Gala” apple fruits in an orchard in image processing to predict yield. *Precis. Agric.* 13, 568–580. doi:10.1007/s11119-012-9269-2

Chapter II. Methodology

All data used in this thesis was acquired in a commercial Fuji apple orchard (*Malus domestica* Borkh. cv. Fuji), located in Agramunt, Catalonia, Spain (E: 336,297 m; N: 4,623,494 m; 312 m a.s.l., UTM 31T - ETRS89). Trials were carried out in September 2017, three weeks before harvesting, corresponding to the BBCH growth stage 85 – advanced ripening, increase in intensity of cultivar-specific colour– (Meier, 2001). Trees grown in the studied orchard were 8 years old, 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 equipment used for data acquisition were: a Mobile Terrestrial Laser Scanner (MTLS) (used in P1, P4 and P5) (Figure 1a); a mobile platform equipped with two Microsoft Kinect V2 sensors (Microsoft, Redmond, WA, USA) (used in P2 and P6) (Figure 1b); and a Canon colour camera with a CMOS APS-C sensor (Canon Inc. Tokyo, Japan) (used in P3 and P7).

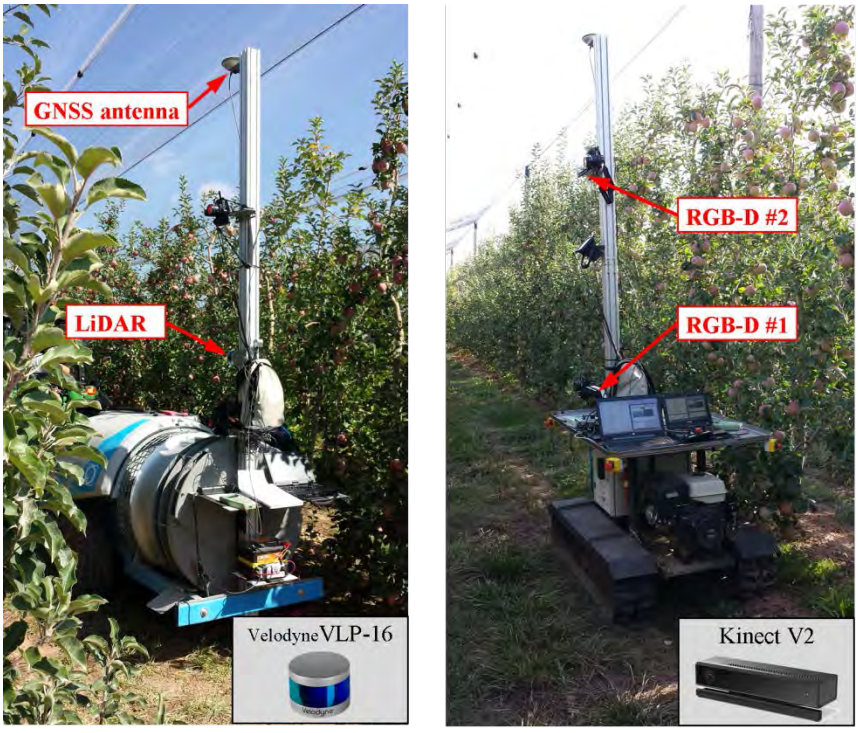


Figure 1. Equipment used for data acquisition. a) Mobile Terrestrial Laser Scanner mounted on an air-assisted sprayer. The LiDAR sensor used is illustrated in the bottom-right corner. b) Mobile platform equipped with two RGB-D sensors. The RGB-D sensor used is illustrated in the bottom-right corner.

The 2D data presented in data article P2 (and used in research article P6) include images of different randomly selected trees, while the 3D data provided in data articles P1 and P3 (and used in research articles P4, P5 and P7) contain information from 11 consecutive trees. A total of five scanning passes were carried out on each side of the 3D measured trees: One pass using SfM photogrammetry and four passes with an MTLS. Each MTLS measurement corresponded to a different sensor height (two heights tested) and to a different wind condition (with and without applying forced air flow). The number of apples produced by this set of 11 trees was manually counted in the field (ground truth field - GT_{field}). Additionally, the 3D point clouds acquired with the MTLS and with SfM photogrammetry were manually annotated (GT_{MTLS} and GT_{SfM}, respectively), placing 3D rectangular bounding boxes around each apple. As shown in Table 1, the number of apples manually counted in the orchard GT_{field} differs from the number of GT_{MTLS} and GT_{SfM} annotations. These differences can be attributed to human error during fruit counting or to fruits that were not visible in the 3D point clouds. The reader is referred to Chapter III for a more extensive explanation of data acquisition and dataset generation.

Table 1. Fruit counting ground truth. Comparison between the number of fruits manually counted in the orchard (GT_{field}), the number of fruits annotated in the MTLS point cloud (GT_{MTLS}) and the number of fruits annotated in the SfM point cloud (GT_{SfM}).

	GT_{field}	GT_{MTLS}	GT_{SfM}
Tree 01	139	138	147
Tree 02	106	100	105
Tree 03	139	131	135
Tree 04	137	129	137
Tree 05	94	85	92
Tree 06	131	119	133
Tree 07	119	114	122
Tree 08	145	137	151
Tree 09	139	131	143
Tree 10	136	122	128
Tree 11	159	147	162
Total	1444	1353	1455

Regarding the data processing, the algorithms tested for fruit detection in LiDAR point clouds (Chapters IV and V) were based on a reflectance threshold followed by other classification and clustering methods such as K-means (Jain, 2010) and SVM (Burges,

1998), while the algorithms tested for fruit detection in RGB-D and RGB images (Chapters VI and VII) were based on the deep neural networks Faster R-CNN (Ren et al., 2017) and Mask R-CNN (He et al., 2017), respectively. A further description of the fruit detection methodologies tested in this thesis is included in the corresponding research papers (Chapters IV to VII).

References

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* 2, 121–167. doi:10.1023/A:1009715923555

He, K., Gkioxari, G., Dollar, P., Girshick, R., 2017. Mask RCNN. *Proc. IEEE Int. Conf. Comput. Vis.* 2017, 2961–2969. doi:10.1109/ICCV.2017.322

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* 31 (8), 651–666. doi:10.1016/j.patrec.2009.09.011

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

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

Chapter III. Data articles

Chapter III.A. P1: LFuji-air dataset

Section submitted for publication in *Data in Brief*

LFuji-air dataset: annotated 3D LiDAR point clouds of Fuji apple trees for fruit detection scanned under different forced air flow conditions.

Jordi Gené-Mola[1], Eduard Gregorio[1], Fernando Auat Cheein[2], Javier Guevara[2], Jordi Llorens[1], Ricardo Sanz-Cortiella[1], Alexandre Escolà[1], Joan R. Rosell-Polo[1]

[1.] Research Group in AgroICT & Precision Agriculture, Department of Agricultural and Forest Engineering, Universitat de Lleida (UdL) – Agrotecnio Center, Lleida, Catalonia, Spain.
[2.] Department of Electronic Engineering, Universidad Técnica Federico Santa María, Valparaíso, Chile.

Abstract

This article presents the LFuji-air dataset, which contains LiDAR based point clouds of 11 Fuji apples trees and the corresponding apples location ground truth. A mobile terrestrial laser scanner (MTLS) comprised of a LiDAR sensor and a real-time kinematics global navigation satellite system was used to acquire the data. The MTLS was mounted on an air-assisted sprayer used to generate different air flow conditions. A total of 8 scans per tree were performed, including scans from different LiDAR sensor positions (multi-view approach) and under different air flow conditions. These variability of the scanning conditions allows to use the LFuji-air dataset not only for training and testing new fruit detection algorithms, but also to study the usefullness of the multi-view approach and the application of forced air flow to reduce the number of fruit oclusions. The data provided in this article is related to the research article entitled "Fruit detection, yield prediction and canopy geometric characterization using LiDAR with forced air flow" (Gené-Mola et al., 2019a).

Keywords: Fruit detection; Fruit location; Yield prediction; LiDAR; MTLS; Fruit reflectance; Forced air flow

Specifications Table

Subject	Agronomy and Crop Science, Horticulture, Computer Vision and Pattern Recognition
Specific subject area	Precision Agriculture, Fruit Detection
Type of data	LiDAR based point clouds Fruit location annotations
How data were acquired	Data was acquired with a Mobile Terrestrial Laser Scanner (MTLS) comprised of a LiDAR Sensor and a real-time kinematics global navigation satellite system (RTK-GNSS).
Data format	Raw LiDAR data: *PCAP* Raw RTK-GNSS data: *TXT* 3D point clouds: *MAT* Annotations: *TXT*
Parameters for data collection	The MTLS forward speed was set to 0.125 m/s. The LiDAR sensor acquired data at a frequency of 10 Hz, while the RTK-GNSS sensor provided positioning measurements with a precision of ± 0.01/0.02 m (horizontal / vertical) at 20 Hz frequency rate. The system was mounted on an air-assisted sprayer which generated an air flow speed of 5.5 ± 2.3 m s^{-1} (measured at 2.4 m from the sprayer fan).
Description of data collection	A MTLS was used to scan 11 Fuji apple trees containing a total of 1444 apples. The MTLS was mounted on an air-assisted sprayer used to generate different air flow conditions. A total of 8 different scanning conditions were tested, corresponding to the following combinations: two different sensor positions (1.8m and 2.5m height); two different air flow conditions (sprayer fan switched on and off) ; scans from the two sides of the row of trees (East and West). The ground truth of the apples locations was manually generated by placing 3D rectangular bounding boxes around each apple position.
Data source location	City/Town/Region: *Agramunt, Catalonia* Country: *Spain* GPS coordinates for collected data: *E: 336297 m, N: 4623494 m, 312 m a.s.l., UTM 31T - ETRS89*
Data accessibility	Repository name: *GRAP datasets / Lfuji-air dataset* Data identification: *Lfuji-air dataset* Direct URL to data: http://www.grap.udl.cat/en/publications/datasets.html
Related research article	Gené-Mola J, Gregorio E, Auat Cheein F, Guevara J, Llorens J, Sanz-Cortiella R, Escolà A, Rosell-Polo JR. Fruit detection, yield prediction and canopy geometric characterization using LiDAR with forced air flow *Computers and Electronics in Agriculture* (In press). https://doi.org/10.1016/j.compag.2019.105121

Value of the data

- First dataset for fruit detection containing annotated LiDAR based 3D data acquired from different sensor positions and under different air flow conditions.
- The dataset allows testing fruit detection algorithms based on LiDAR based 3D data.
- Precision horticulture community can benefit from these data to test methodologies with applications in yield prediction, yield mapping and canopy geometric characterization.
- Presented data can be used for analysing the effect of applying forced air flow and multi-view sensing for reducing the number of occlusions in fruit detection.

1. Data description

1.1. Data repository

The repository Lfuji-air dataset (http://www.grap.udl.cat/en/publications/LFuji_air_dataset.html) includes 3D LiDAR point clouds of 11 Fuji apple trees (*Malus domestica* Borkh. Cv. Fuji) containing 1444 apples (Figure 1). A total of 8 point clouds are provided for each tree, corresponding to the combinations of the following scanning conditions:

- *H1*: LiDAR sensor positioned at the height of 1.8 m from the floor. This LiDAR position corresponds (approximately) to the half of the tree heights.
- *H2*: LiDAR sensor positioned at the height of 2.5 m from the floor.
- *n*: Trees scanned without forced air flow application.
- *af*: Trees scanned under forced air flow conditions.
- *E*: Data acquired from the East side of the row of trees.
- *W*: Data acquired from the West side of the row of trees.

Point clouds were saved in MAT format. Each MAT file contains the data from one tree *#T#* (*01-11*), scanned with the LiDAR sensor at height *#H#* (*H1* or *H2*), under air flow conditions *#F#* (*n* or *af*), and from the side *#S#* (*E* or *W*). From that, the point clouds files are named as "Tree*#T#_#H#_#F#_#S#*.mat". For instance, the file "Tree07_H2_af_E.mat" contains the point cloud of Tree 7, obtained with the LiDAR sensor at height 2.5m (H2), by applying forced air flow (af), and scanned from the east (E) side. Data inside the MAT files is organized in an *m*4* matrix, where the three first columns give the position of the points in global world coordinates ($[X, Y, Z]_{<Global>}$), and last column corresponds to each point reflectance (*R*).

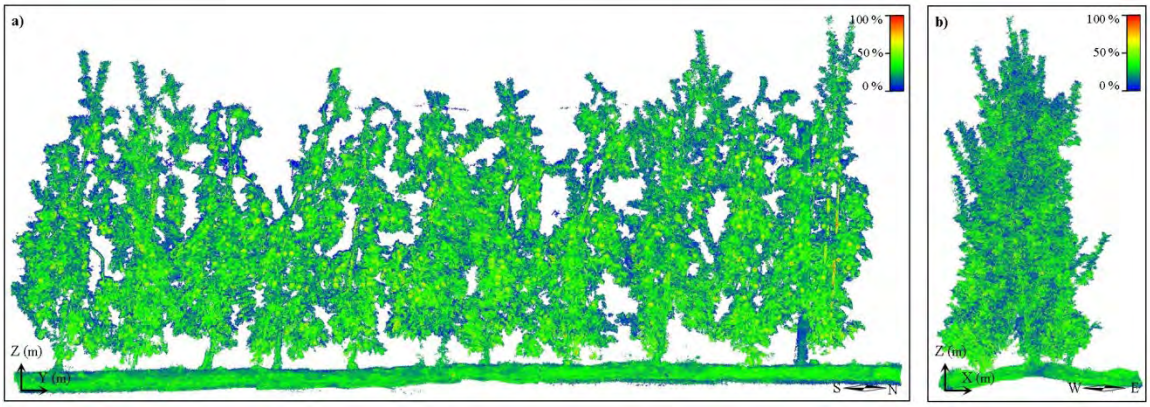


Figure 1. 3D point cloud of all trees included in the dataset (11 trees). a) Front elevation view. b) Left side elevation view. The color scale illustrates the points reflectance, ranging from 0 % (blue) to 100% (red). *X*, *Y*, and *Z* indicate the direction of the global axis, while *N*, *S, E,* and *W* represent the cardinal directions.

The dataset includes a total of 1353 apple annotations (out of 1444 apples manually counted in the field). The remaining 6.3% apples could not be identified in the point cloud because they were not visible (from a human/visual inspection). Annotations are provided in TXT format, where the first row indicates the position of the apple centre, while the following eight rows correspond to the positions of the bounding box corners.

Raw data used to generate the 3D point clouds is also provided in the dataset. This includes LiDAR data in PCAP format, and the positions of the real-time kinematics global navigation satellite system (RTK-GNSS) system in TXT format. Section 2 describes how data was acquired and processed to generate the described point clouds.

1.2. Code repository

The code used to process the row data and generate the georeferenced point clouds has been made publicly available at https://github.com/GRAP-UdL-AT/MTLS_point_cloud_generation. This Matlab code combines the LiDAR and RTK-GNSS raw data to obtain the 3D model of the measured trees. Section 2.3 describes the transformation matrices implemented in this code.

Additionally, the code used in [1] for fruit detection using the present dataset has also been made publicly available at https://github.com/GRAP-UdL-AT/fruit_detection_in_LiDAR_pointClouds. This code was developed to train and test the fruit detection algorithm as well as studying different sensor heights and air flow conditions to reduce the number of fruit occlusions. Both processing codes presented in

this section were implemented using MATLAB® (R2018a, Math Works Inc., Natick, Massachusetts, USA).

2. Materials and Methods

2.1. Experimental Design

Data was collected in a commercial Fuji apple orchard (*Malus domestica* Borkh. cv. Fuji). A total of 11 consecutive trees containing 1444 apples were scanned 3 weeks before harvesting, at 85 BBCH growth stage (Meier, 2001). The experimental setup used for data acquisition was a mobile terrestrial laser scanner (MTLS) comprised of a LiDAR sensor and a real-time kinematics global navigation satellite system (RTK-GNSS) (Figure 2). Both sensors were connected to a rugged laptop used to acquire and synchronise the data by means of the acquisition time.

The LiDAR sensor used was a Puck VLP-16 (Velodyne LIDAR Inc., San José, CA, USA), which generates a 3D point cloud of the scanned scene in the <*LiDAR*> coordinate system (Figure 2) with an accuracy of ±0.03 m (typical) at a frequency of 10 Hz (manually set). Additionally, this sensor provides the calibrated reflectance of each point (R) (Velodyne, 2016), which is a valuable information for fruit detection due to the different reflectance of apples and background (Gené-Mola et al., 2019c). The RTK-GNSS system used was a GPS1200+ (Leida Geosystems AG, Heerbrugg, Swizeland), which provides position measurements of the MTLS in <*Global*> world coordinates (Figure 2) at a frequency of 20 Hz with an absolute error of 0.01/0.02 m (horizontal / vertical). Further specifications of the LiDAR and RTK-GNSS sensors used are detailed in Table 1.

Table 1. Mobile terrestrial laser scanner set up specifications

LiDAR sensor	Manufacturer and model	Velodyne Puck VLP-16
	Number of laser beams	16
	Measurement Range	100 m
	Measurement accuracy	±30 mm
	Field of View (Horizontal // Vertical)	30º // 150º (manually set)
	Angular Resolution (Horizontal // Vertical)	2.0º // 0.2º
	Scan Rate	10 Hz (manually set)
	Wavelength	903 nm
RTK-GNSS	Manufacturer and model	Leica GPS1200+
	Measurement accuracy	20 mm
	Measurement Rate	20 Hz

The MTLS system was mounted on an air-assisted sprayer, next to the sprayer fan, which was used to generate forced air flow and move the tree foliage. The GNSS antenna was installed at a height of 3.5 m. The LiDAR sensor was mounted vertically, with the $Z_{<LiDAR>}$ axis pointing to the forward direction (Figure 2), and placed at heights of 1.8 m (*H1*) and 2.5 m (*H2*). The experimental setup was pulled by a tractor at 0.125 m s^{-1} forward speed and following a linear trajectory parallel to the row of trees.

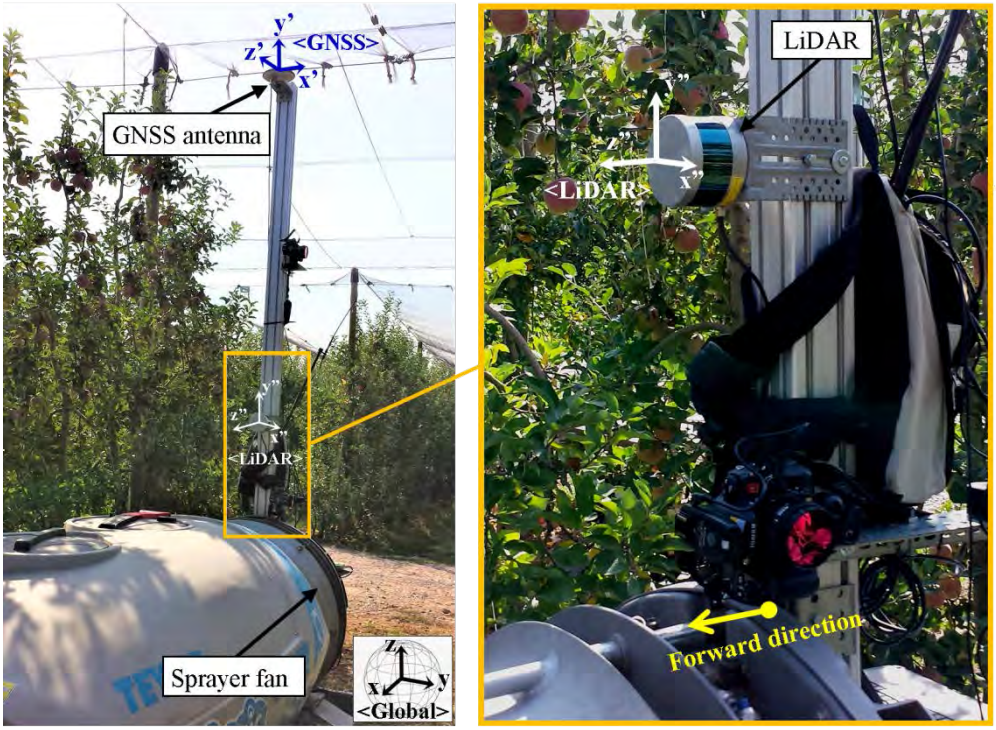


Figure 2. Global scheme of the experimental setup used for data acquisition. The orientation of <LiDAR>, <GNSS>, and <Global> coordinate systems used for point cloud generation are represented.

2.2. Sprayer fan characterization

In order to generate forced air flow, the air-assisted sprayer operated at 18π rad s^{-1} (540 rpm of PTO, power take-off angular speed). At these conditions, the air flow speed at different heights and widths was characterized using an AIRMAR 200WX weather station (AIRMAR Technology Corporation, Milford, NH, USA), which measures the wind speed with an accuracy of ±0.5 m s^{-1}. A total of 35 measurements from a distance of 2.4 m (distance between sprayer fan and scanned trees) were performed, corresponding to the measurement of 7 height and 5 width intervals (Figure 3). The 7 height intervals were equally distributed from 0 m to 3.5 m height, corresponding to the maximum trees height. On the other hand, the 5 width intervals were equally distributed

along 1.4 m width, which corresponds to the field-of-view of the LiDAR sensor. The speed values shown in Figure 3 are the result of averaging 10 measurements in each position.

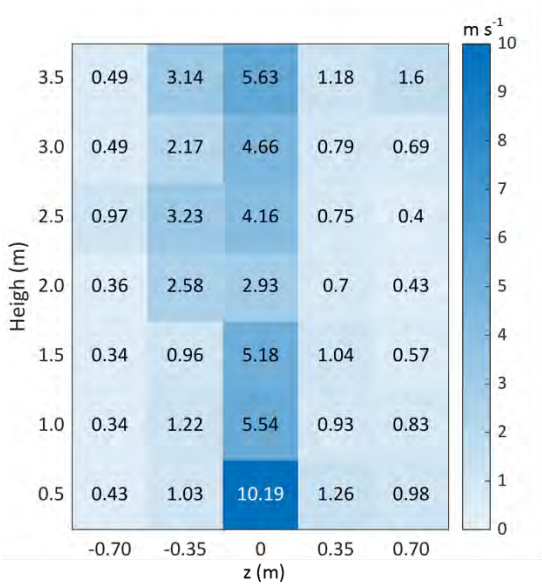


Figure 3. Air flow speed in $m\ s^{-1}$ at different heights and widths (z) measured at a distance of 2.4 m from the sprayer fan.

2.3. Point cloud generation

The LiDAR row data consist on a set of frames acquired from different positions, where each frame $P_{<LiDAR>}$ is a point cloud in the $<LiDAR>$ coordinate system:

$$P_{<LiDAR>} = \begin{bmatrix} x_1 & x_2 & \dots & x_n \\ y_1 & y_2 & \dots & y_n \\ z_1 & z_2 & \dots & z_n \\ 1 & 1 & \dots & 1 \end{bmatrix}, \quad (1)$$

where n denotes the number of points in the LiDAR frame.

For the generation of the 3D point cloud of all trees (Figure 1), each frame $P_{<LiDAR>}$ was transformed into $<Global>$ coordinates as follows:

$$P_{<Global>} = T_{<LiDAR>\rightarrow<Global>} \times P_{<LiDAR>}, \quad (2)$$

where the transformation matrix $T_{<LiDAR>\rightarrow<Global>}$ can be expanded as:

$$T_{<LiDAR>\rightarrow<Global>} = T_{<GNSS>\rightarrow<Global>} \times T_{<LiDAR>\rightarrow<GNSS>} \quad (3)$$

Because the LiDAR and the GNSS antenna were assembled in a rigid structure, the rigid transformation matrix $T_{<LiDAR>\rightarrow<GNSS>}$ only has a translational offset:

$$T_{<LiDAR>\rightarrow<GNSS>} = I|\Delta xyz_{H1<LiDAR><GNSS>} = \begin{bmatrix} 1 & 0 & 0 & \Delta x_{<LiDAR><GNSS>} \\ 0 & 1 & 0 & \Delta y_{<LiDAR><GNSS>} \\ 0 & 0 & 1 & \Delta z_{<LiDAR><GNSS>} \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad (4)$$

where $\tau = [\Delta x_{<LiDAR><GNSS>}, \Delta y_{<LiDAR><GNSS>}, \Delta z_{<LiDAR><GNSS>}]'$ denotes the offset between each axis of the GNSS and the LiDAR sensor. Considering the distribution of the sensors in the experimental setup, the translation offsets for the *H1* and *H2* trials were $\tau = [0\ , 1.768, 0.058]'\ m$ and $\tau = [0\ , 1.07, 0.058]'\ m$, respectively.

Meanwhile, the rigid transformation matrix $T_{<GNSS>\rightarrow<Global>}$ includes a rotational, $R_{<GNSS>}$, and a translational, $T_{<GNSS>}$, component. As depicted in Figure 2, the forward direction is $z'_{<GNSS>}$. Being θ and φ the orientation angles of the vehicle around the $y'_{<GNSS>}$(Yaw) and $x'_{<GNSS>}$ (Pitch) axes, respectively, the transformation matrix $T_{<GNSS>\rightarrow<Global>}$ is obtained according to:

$$T_{<GNSS>\rightarrow<Global>} = R_{<GNSS>}|T_{<GNSS>}$$

$$T_{<GNSS>\rightarrow<Global>} = \begin{bmatrix} \cos\theta & -\sin\theta\cos\varphi & \sin\theta\sin\varphi & X_{<GNSS>} \\ \sin\theta & \cos\theta\cos\varphi & -\cos\theta\sin\varphi & Y_{<GNSS>} \\ 0 & \sin\varphi & \cos\varphi & Z_{<GNSS>} \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (5)$$

where $X_{<GNSS>}, Y_{<GNSS>}$ and $Z_{<GNSS>}$ denote the position of the GNSS antenna in each axis of the global coordinate system. It is worth to mention that the experimental setup did not include an inertial measurement unit (IMU); therefore, the orientation angles θ y φ were obtained by using the forward direction computed from the measurements of the RTK-GNSS receiver. Since trials were conducted in short rectilinear trajectories, the orientation of the system was assumed to be constant along the path.

The resulting point clouds were manually split into a single point cloud per tree. Then each tree was manually annotated by placing 3D rectangular bounding boxes around each apple position. This process was carried out using the software CloudCompare (Cloud Compare [GPL software] v.9 Omnia) and supported by additional RGB images of the tested trees.

Acknowledgments
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). The Spanish Ministry of Education is thanked for Mr. J. Gené's pre-doctoral fellowships (FPU15/03355). The work of Jordi Llorens was supported by the Spanish Ministry of Economy, Industry and Competitiveness through a postdoctoral

position named Juan de la Cierva Incorporación (JDCI-2016-29464_N18003). We would also like to thank CONICYT FONDECYT 1171431 and CONICYT FB0008. Nufri (especially Santiago Salamero and Oriol Morreres) and Vicens Maquinària Agrícola S.A. are also thanked for their support during data acquisition, and Ernesto Membrillo and Roberto Maturino for their support in dataset labelling

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Gené-Mola, J., Gregorio, E., Auat Cheein, F., Guevara, J., Llorens, J., Sanz-Cortiellaa, R., Escolà, A., Rosell-Polo, J.R., 2019a. Fruit detection, yield prediction and canopy geometric characterization using LiDAR with forced air flow. *Computers and Electronics in Agriculture* (In press). https://doi.org/10.1016/j.compag.2019.105121

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., 2019b. 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., 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

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

Velodyne, L., 2016. VLP-16 In VLP-16 Manual: User's Manual and Programming Guide; Velodyne LiDAR.

Chapter III.B. P2: KFuji RGB-DS database

This section was published in *Data in Brief* 25 (2019) 104289, https://doi.org/10.1016/j.compag.2019.05.016:

Data Article

KFuji RGB-DS database: Fuji apple multi-modal images for fruit detection with color, depth and range-corrected IR data

Jordi Gené-Mola [a], Verónica Vilaplana [b], Joan R. Rosell-Polo [a], Josep-Ramon Morros [b], Javier Ruiz-Hidalgo [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

This article contains data related to the research article entitle “Multi-modal Deep Learning for Fruit Detection Using RGB-D Cameras and their Radiometric Capabilities” (Gené-Mola et al., 2019c). The development of reliable fruit detection and localization systems is essential for future sustainable agronomic management of high-value crops. RGB-D sensors have shown potential for fruit detection and localization since they provide 3D information with color data. However, the lack of substantial datasets is a barrier for exploiting the use of these sensors. This article presents the KFuji RGB-DS database which is composed by 967 multi-modal images of Fuji apples on trees captured using Microsoft Kinect v2 (Microsoft, Redmond, WA, USA). Each image contains information from 3 different modalities: color (RGB), depth (D) and range corrected IR intensity (S). Ground truth fruit locations were manually annotated, labeling a total of 12,839 apples in all the dataset. The current dataset is publicly available at http://www.grap.udl.cat/publicacions/datasets.html.

Keywords: Multi-modal dataset; Fruit detection; Depth cameras; RGB-D; Fruit reflectance; Fuji apple

Specifications Table

Subject	Machine learning, computer vision, deep learning, agronomy
Specific subject area	Image fusion, Precision agriculture.
Type of data	Multi-modal images with colour (RGB), depth (D), and range-corrected IR intensity (S).
How data were acquired	The images were acquired using Microsoft Kinect v2.
Data format	Raw images: *JPG* Raw point clouds: *MAT* Pre-processed images: *JPG* (colour channels) and *MAT* (depth and range-corrected IR channels) Annotations: *CSV* and *XLM*.
Experimental factors	Different image modalities have been registered to have pixel-wise correspondence between image channels.
Experimental features	All captures were carried out during the night, using artificial lighting.
Data source location	Data were acquired in Tarassó Farm, a commercial apple field located in Agramunt, Catalonia, Spain (E: 336297 m N: 4623494 m 31N 312 m a.s.l., UTM31T - ETRS89).
Data accessibility	http://www.grap.udl.cat/en/publications/datasets.html
Related research article	Gené-Mola J, Vilaplana V, Rosell-Polo J.R, Morros J.R, Ruiz-Hidalgo J, Gregorio E. Multi-modal Deep Learning for Fruit Detections Using RGB-D Cameras and their Radiometric Capabilites. *Computers and Electronics in Agriculture* (2018) 162, 689-698. DOI: 10.1016/j.compag.2019.05.016

Value of the data

- First dataset for fruit detection that contains 3 different modalities: color, depth and range corrected IR intensity.
- The presented dataset could be used in the development and training of fruit detection systems with applications in yield prediction, yield mapping and automated harvesting.
- Compilation of this database allows fusing RGB-D and radiometric information obtained with Kinect v2 for fruit detection.

1. Data

The KFuji RGB-DS database contains a total of 967 multi-modal images of Fuji apples on trees and the corresponding ground truth fruit location annotations. Each image contains data from three different modalities: color (RGB), depth (D), and range-corrected IR intensity (S). Figure 1 illustrates three selected images from de dataset, showing ground truth annotations and the modalities that composes each image.

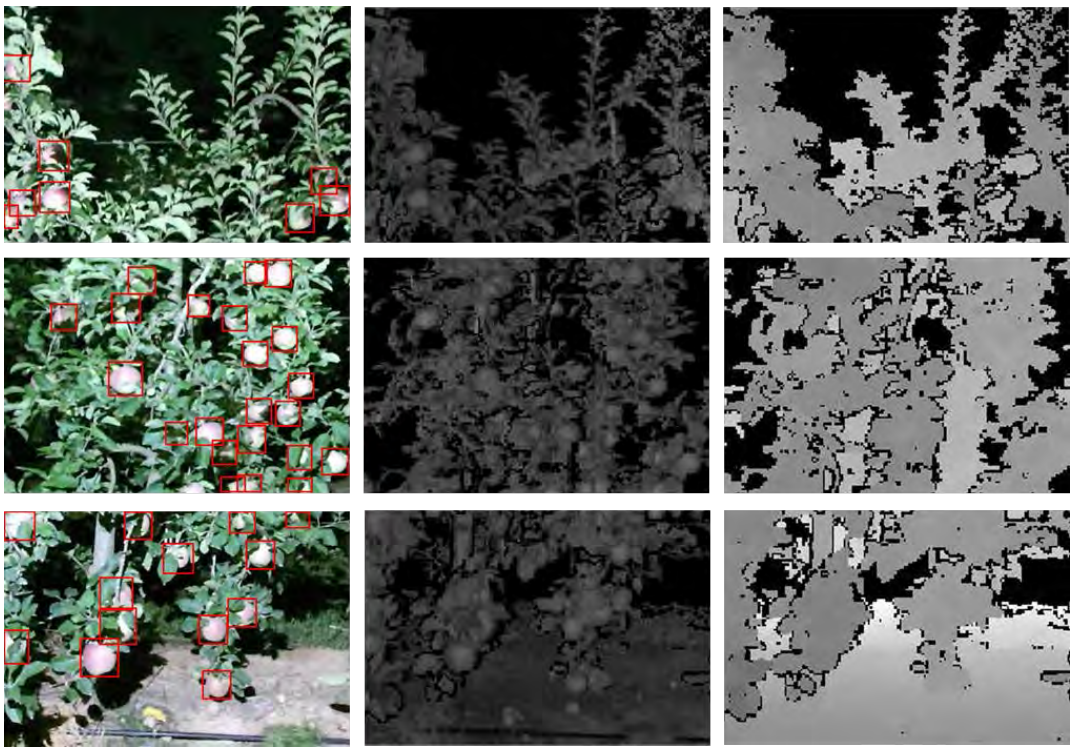


Figure 1. Selection of 3 multi-modal images and the corresponding ground truth fruit locations (red bounding boxes). Each image column corresponds to a different image modality: RGB, S and D, respectively.

This dataset was built to be used for training, validation and benchmarking of fruit detection algorithms using RGB-D sensors. For instance, in Gené-Mola et al. (2019), the deep convolutional neural network Faster R-CNN (Ren et al., 2017) was used to detect and localize fruits from the presented dataset.

Images are 548x373px and were saved in three different files:

- RGB_{hr} (high resolution color image): Raw color image. These images are saved in 8-bit JPG files.

- RGB_p (projected color image): Projection of the color 3D point cloud onto the camera focal plane. The RGB_p and the D-S modalities are obtained following the same procedure, allowing the comparison between these modalities for fruit detection. These images are saved in 8-bit JPG files.

- DS (depth and range-corrected IR image): Projection of the range-corrected IR 3D point cloud onto the camera focal plane. The D channel corresponds to the depth values, while the S channel corresponds to the range-corrected IR intensity values. These modalities are saved in a unique 64-bit MAT file.

S and D data were normalized between 0 and 255 –like RGB images- to achieve similar mean and variance between channels. This normalization allows a faster learning convergence of machine learning algorithms (such as deep convolutional neural networks).

All images were manually annotated with rectangular bounding boxes, labelling a total of 12,839 apples in all the dataset. Annotations are provided in XLM and CSV formats, where each row corresponds to an apple annotation, giving the following information: item, topleft-x, topleft-y, width, height, label id.

2. Experimental Design, Materials, and Methods

The data acquisition was carried out in a commercial Fuji apple orchard (Malus domestica Borkh. cv. Fuji), three weeks before harvesting -85 BBCH growth stage (Meier, 2001)-. The RGB-D sensors used were two Microsoft Kinect v2 (Microsoft, Redmond, WA, USA), which are composed by an RGB camera and a time-of-flight (ToF) depth sensor. For each capture, the sensor provides a 3D point cloud with RGB and backscattered IR intensity data, and a raw RGB image. Due to the performance of the depth sensor drops under direct sunlight exposure (Rosell-Polo et al., 2015), data was acquired at night using artificial lighting.

Pre-processing of data was carried out to build the multi-modal images with pixel-wise correspondence between channels. Figure 2 shows an outline of the data preparation steps. To overcome the IR signal attenuation, the IR intensity data was range-corrected (Figure 2a) following the methodology described in Gené-Mola et al. (2019). Then the

acquired 3D point clouds were projected onto the camera focal plane (Figure 2b), generating the RGB, range-corrected IR and depth projected images. These images were geometrically wrapped and registered Figure 2c) with RGB_{hr} so that different image modalities have pixel-wise correspondence. Finally, to reduce the number of fruits per image, and considering that fruit size is small compared with the image size, each capture was split into 9 images of 548 x 373 px (Figure 2d).

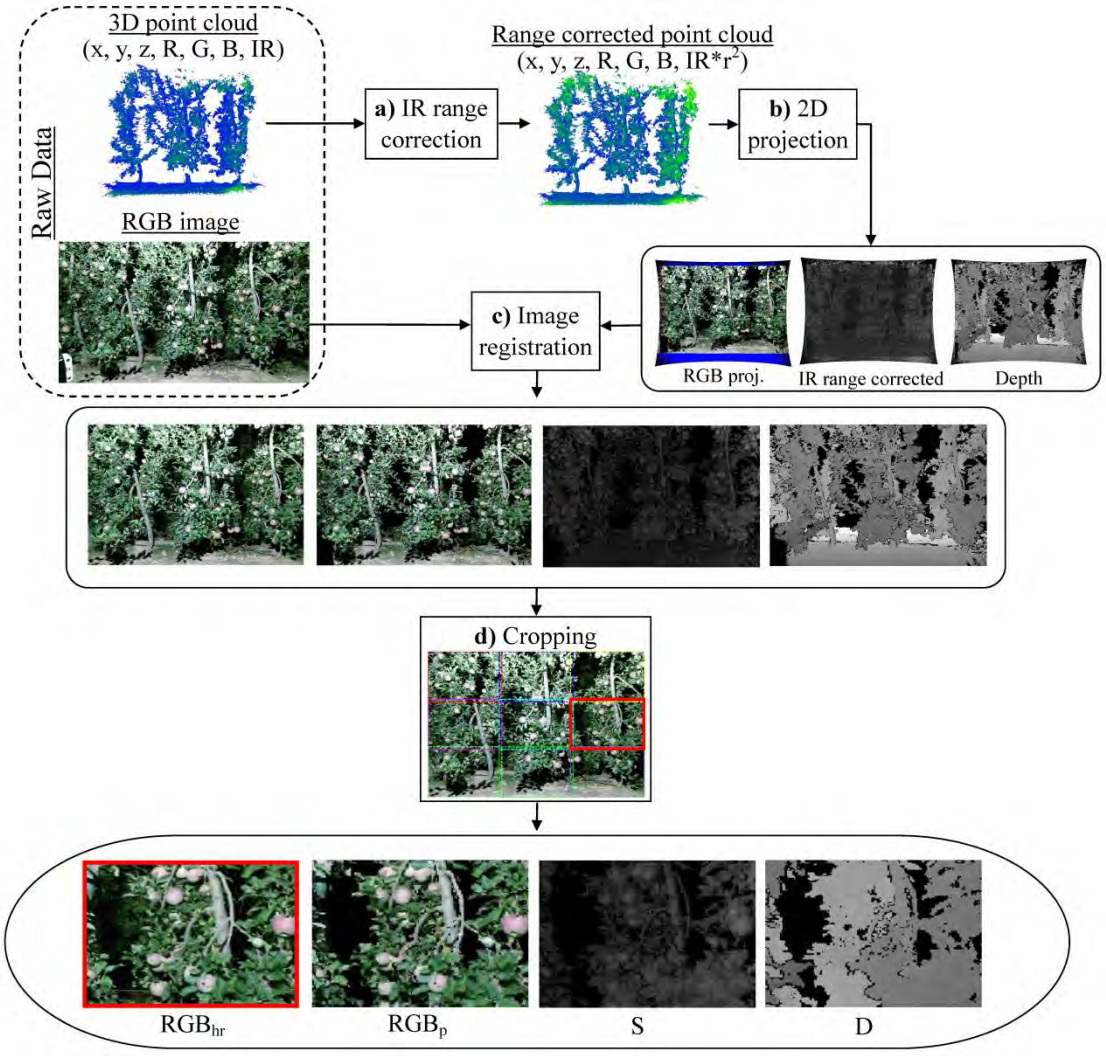


Figure 2. Data preparation outline.

Acknowledgments
This work was partly funded by the Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya, the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund (ERDF) under Grants 2017 SGR 646, AGL2013-48297-C2-2-R and MALEGRA, TEC2016-75976-R. The Spanish Ministry of Education is thanked for Mr. J. Gené's pre-doctoral

fellowships (FPU15/03355). We would also like to thank Nufri and Vicens Maquinària Agrícola S.A. for their support during data acquisition.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Gené-Mola, J., Vilaplana, V., Rosell-Polo, J.R., Morros, J.R., Ruiz-Hidalgo, J., Gregorio, E., 2019. 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

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

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. *Adv. Agron.* 133, 71–112. doi:10.1016/bs.agron.2015.05.002

Chapter III.C. P3: Fuji-SfM dataset

Section submitted for publication in *Data in Brief:*

Fuji-SfM dataset: a collection of annotated images and point clouds for Fuji apple detection and location using structure-from-motion photogrammetry.

Jordi Gené-Mola[1,*], Ricardo Sanz-Cortiella[1], Joan R. Rosell-Polo[1], Josep-Ramon Morros[2], Javier Ruiz-Hidalgo[2], Verónica Vilaplana[2], Eduard Gregorio[1]

[1] Research Group in AgroICT & Precision Agriculture, Department of Agricultural and Forest Engineering, Universitat de Lleida (UdL) – Agrotecnio Center, Lleida, Catalonia, Spain.
[2] Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, Barcelona, Catalonia, Spain.

Abstract

The present dataset contains colour images acquired in a commercial Fuji apple orchard (*Malus domestica* Borkh. cv. Fuji) to reconstruct the 3D model of 11 trees by using structure-from-motion (SfM) photogrammetry. The data provided in this article is related to the research article entitled “Fruit detection and 3D location using instance segmentation neural networks and structure-from-motion photogrammetry” (Gené-Mola et al., n.d.). The Fuji-SfM dataset includes: (1) a set of 288 colour images and the corresponding annotations (apples segmentation masks) for training instance segmentation neural networks such as Mask-RCNN; (2) a set of 582 images used to generate the 3D model of 11 Fuji apple trees containing 1455 apples by using SfM; (3) the 3D point cloud of the scanned scene with the corresponding apple positions ground truth in global coordinates. This data allows the development, training, and test of fruit detection algorithms either based on RGB images, on coloured point clouds or on the combination of both types of data.

Keywords: Structure-from-motion; SfM; Fruit detection; Fruit location; Mask R-CNN; Photogrammetry; Terrestrial remote sensing

Specifications Table

Subject	Agronomy and Crop Science, Horticulture, Computer Vision and Pattern Recognition
Specific subject area	Precision Agriculture, Fruit Detection, Remote sensing
Type of data	Images Instance segmentation masks Coloured point clouds Fruit location annotations
How data were acquired	Images were taken freehand, using an EOS 60D DSLR Canon camera with an 18 MP (5184 x 3456 px) CMOS APS-C sensor (22.3 x 14.8mm), and a Canon EF-S 24mm f/2.8 STM lens.
Data format	Raw images: *JPG* Instance segmentation masks: *CSV* and *JSON* Point clouds: *TXT* Fruit location annotations: *TXT*
Parameters for data collection	The camera focal length was 35 mm (38mm film equivalent focal length), which corresponded to a field of view of [59° 10', 50° 35'] (horizontal, vertical). Images were taken from a distance of 3m between the camera and the middle plane of the row. The vertical and horizontal overlapping between neighbouring images was higher than 30% and 90%, respectively.
Description of data collection	A total of 11 Fuji apple trees containing 1455 apples were photographed from 53 position (per side) distributed along the row of trees, having a separation of approximately 22 cm between two consecutive positions. In each position, a vertical sweep of 5-6 images was practiced, obtaining images of all tree heights -from the soil/trunk to the upper part of the trees-. The East side of the row of trees was photographed in the morning, while the West face in the afternoon, obtaining a similar illumination conditions in both faces.
Data source location	City/Town/Region: *Agramunt, Catalonia* Country: *Spain* GPS coordinates for collected data: *E: 336297 m, N: 4623494 m, 312 m a.s.l., UTM 31T - ETRS89*
Data accessibility	Repository name: *GRAP datasets / Lfuji-air dataset* Data identification: *Lfuji-air dataset* Direct URL to data: http://www.grap.udl.cat/en/publications/datasets.html
Related research article	Gené-Mola J, Sanz-Cortiella R, Rosell-Polo JR, Morros J-R, Ruiz-Hidalgo J, Vilaplana V, Gregorio E. 2019. Fruit detection and 3D location using instance segmentation neural networks and structure-from-motion photogrammetry. (Submitted)

Value of the data

- First dataset for fruit detection with 3D coloured point clouds generated by applying structure-from-motion photogrammetry.
- Computer vision community can benefit from these data to test new object detection and segmentation algorithms either based on 2D or on 3D data.
- Annotations provided can be used for training machine learning systems used in agriculture with applications such as yield prediction or yield mapping.
- This dataset can be used for benchmarking fruit detection and structure-from-motion algorithms.

1. Data

The Fuji-SfM dataset includes annotated data for 2D and 3D fruit detection. Raw data consist of 582 images (291 per row of trees side) of 11 consecutive Fuji apple trees (*SfM-set*), and 12 additional images (*Mask-set*) used in (Gené-Mola et al., n.d.) to train and validate the Mask-RCNN. All raw images are 5184 x 3456 pixels (px) size and were saved in 8-bit JPG format. Since the performance of object detection and segmentation neural networks decreases when detecting small objects in the images (Gené-Mola et al., 2019b), each *Mask-set* image was divided into 24 sub-images of 1024 x 1024 px (Figure 1a and Figure 1b). The resulting 288 sub-images were manually annotated generating the apples segmentation masks ground truth (Figure 1c). Image annotations were saved in CSV and JSON file formats, where each mask is a set of polygon points enclosing the pixels of an apple.

The *SfM-set* images were used to generate the 3D model of the scanned scene by applying structure-from-motion photogrammetry. The obtained 3D model was georeferenced in global world coordinates and saved as a point cloud in TXT format. Each row of the point cloud file correspond to a single 3D point, giving the information of [x, y, z, R, G, B], where, [x, y, z] is the point position in global coordinates, and [R,G,B] is the point colour with 8-bit precision values ranging from 0 to 255. The point cloud was manually labelled by placing 3D rectangular bounding boxes around each apple position (blue bounding boxes illustrated in Figure 2). A total of 1455 apples were annotated. Each fruit location annotation was saved in a TXT file where the first row corresponds to the position [x, y, z] of the apple centre, while the following eight rows indicate the positions of the bounding box corners.

Figure 1. Illustration of an image from the *Mask-set*. a) Image cropping borders. b) Example of two sub-images. c) Ground truth segmentation masks.

2. Experimental Design, Materials, and Methods

Images provided in Fuji-SfM dataset were acquired on September 2017 in a commercial Fuji apple orchard located in Agramunt, Catalonia, Spain (E: 336,297 m; N: 4,623,494 m; 312 m a.s.l., UTM 31T - ETRS89). The scanned trees were trained in a tall spindle system, with a maximum canopy height of 3.5m and width of 1.5 m, approximately. *Mask-set* images were taken from different randomly selected zones of the orchard, while *SfM-set* images were acquired from both sides of 11 consecutive trees containing a total of 1455 apples. All data was acquired three weeks before harvesting, at BBCH phenological growth stage 85 (Meier, 2001).

The camera used for data acquisition was an EOS 60D DSLR Canon camera (Canon Inc. Tokyo, Japan), with an 18 MP (5184 x 3456 px) CMOS APS-C sensor (22.3x14.9 mm), and a Canon EF-S 24mm f/2.8 STM lens (35 mm film equivalent focal length of 38 mm). All images were taken freehand from a distance of approximately 3 m from the trees centre, and at a height of 1.7 m (Figure 2). Images from the east side of the row of trees were photographed in the morning (11:53 – 12:26h), while the west side was photographed in the afternoon (15:27 – 16:05h) under natural illumination conditions.

Figure 2 illustrates the data acquisition process followed for the *SfM-set*. Yellow circles represent the camera centre of different photographic positions. The separation between two consecutive positions was 0.2 m, corresponding to a total of 53 photographic positions per row of trees side. From each camera position, a vertical sweep of 5-6

photographs was taken (black lines). With this configuration, a total of 291 images were taken per side, with a vertical/horizontal overlapping between neighbouring images higher than a 30/90 %, respectively (as shown in (Gené-Mola et al., n.d.), figure 2).

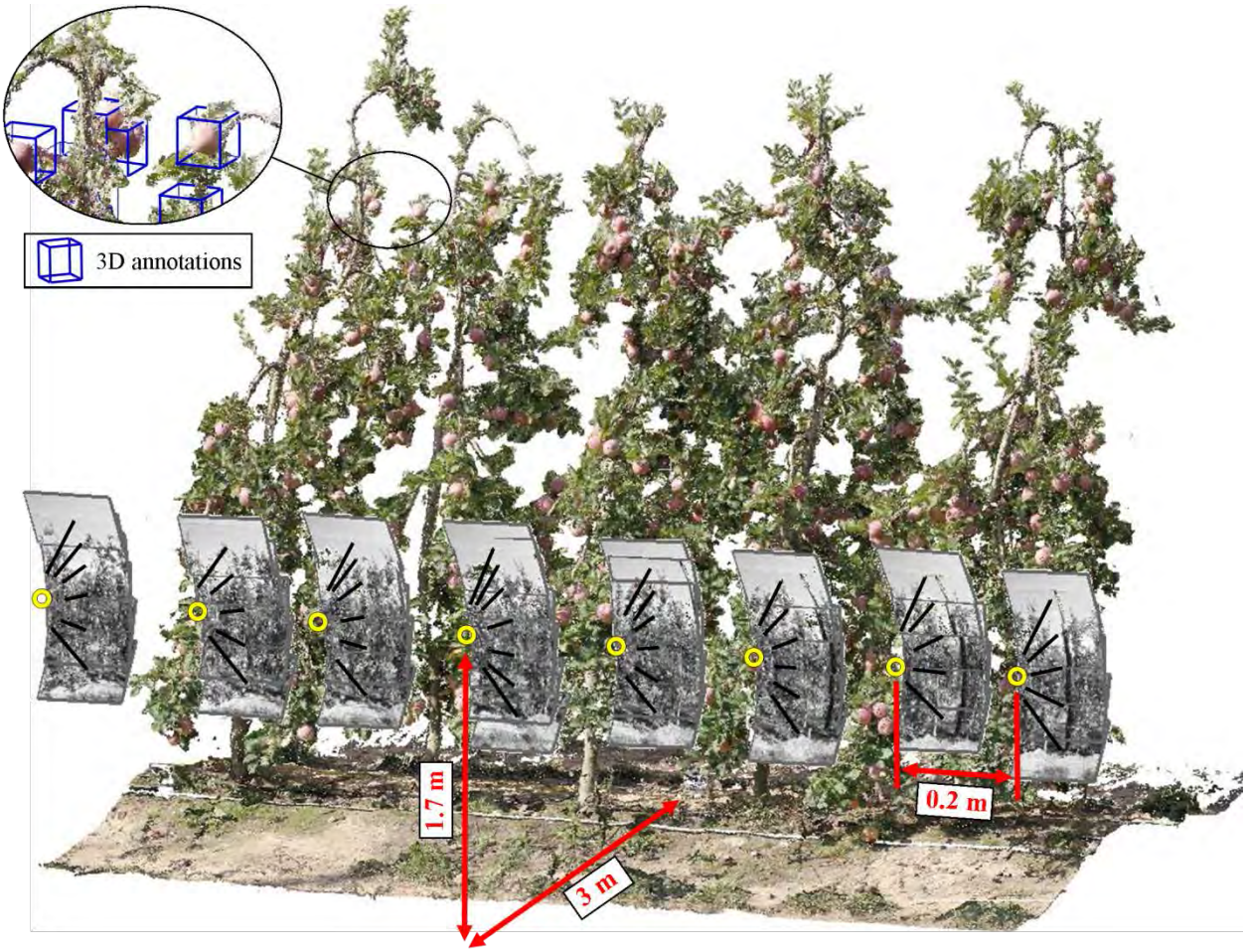


Figure 2. Isometric view of five scanned trees and illustration of the photographic process layout. Yellow circles show the photographic position. Blue 3D rectangular bounding boxes illustrate the apple position ground truth annotations.

SfM-set images were used to reconstruct the 3D model of the 11 scanned trees. A multi-view structure-from-motion photogrammetry based on bundle adjustment (Triggs et al., 2000) was applied to generate the 3D point cloud of each side of the row of trees. This 3D model generation was carried out using Agisoft Professional Photoscan software (v1.4, Agisoft LLC, St. Petersburg, Russia). A set of known markers (depicted in Chapter VII, Figure 5d) in the scene was used to scale and georeferencing the obtained point clouds. Then, point clouds from both sides of the row of trees were merged, obtaining a complete representation of the scanned trees in a single point cloud.

Mask-set images were manually labelled with apple segmentation masks, allowing the use of this set of images to train and test 2D instance segmentation algorithms. This annotation was performed using the VIA annotation software (Dutta and Zisserman, 2019), enclosing individual apples with polygon region shapes. The point cloud of the 11 scanned trees was also manually labelled. Similarly than in (Gené-Mola et al., 2019a), the 3D annotation was carried out using the software CloudCompare (Cloud Compare [GPL software] v2.9 Omnia), placing 3D rectangular bounding boxes around each apple, as can be seen in the zoomed-in region of Figure 2.

Acknowledgments

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.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Dutta, A., Zisserman, A., 2019. The VIA Annotation Software for Images, Audio and Video, in: *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, New York, NY, USA. doi:10.1145/3343031.3350535

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., Sanz-Cortiella, R., Rosell-Polo, J.R., Morros, J.-R., Ruiz-Hidalgo, J., Vilaplana, V., Gregorio, E., n.d. Fruit detection and 3D location using instance segmentation neural networks and structure-from-motion photogrammetry. Submitted.

Gené-Mola, J., Vilaplana, V., Rosell-Polo, J.R., Morros, J.R., Ruiz-Hidalgo, J., Gregorio, E., 2019b. 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

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

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

Chapter IV. P4: Fruit detection in an apple orchard using a mobile terrestrial laser scanner

This chapter was published in *Biosystems Engineering 187* (2019) 171-184, https://doi.org/10.1016/j.biosystemseng.2019.08.017:

Research Paper

Fruit detection in an apple orchard using a mobile terrestrial laser scanner

Jordi Gené-Mola [a], *Eduard Gregorio* [a,*], *Javier Guevara* [b], *Fernando Auat* [b], *Ricardo Sanz-Cortiella* [a], *Alexandre Escolà* [a], *Jordi Llorens* [a], *Josep-Ramon Morros* [c], *Javier Ruiz-Hidalgo* [c], *Verónica Vilaplana* [c], *Joan R. Rosell-Polo* [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 Electronic Engineering, Universidad Técnica Federico Santa María, Valparaíso, Chile
[c] Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, Barcelona, Catalonia, Spain

Abstract

The development of reliable fruit detection and localization systems provides an opportunity to improve the crop value and management by limiting fruit spoilage and optimized harvesting practices. Most proposed systems for fruit detection are based on RGB cameras and thus are affected by intrinsic constraints, such as variable lighting conditions and camera calibration. This work presents a new technique that uses a mobile terrestrial laser scanner (MTLS) to detect and localise Fuji apples. An experimental test focused on Fuji apple trees (*Malus domestica* Borkh. cv. Fuji) was carried out. A 3D point cloud of the scene was generated using an MTLS composed of a Velodyne VLP-16 LiDAR sensor synchronized with an RTK-GNSS satellite navigation receiver. A reflectance analysis of tree elements was performed, obtaining mean apparent reflectance values of 28.9%, 29.1%, and 44.3% for leaves, branches and trunks, and apples, respectively. These results suggest that the apparent reflectance parameter (at 905 nm wavelength) can be useful to detect apples in the tree. For that purpose, a four-step fruit detection algorithm was developed. By applying this algorithm, a localization success of 87.5%, an identification success of 82.4%, and an

F1-score of 0.858 were obtained in relation to the total amount of fruits. These detection rates are similar to those obtained by RGB-based systems, but with the additional advantages of providing direct 3D fruit location information, which is not affected by sunlight variations. From the experimental results, it can be concluded that LiDAR-based technology and, particularly, its reflectance information, has potential for remote apple detection and 3D location.

Keywords: LiDAR; Mobile Terrestrial Laser Scanning; Fruit detection; Agricultural robotics; Fruit reflectance.

Nomenclature

FDR_{ID}	False detection rate identification
FDR_L	False detection rate localization
FoV	Field of View *[º]*
FP_{ID}	False positive identification
FP_L	False positive localization
GT_{field}	Number of fruits manually-counted in field
GT_{labels}	Number of fruits labelled
IoD_i	Intersection over detection
K	Number of fruits in a cluster
$MTLS$	Mobile Terrestrial Laser Scanner
n	Number of clusters that detect the same fruit
N_m	Fruit multi-detections (n-1)
P	Number of points of a cluster
P_{kj}	Number of points threshold used to find clusters with *j* apples
R	Apparent reflectance *[%]*
R_{th}	Reflectance threshold *[%]*
$RTK\text{-}GNSS$	Real-Time Kinematics Global Navigation Satellite System
$\bar{R}$	Mean apparent reflectance of the points of a cluster *[%]*
$\bar{R}_{FP}$	Mean apparent reflectance threshold used to find false positive clusters *[%]*
$\bar{R}_{kj}$	Mean apparent reflectance threshold used to find clusters with *j* apples *[%]*
$Success_{ID}$	Identification success (recall)
$Success_L$	Localization success
SVD	Singular Value Decomposition
TOF	Time of flight
TP_{ID}	True positive identification
TP_L	True positive localization
V	Volume of a cluster *[m^3]*
V_{FP}	Volume threshold used to find false positive clusters *[m^3]*
V_{kj}	Volume threshold used to find clusters with *j* apples *[m^3]*
[x, y, z]	3D point with UTM coordinates *[m]*
α	Sparse outlier removal tuning parameter
λ_{in}	Normalized principal value *i*
λ_i	Principal value *i* of a cluster
Ψ	Geometric parameter
Ψ_{FP}	Geometric parameter value used to find false positive clusters
Ψ_{kj}	Geometric parameter value used to find clusters with *j* apples

1. Introduction

Fruticulture is under constant pressure to increase fruit production and quality, as demanded by a growing world population. To this end, farmers need to find new ways to improve fruit productivity and, at the same time, reduce economic and environmental costs (Siegel *et al.*, 2014). Agricultural robotics takes advantage of new technologies to respond to this challenge (Bac *et al*., 2014; Bechar and Vigneault, 2017, 2016; Gongal *et al*., 2015; Y. Zhao *et al*., 2016). The use of robotics in agricultural fields and orchards is increasing, particularly in tasks related to guidance (seeding or harvesting), detection (weed monitoring and control, extraction of biological features), and mapping (Auat Cheein *et al*., 2017; Auat Cheein and Carelli, 2013; Foglia and Reina, 2006). In general, the development of intelligent robots interacting with agricultural fields increases the accuracy of tasks and reduces the consumption of resources without decreasing yield, making it a reasonable option for repeatable tasks (Cariou *et al*., 2009; Foglia and Reina, 2006; Zhang and Pierce, 2016).

Fruit detection and localization are complex tasks that can be handled by agricultural robotics, with applications related to yield prediction, yield mapping, and automated harvesting. Nowadays, yield prediction is done by manual counting of selected sample trees, leading to inaccurate predictions due to the high variability in orchards (Payne *et al*., 2014; Stein *et al*., 2016). Crop monitoring using new technologies could provide more accurate and efficient predictions (Bechar and Vigneault, 2017, 2016). Another application of fruit detection is yield mapping. The fruit load of an orchard is influenced by in-field spatial variability (due to soil type variations), fertility, and water content, among other factors. In precision agriculture, yield mapping helps to determine the reasons for and find solutions to cope with this variability (Kurtulmus *et al*., 2014). Finally, fruit localization is the basis for future automated harvesting. Manual picking is a bottleneck in fruit production management, because it requires lots of resources in the context of decreasing farming labour force. In addition, hand harvesting exposes farmers to awkward postures on ladders and platforms with heavy loads, making manual harvesting dangerous and inefficient (De-An *et al*., 2011; Gongal *et al*., 2015).

The detection of fruits can involve many fruit properties of different complexity, from the simplest, such as the presence/absence of a fruit, to properties that are more challenging to measure, including size, volume, diameter, maturation stage, sugar, and other substance contents, defects and disease/pest affectation, etc. There are multiple technologies available for fruit detection and localization, each with its advantages and disadvantages (Gongal *et al*., 2015). All approaches have to solve problems derived from occlusions (Stein *et al*., 2016; Wachs *et al*., 2010), clustering (Gong *et al*., 2013; Xiang *et al*., 2014), and variable lighting conditions (Gongal *et al*., 2016; C. Zhao *et al*., 2016).

The most commonly used sensors are RGB cameras (Linker, 2017; Maldonado and Barbosa, 2016; C. Zhao *et al*., 2016). These are affordable sensors, which allow fruits to be distinguished from other elements by colour (Linker *et al*., 2012; Liu *et al*., 2016), geometric shape (Barnea *et al*., 2016; Lak *et al*., 2010), texture (Chaivivatrakul and Dailey, 2014; Qureshi *et al*., 2017), or by using machine learning techniques like, e.g., deep neural networks (Bargoti and Underwood, 2017). The two main drawbacks to RGB cameras are their sensitivity to lighting conditions and the fact that they only provide 2D information (unless using stereoscopic techniques). Other, more expensive, cameras include thermal cameras (Bulanon *et al*., 2009, 2008; Stajnko *et al*., 2004; Wachs *et al*., 2010), multispectral cameras (Sa *et al*., 2016; Zhang *et al*., 2015), and hyperspectral cameras (Okamoto and Lee, 2009; Safren *et al*., 2007). The former allows fruits to be distinguished from the background through their temperature, while the latter detect fruits from their reflectance at different wavelengths. Like RGB cameras, thermal, multispectral, and hyperspectral cameras do not provide 3D information, unless a stereoscopic approach is implemented.

There are several solutions to obtain three-dimensional information. One of them is based on using two (stereovision) or more cameras (Font *et al*., 2014; Si *et al*., 2015; Xiang *et al*., 2014). By applying triangulation techniques, it is possible to obtain the depth of each pixel and reconstruct the 3D structure. The major advantage of this technique is that it allows us to obtain accurate 3D models with RGB information, while the main disadvantages are that 3D model generation is computationally expensive and the performance is affected by lighting conditions. Another more recent technique is the

use of laser range finders and LiDAR-based (Light Detection and Ranging) systems. These are more expensive sensors that generally operate under the principle of time-of-flight (TOF) (Wehr and Lohr, 1999). This type of sensor typically also provides the amount of energy backscattered from the impacted object. Very few studies have used LiDAR-based systems in fruit detection and, to the best of the authors' knowledge, none of them have been tested in a real orchard environment. For example, Jiménez *et al*. (2000, 1999) developed a vision system based on a laser range-finder, with the aim of detecting spherical objects in non-structured environments. They report good detection performances, although the tests were carried out on a limited number of oranges suspended from an artificial tree. Finally, another technology derived from photogrammetry and LiDAR, and also used in fruit growing, are the RGB-D (depth) cameras, where each pixel of the image contains colour and depth data, generating 3D colour images (Barnea *et al*., 2016; Nguyen *et al*., 2016; Rosell-Polo *et al*., 2017, 2015). These systems are based on the simultaneous combination of RGB cameras and depth sensors based on laser light (either through structured laser light or TOF flash-type LiDAR-based systems).

This work presents a proof of concept of using LiDAR in detecting Fuji apples in producing orchard trees. The methodology is founded on the fact that apples have higher apparent reflectance than leaves and trunks at 905 nm laser wavelength. The main contributions of this paper are: (1) analysis of apple reflectivity on 3D point clouds from LiDAR sensors; (2) development of an apple detection and localization algorithm based on three stages (point cloud segmentation; fruit separation, and false positive removal); and (3) experimental validation of the proposed technique on a real Fuji apple orchard. The principal advantage of this technique over previously published efforts would be its capacity to provide direct 3D fruit localization information without being affected by illumination conditions. The paper is structured as follows. Section 4.2 presents the experimental data set, the point cloud generation procedure, the reflectance analysis, and the developed apple detection algorithm. Section 4.3 shows the results of the first experimental tests performed on three Fuji apple trees of a commercial orchard. Finally, the conclusions are presented in Section 4.4.

2. Materials and methods

2.1. Experimental set up

A fruit detection experiment was carried out on September 28th of 2017 in Tarassó farm, a commercial apple orchard located in Agramunt, Catalonia, Spain (E: 336,297 m; N: 4,623,494 m; 312 m a.s.l., UTM 31T - ETRS89). The trials were carried out in an 8-year-old Fuji apple orchard (*Malus domestica* Borkh. cv. Fuji), trained in a tall spindle system with a maximum tree height of 3.75 m. The three analysed trees were at BBCH (Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie) growth stage 85 (Meier *et al.*, 2001), three weeks before harvesting.

The measurement equipment consisted of a mobile Terrestrial Laser Scanner (MTLS), comprised of a LiDAR sensor and a real-time kinematics global navigation satellite system (RTK-GNSS), connected to a rugged laptop suitable for working in field conditions. The LiDAR sensor used was a Puck VLP-16 (Velodyne LIDAR Inc., San José, CA, USA), which generates a 3D point cloud (x-y-z positions) of the scanned scene, as well as calibrated apparent reflectance (*R*) of each point in the 3D point cloud. This calibration was carried out by sensor manufacturer using a set of calibration targets, and implies a conversion of the backscattered range-corrected intensity (digital numbers) into apparent reflectance values independently of laser power and distance (Velodyne, 2016). Note that the measured apparent reflectance (hereinafter referred to as reflectance) is an approximation of the actual hemispherical reflectance, considering that the measured objects are Lambertian (diffuse reflectors), and not considering the incidence angle (Kaasalainen et al., 2011; Kukko et al., 2008; Ray, 1994). The VLP-16 sensor emits 16 laser beams (905 nm wavelength) with a horizontal angular resolution of 2º (30º horizontal FoV) when mounted on a vertical plane as shown in Figure 1. Although the vertical FoV can be set up to 360º, in this experiment it was set to 150º, since only one row of trees was scanned. The scanning frequency rate was set to 10 Hz, corresponding to a vertical angular resolution of 0.2º, so that a maximum of 12,000 points were obtained from each scan (acquisition speed of 120,000 points/second). Even though this sensor has a range of 100 m, points further than 4 m where not considered for 3D point cloud generation, thus only the tree row of interest was modelled. The

acquisition of Coordinated Universal Time (UTC) of each point was obtained via a GPS 18x LVC receiver (Garmin International Inc., Olathe, KS, USA), connected to the VLP-16 sensor. The RTK-GNSS system used was the GPS1200+ (Leica Geosystems AG, Heerbrugg, Swizeland), which provides absolute coordinates and UTC time (synchronized with the LiDAR) with a frequency of 20 Hz and a precision of approx. 20 mm.

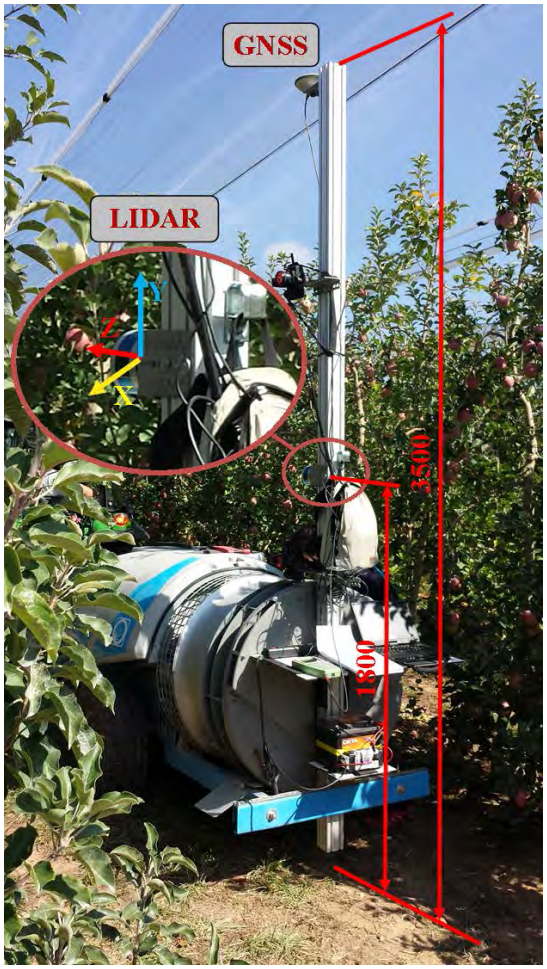


Fig. 1. View of the MTLS equipment showing the GNSS antenna placement and the mounting orientation of the LiDAR sensor. Distance data are in mm.

As shown in Figure 1, the MTLS measurement system was mounted on the rear of an air-assisted sprayer by means of an aluminium structure. The sprayer was pulled at low gear by a farm tractor equipped with an electronic speedometer. The GNSS rover receiver antenna was installed on top of the mast, at a height of 3.5 m. The LiDAR sensor was mounted vertically (Figure 1) and placed at a height of 1.8 m, that is about half the maximum height of studied trees. This position was selected to have similar

detection performance along the tree height. The field test was performed by moving the MTLS along a rectilinear trajectory parallel to the tree row axis, at a distance of 2.4 m. Due to the fact that the system did not include an inertial measurement unit (IMU), moving the MTLS along a linear trajectory was important to improve the point cloud consistency. The forward speed was 0.125 m s^{-1}, corresponding to a resolution of 12.5 mm between consecutive scans (~53,600 points m^{-2} in a vertical plane at the distance of 2.4m). The tree row was scanned from both sides in order to obtain a complete 3D model.

2.2. 3D point cloud model

A rigid transformation was performed by applying a rotation and translation matrix to each point, in order to build the point cloud with absolute coordinates. The translation matrix was built using the Universal Transverse Mercator (UTM) coordinates of the RTK-GNSS system, by considering the relative distance between the optical centre of the LiDAR sensor and the GNSS receiver. The rotation matrix depends on the orientation of the MTLS at each time instant and was obtained by the forward direction computed from the measurements of the RTK-GNSS receiver. Given that the trials were performed with a short rectilinear trajectory, the tilt of the platform can be ignored, assuming a constant orientation along the path. An illustration of the 3D point cloud models generated is shown in Figure 2.

The resulting 3D point cloud was manually labelled in order to generate ground truth of the apples locations. This enables a study of the features that characterize the apples, as well as the possibility to evaluate the performance of the developed apple detection techniques. The annotation was carried out using the software CloudCompare (Cloud Compare [GPL software] v2.9 Omnia), placing 3D rectangular bounding boxes on each apple, as can be seen in the third tree of Figure 2. This annotation was supported by additional RGB images to localise the apples in the 3D point cloud. The actual number of apples counted in field (ground truth field or *GT_field*) were 139 in tree 1, 145 in tree 2, and 139 in tree 3, of which 133, 138 and 134, respectively, could be labelled in the 3D point cloud (ground truth labels or *GT_labels*) due to occlusions or in field counting errors. Trees 1 and 2 were used as the training dataset to select and tune the algorithm

parameters, while tree 3 was used as the test dataset to evaluate the performance of the developed algorithm. From the labelled scene and the reflectance data extracted from the LiDAR for each point, a reflectance study of the different elements of the tree was carried out (Section 3.1).

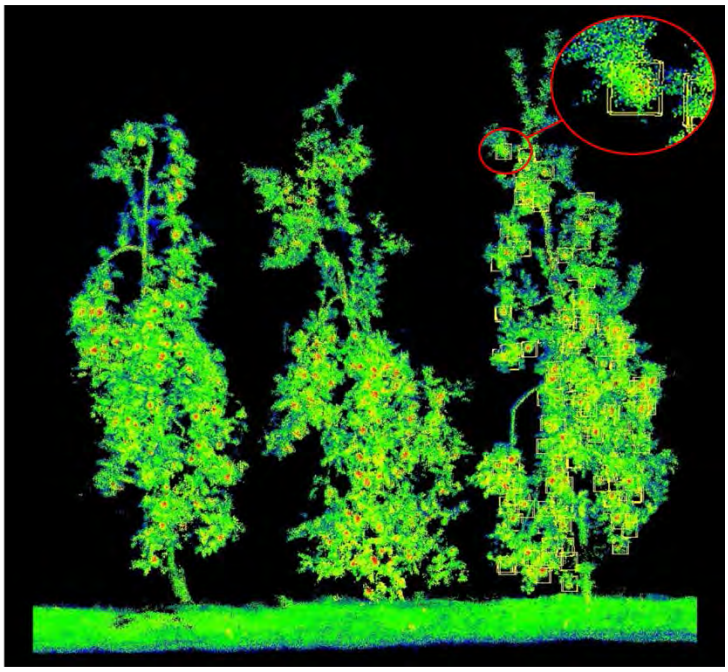


Figure 2. 3D point cloud models obtained for trees 1, 2, and 3. First two trees were used as training dataset, while the third was kept as test dataset. Ground truth bounding boxes of tree 3 are shown, while the zoom bounding box (red circle) shows its shape.

2.3. Apple detection algorithm

As shown in Figure 3, the algorithm proposed in this paper is structured as follows: 1) Point cloud segmentation; 2) fruit separation; and 3) false positive removal. The segmentation is based on the reflectance of measured elements and aims at removing points corresponding to leaves, branches, and the trunk, and grouping the remaining points -likely to be an apple- in clusters. The fruit separation uses features of clusters in order to identify and split those that contain more than one apple. False positive removal is based on the geometry and reflectance of the clusters. All data processing was implemented in MATLAB (R2018a, Math Works Inc., Natick, Massachusetts, USA). The different implemented steps are detailed below.

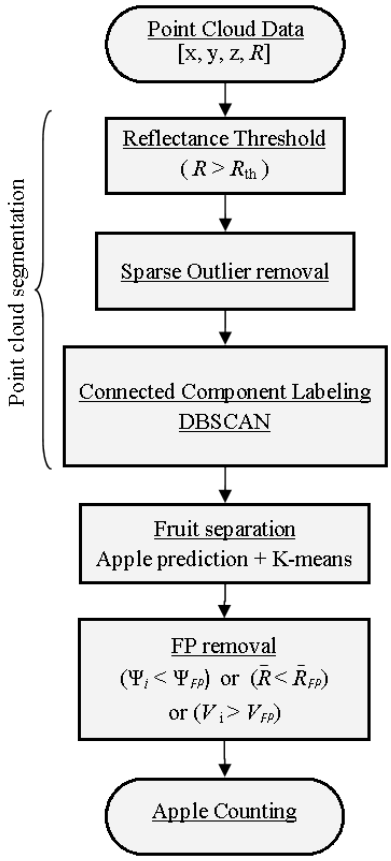


Figure 3. Apple detection algorithm flowchart.

2.3.1. Point cloud segmentation

The objective of this step is to segment the 3D point cloud and obtain a set of clusters with points that could be apple candidates. Since some groups of apples could be touching, the clusters obtained in this first step could contain one or more apples. The 3D model acquired with the MTLS consists of a set of 3D points with UTM coordinates and their reflectance *[x, y, z, R]*. The reflectance analysis (Section 3.1) shows that apple reflectance at the 905 nm laser wavelength is higher than that for leaves and the trunk and, therefore, this parameter is used for apple detection. To remove the points that do not correspond to apples, a threshold, R_{th}, is applied. This is followed by Sparse Outlier Removal (Rusu *et al*., 2008) to reduce the noise; this approach removes the points which fall outside $\mu + \alpha \cdot \sigma$, with μ and σ being the mean and standard deviation, respectively, of the k nearest neighbour distances, while α is a tuning parameter. The point cloud segmentation ends with a connected components labelling using a density-based scan algorithm, DBSCAN (Ester *et al*., 1996), which clusters points that have more than minPts points closer than a distance, ε. Outlier Removal is first applied to

delete noisy points that otherwise would connect clusters from different apples. All the parameters used in this step were selected through a hyperparameter optimization procedure, using the training data set to search for the combination of parameters that best suits our data. In this search, we found that the results were stable against small variations in the different parameter values, except the reflectance threshold, the behaviour of which is shown in Section 3.1, Figure 7. The parameter values used are detailed in Appendix A.

2.3.2. Apple separation

If apples are properly separated, the results obtained in the previous step would consist of a set of clusters of one apple in each. Nevertheless, it was found that groups of apples touching will result in clusters of more than one apple. The aim of this second step is to identify the clusters containing more than one apple and split them into sub-clusters, each containing one apple. First, the number of apples, *K*, that make up a cluster has to be predicted, and then the cluster is split using the K-means algorithm. This clustering method aims to partition the 3D points into K sub-clusters in which each 3D point belongs to the sub-cluster with nearest mean (Jain, 2010).

To predict the number of apples contained in each cluster (the K number used in the K-means algorithm), three different methods were tested. The first one is inspired by a template matching technique (Brunelli, 2009). The second method applies a decision tree, based on cluster features such as volume, density of points, reflectance, and shape. Finally, the third method is a combination of the previous two approaches. These methods are explained in more detail below.

Method 1

The first approach projects the 3D point clouds of each cluster (Figure 4b) onto a 2D plane, obtaining an image of the cluster with reflectance data at a resolution of 4x4 mm per pixel (Figure 4c). The cluster image then is convolved with a Gaussian filter of size 20x20 pixels and standard deviation of 3.5. These parameters correspond to the measured fruit size, so that the dimension of this filter is 80x80 mm, similar to the mean size of the tested apples. Since the apples have an approximately spherical shape, when the cluster image is convolved with a Gaussian filter, the local maxima of the obtained

image correspond to the centres of the apples (Figure 4d). The value K, to be used in the K-means algorithm, corresponds to the number of local maxima found in the convolved image (Figure 4e).

The result of this method could vary with the 2D projection plane used (e.g., the projection may produce occlusions). The technique is applied in four different planes to prevent this projection-induced variability: frontal, lateral, top, and the plane defined by the first two principal axes of the cluster (Figure 4, b and c). The value of *K* will be the maximum obtained in these four planes. The principal axes are the directions where the variance of data is maximized and, therefore, where the points exhibit the largest range. The first two principal axes define the principal plane of the cluster and are obtained by applying singular value decomposition (SVD) to the set of points forming the cluster.

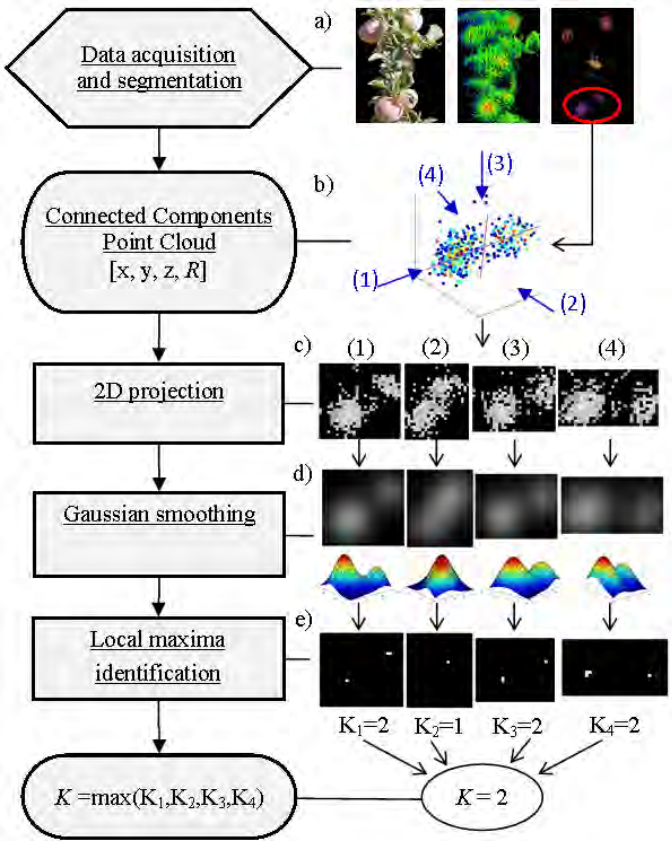


Figure 4. Method 1 - Cluster splitting by Gaussian smoothing. The aim of this method is to determine the number of apples, K, that are contained in a cluster. a) Actual data before applying method 1: the real scene is scanned and the resulting point cloud is segmented, obtaining clusters likely to contain apples. b) Cluster containing 2 apples. c) 2D projection in four planes: (1) frontal; (2) lateral, (3) top; and (4) plane defined by 2 principal axes. d) Gaussian smoothing. e) Local maxima identification.

Method 2

The second method applies a decision tree based on cluster features. The first step is to extract the following features for each cluster: volume (V), number of cluster points (P), mean reflectance of cluster points ($\bar{R}$), and a geometric parameter, Ψ, computed as the product of normalized eigenvalues $[\lambda_{1n}, \lambda_{2n}, \lambda_{3n}]$. The volume ($V$) was defined as the volume enclosed by the boundary points of the cluster. Clusters that contain more than one fruit are expected to have a larger volume (V) and more points (P). However, when a fruit is placed next to a leaf or trunk (not filtered in previous steps), the cluster volume (V) and the number of points (P) could increase as well. Due to this fact, the threshold, $\bar{R}$, is applied, as it was observed that the mean reflectance of this kind of trunk/leaf co-located cluster is lower than clusters containing grouped fruits. The last features used are the eigenvalues, which provide information about the cluster shape. Spherical shapes (clusters with only one apple) will have similar eigenvalues, while elongated shapes will have different eigenvalues. Eigenvalues are obtained with SVD, and their values depend on the variance of the points projected on the principal axes. In order to compare eigenvalues of different clusters, a normalization step is applied so that the eigenvalues sum to one. From that, the geometric parameter Ψ is defined as the product of eigenvalues and a normalization factor. The normalization factor of 27 allows the geometrical parameter, Ψ, to be bound between 0 and 1:

$$\lambda_{in} = \frac{\lambda_i}{\lambda_1+\lambda_2+\lambda_3} \quad \text{so that} \quad \lambda_{1n} + \lambda_{2n} + \lambda_{3n} = 1 \tag{1}$$

$$\Psi = 27 \cdot \lambda_{1n} \cdot \lambda_{2n} \cdot \lambda_{3n} \quad \text{where} \begin{cases} \Psi = 1 & \text{for spherical distributions} \\ 1 > \Psi \geq 0 & \text{otherwise} \end{cases} \tag{2}$$

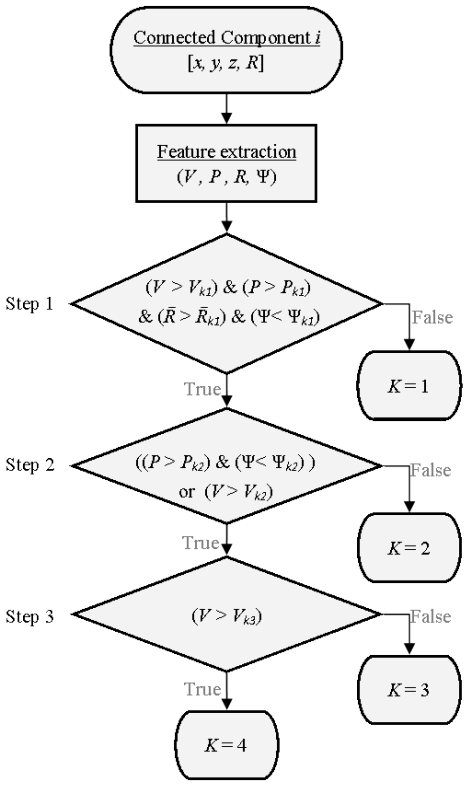


Figure 5. Method 2 - Decision tree used to predict the number of apples in a cluster.

The implemented decision tree is based on the analysed features in the training data set and is composed of the following steps (Figure 5):

- Feature extraction: Compute V ,P ,$\bar{R}$, and Ψ of the studied cluster.
- Step 1: If V, P, and $\bar{R}$ are higher than the corresponding thresholds V_{k1} , P_{k1}, $\bar{R}_{k1}$, and Ψ is smaller than Ψ_{k1}, it is concluded that the cluster contains more than one apple. Otherwise, K is assigned the value *1*.
- Step 2: A cluster will have more than two apples if P is higher than P_{k2} and Ψ is lower than Ψ_{k2} , or if V is higher than V_{k2}. Otherwise, K is assigned the value of *2*.
- Step 3: $K=4$ when a cluster meets both previous conditions and has a volume (V) higher than V_{k3}. Otherwise, K is assigned the value *3*.

All threshold values used in the decision tree were empirically selected by the graphical representation of four analysed features using the training dataset. The values used and the graphical representation of these features are presented in Appendix A, Table A1 and Figure A1.

Method 3

By applying method 1, some single-fruit clusters are split into multiple detections due to partial occlusions of apples by leaves. Method 3 addresses this concern by combining methods 1 and 2. First, step 1 of method 2 is applied to distinguish between clusters with single or multiple apples. For those clusters that contain more than one apple, method 1 is applied to determine the value of K.

2.3.3. False Positive removal

After implementing the first two steps of the algorithm (segmentation and apple separation), it was observed that some detections do not actually correspond to apples, i.e., these were false positive detections. That is because some leaves and trunks have a texture or shape that result in a high reflectance. It was found that some of these erroneous detections had a different geometric shape (Ψ), volume (V), and mean reflectance ($\bar{R}$) compared to the successful detections. In order to reduce these false positives, the clusters that met the condition $(\Psi < \Psi_{FP}) \mid (\bar{R} < \bar{R}_{FP}) \mid (V > V_{FP})$ were removed. In the same manner as with method 2, the thresholds were empirically selected from a graphical representation of these three features using the training dataset. The values used and the graphical representation of these features are presented in Appendix A, Table A1 and Figure A1.

2.4. Performance evaluation

In this work, the results were evaluated using two different approaches: localization and identification. The *localization evaluation* aims to assess the system in the context of harvesting automation. This approach assumes that a robotic arm, when it gets close to a group of apples, is able to separate different apples that have been detected within the same cluster, or to unify the multi-detections that correspond to the same apple. Thus, a detection that contains K apples counts as K true positives (Figure 6a), while multi-detections are counted as one true positive and no false positives (Figure 6e).

The *identification evaluation* aims to assess the system for use in yield prediction or mapping. This assessment is performed cluster-by-cluster, so that a single detection containing K apples counts as only one true positive (Figure 6a), while a single apple

detected *n* times (multi-detection) is counted as one true positive and $N_m = n-1$ false positives (Figure 6e).

To evaluate object detection in images, the metric intersection over union (IoU) is commonly used. This is possible when both bounding-box and object detection can be seen as a group of pixels. In this study, the detections are groups of 3D points, while ground truth bounding boxes are cube regions. The metric IoU has been substituted by the intersection over detection (IoD) for this reason; IoD is defined as the percentage of detected points that are placed inside ground truth bounding boxes.

The following defines the metrics used for each approach, namely localization (subscript L) and identification (subscript ID).

- Intersection over detection (IoD_i): Percentage of points, P_i, of a detection, *i*, that are placed inside ground truth bounding-boxes (*GT*). $IoD_i = \frac{P_i \cap GT}{P_i}$
- True positive localization (TP_L): Number of ground truth apples that are detected with an $IoD_i \geq 0.5$.
- False positive localization (FP_L): Number of detections with an $IoD_i < 0.5$.
- Localization success ($Success_L$): Quotient between TP_L and the number of labelled apples (GT_{labels}). $Success_L = \frac{TP_L}{GT_{labels}}$
- False detection rate localization (FDR_L): Ratio between FP_L and the total positive ($TP_L + FP_L$) $FDR_L = \frac{FP_L}{TP_L + FP_L}$
- True positive identification (TP_{ID}): Number of clusters with an $IoD_i \geq 0.5$, minus multi-detections ($\sum N_m$).
- False positive identification (FP_{ID}): Sum of the number of detections with an $IoD_i < 0.5$ (FP_L), plus multi-detections. $FP_{ID} = FP_L + \sum N_m$.
- Identification success ($Success_{ID}$ or recall): Quotient between TP_{ID} and GT_{labels}. $Success_{ID} = recall = \frac{TP_{ID}}{GT_{labels}}$
- False detection rate identification (FDR_{ID}): Ratio between FP_{ID} and the total positive. $FDR_{ID} = \frac{FP_{ID}}{TP_{ID} + FP_{ID}}$

- Precision: Percentage of TP_{ID} with respect to the total positive (TP_{ID} + FP_{ID})

$$Precision = \frac{TP_{ID}}{TP_{ID}+FP_{ID}}$$

- F1-score: Harmonic mean of precision and recall. $F1 = 2 \cdot \frac{precision \cdot recall}{precision+recall}$

Selected examples of the evaluation criteria can be seen in Figure 6. Intersection over detection (IoD) is given for different scenarios, while true positive and false positive rates are calculated for localization and identification assessment approaches. Red shapes are apple detections, while green squares correspond to the ground truth labels. Note that actual clusters and bounding-boxes are in 3D (as shown in Figure 2), although for the sake of simplicity this figure shows the 2D projection. The examples shown are: a) One cluster with K=2 apples and three GT bounding-boxes ; b) One cluster with K=1 apple and one GT bounding-box ; c) Two clusters of K=1 apple each and two GT bounding-boxes ; d) One GT bounding-box not detected ; e) Two clusters detecting the same GT bounding-box (multi-detection) ; f) One cluster that does not correspond to any GT object ; and g) One cluster detecting an apple with an IoD<0.5.

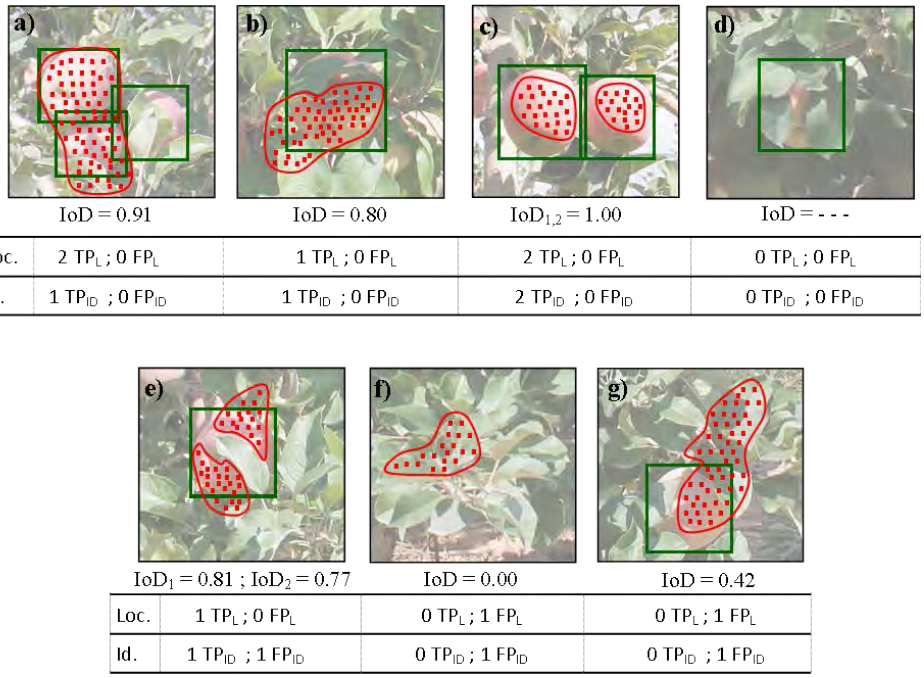


Figure 6. Localization and identification performance evaluation criteria. Intersection over detection (IoD) is given for different scenarios, while true positive and false positive are calculated for localization and identification assessment approaches. Red shapes are apple detections, while green squares correspond to the ground truth labels.

3. Results and discussions

3.1. Reflectance analysis

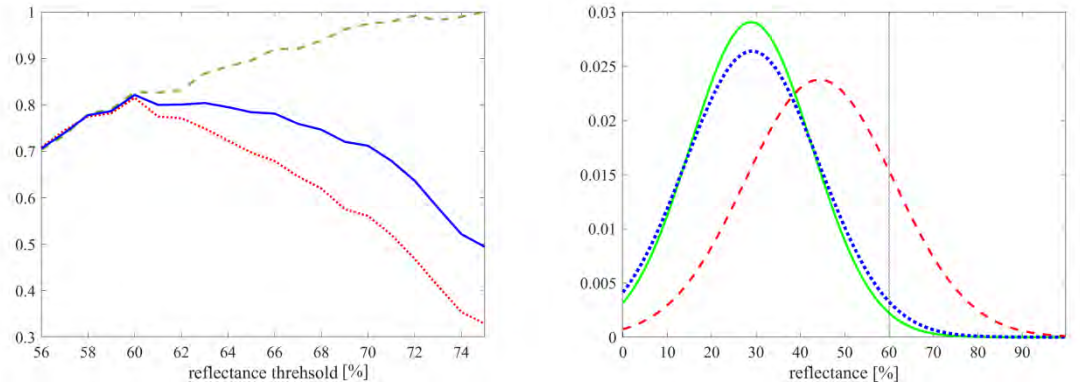
Table 1 shows the reflectance analysis results for both trees used in the training dataset. Mean apparent reflectance values of 28.9%, 29.1%, and 44.3% were obtained for leaves, trunks, and Fuji apples, respectively. These results indicate that the reflectance is higher than other tree elements. Hence, this characteristic will be used as a valuable feature for Fuji apple detection. Note that these results were obtained using a LiDAR system operating with a laser source at 905nm wavelength. Further studies should be carried out to ensure that the present methodology could be extended to other laser systems (operating at different wavelengths) and other fruit varieties or branching structures.

Table 1. Reflectance analysis: The mean apparent reflectance and standard deviation of different elements in an apple orchard.

Tree	Elements	mean(R) [%]	std(R) [%]
T1	Leaves	29.23	13.57
T2	Leaves	28.69	13.88
T1	Trunks	29.67	14.83
T2	Trunks	28.52	15.41
T1	Apples	43.59	16.81
T2	Apples	45.10	16.78

The results of this analysis are the basis of the proposed detection algorithm, with reflectance being the principal feature used in the segmentation step. Although Fuji apples have higher reflectance than leaves and trunks at 905 nm, the standard deviation is high enough to create overlap between classes (Figure 7b). In order to find the optimal threshold, R_{th}, that will remove the points corresponding to leaves, branches, and trunks, a performance evaluation of the detection algorithm (Section 2.3) was carried out using different reflectance thresholds. Figure 7a plots the evolution of precision, recall, and F1-score metrics, computed before applying the false positive removal step, under different reflectance thresholds, R_{th}. The best results were obtained with an R_{th} = *60%,* resulting in an *F1-score=82.16%* for the training dataset. Figure 7b

shows the reflectance distributions for leaves, trunks, and apples. As can be seen, most of the 3D points belonging to apples were below the threshold value. This is because our restrictive threshold minimizes the false positives of leaves and trunks being selected as apples. Furthermore, omitting points as apple is not as critical as having a few apple points in a cluster, which are sufficient for detection.



a) **b)**

Figure 7. a) Precision (green dashed line), recall (red dotted line), and F1-score (blue solid line) versus the applied reflectance threshold; b) Gaussian distributions obtained for each tree element in the reflectance analysis of the training dataset. Green solid line corresponds to leaves, blue dotted line to trunks and red dashed line to fruits. The vertical dash-dotted line indicates the reflectance threshold used for fruit detection.

3.2. Step-by-step algorithm performance evaluation

This section presents a qualitative and quantitative evaluation of the different steps and methods implemented in this paper. Regarding the qualitative evaluation, Figure 8 illustrates the evolution after each processing step. First, Figure 8a shows an RGB image of one of the trees, which is incorporated to assist in visualization, but was not used in the algorithm. Figure 8b renders the 3D model obtained with the MTLS. The colour scale indicates the reflectance of each point, where blue corresponds to low values and red implies high reflectance. It is evident from this representation how Fuji apples exhibit higher reflectance than other tree elements. Figure 8c shows the results after applying the reflectance threshold, R_{th}, to the original point cloud. In this step, many of the leaf and trunk points were removed. Once the sparse outlier removal is applied (Figure 8d), zones with low point density were removed, leaving only groups of points which are candidates for apple detection. Figure 8e illustrates the segmentation output, which terminates the clustering of connected points. This result has clusters with

one apple (red, orange, blue, and purple), clusters with more than one apple (green), and false positives (grey). The apple detection algorithm ends by splitting clusters with more than one apple and removing false positives. The final result is presented in Figure 8f.

Table 2 presents the results of the test dataset for each step and method implemented. The first row shows the results after point cloud segmentation (Section 2.3.1); rows 2-4 indicate the results obtained when applying the splitting techniques presented in Section 2.3.2; and the last three rows present the final results after removing the false positives detected (Section 2.3.3).

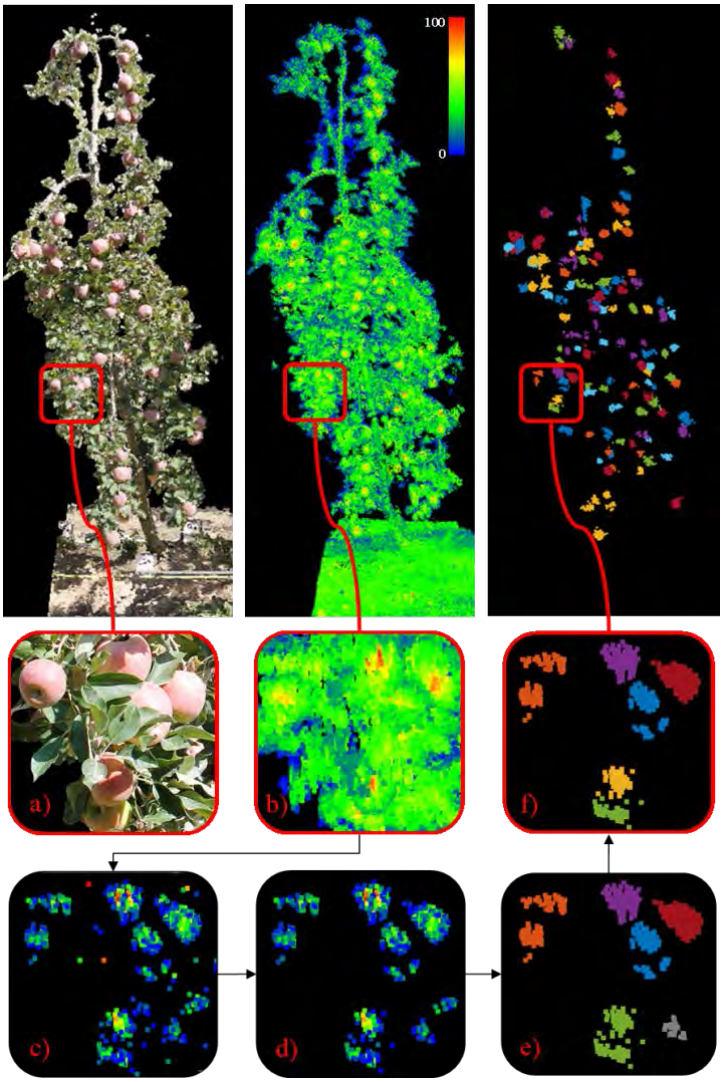


Figure 8. Illustration of the different processing steps (tree 2). a) RGB image. b) Point cloud obtained with the MTLS. c) Point cloud after applying the reflectance threshold. d) Sparse outlier removal. e) Connected component labelling (DBSCAN). f) Apple separation and false positive removal. For better visualization purposes, in b), scale ranges from 0 (blue) to 100 (red), while in c) and d) the scale ranges from 60 (blue) to 100 (red).

The localization success values obtained after point cloud segmentation (before apple separation and false positive removal) are slightly higher than 87% (first row). These results are similar to other methodologies using colour cameras (Gongal *et al*., 2015). The identification success presents significantly lower results (~73%), because of some detections containing more than one apple. Methods 1, 2, and 3 therefore were applied, in order to split these clusters, methods (Section 2.3.2). As a result, the identification success increased by more than 8% (rows 2 to 4), although the number of false positives also increased due to multi-detections. Method 1 performed best in terms of increasing the identification success (+11%), but also generated more multi-detections. Method 2 increased identification success by more than 8%, while false positives only increased 3%. The results of method 3 are a trade-off between the previous two methods. Since localization success performs an evaluation on a point-by-point basis, applying separation methods does not vary the results of this metric.

When applying false positive removal (rows 5 to 7), it is observed that the false detection rate fell by more than 5%, while the localization and identification successes were not affected (except for method 3, with a decrease of less than 1%). The best results were obtained by combining method 2 with false positive removal, resulting in a lower number of false positives, without affecting the performance of the apple detection algorithm.

The processing times indicated in Table 2 correspond to processing the data with a 64-bit operating system, with 8GB of RAM and an Intel ® Core(TM) i7-4500U processor (1.80 GHz, boosted to 2.40 GHz). Although method 2 was slightly more efficient than the other two approaches, no significant differences were observed in the processing time. This is because the most computationally intensive operation is in the DBSCAN clustering algorithm (9.1 seconds), which is part of the segmentation step included in all methods.

Table 2. Performance assessment of the different implemented steps and methods: point cloud segmentation (S); apple separation methods 1, 2, and 3 (M1, M2, and M3, respectively); and false positive removal step (FPr). Results include information from the test dataset (tree 3).

	Localization		Identification		
Method	$Success_L$ [%]	FDR_L [%]	$Success_{ID}$ [%]	FDR_{ID} [%]	Processing Time [s]
S	87.5	11.9	73.5	13.8	11.0
S + M1	87.5	20.7	85.3	26.1	11.9
S + M2	87.5	15.6	82.4	17.0	11.0
S + M3	87.5	16.8	84.6	20.7	12.0
S + M1 + FPr	87.5	12.5	85.3	18.3	12.1
S + M2 + FPr	87.5	9.8	82.4	10.4	11.1
S + M3 + FPr	86.8	11.3	83.8	14.9	12.4

3.3. Detection results

Table 3 shows the apple detection algorithm, as evaluated individually for each tree. These results were generated by applying the point cloud segmentation, followed by an apple separation using method 2, and removing false positives using the condition expressed in Section 2.3. The detection rate is similar for processed trees despite being slightly better for tree 1 and 3. A localization success of 87.5% with a 9.8% of FDR_L, an identification success of 82.4% with a 10.4% of $FDR_{ID,}$ and an F1-score of 85.8% were obtained using the test dataset. These results are comparable with those obtained with other methodologies used in the state of the art. So far, the best detection rates have been reported with image processing, obtaining accuracies of between 80% and 85% using colour features (Gongal *et al*., 2015), and up to 86% of recall using deep learning (Bargoti and Underwood, 2017). However, the vision systems used in harvesting robots in orchard environments report a mean value of 80% in localization success and a mean value of 70% in identification success (Bac *et al*., 2014). Although it is difficult to compare the research found in the state of the art review, given that they are evaluated with different datasets, the methodology presented in this paper yields similar detection rates to previous work based on colour cameras, with the advantage that LiDAR-based measurements are not affected by illumination conditions. Furthermore, the location of each detected apple is obtained directly, which makes the presented system very

interesting for autonomous harvesting or fruit load assessment for yield mapping applications.

Table 3. Apple detection assessment using method 2. Trees 1 and 2 were used as training dataset and tree 3 as test dataset. GT_{field} corresponds to the number of apples hand-counted in field, while GT_{labels} corresponds to the number of apples labelled in data. Other metrics are defined in Section 2.4.

			Localization				Identification				
Tree	GT_{field}	GT_{labels}	TP_L	FP_L	$Success_L$	FDR_L	TP_{ID}	FP_{ID}	$Success_{ID}$	FDR_{ID}	F1-score
Tree 1	139	133	116	12	87.2%	9.4%	110	16	82.7%	12.7%	0.849
Tree 2	145	138	118	13	85.5%	9.9%	109	15	79.0%	12.1%	0.832
Tree 3	139	136	119	13	87.5%	9.8%	112	13	82.4%	10.4%	0.858

Regarding the computational cost, Table 4 includes the inference time, processing each tree separately, and all trees combined. The number of points of each test is also reported. As expected, the computational time increases with the number of points processed. Results show that processing trees individually is much more efficient than processing all trees at once. This is because the average run time complexity of DBSCAN is not linear with the number of points (Ester et al., 1996), resulting in higher efficiency when processing small point clouds.

Table 4. Computational cost according to the number of points in the point cloud.

Tree	Nº of points	Processing Time [s]
Tree 1	438.260	8.0
Tree 2	460.847	9.6
Tree 3	526.136	11.2
Tree 1+2+3	1.425.243	68.8

4. Conclusions

This work presents a new methodology for Fuji apple detection and localization in real commercial orchard environments using a LiDAR-based mobile terrestrial laser scanner (MTLS) with reflectance capabilities. A reflectance analysis of the different apple tree elements was carried out, which showed that apples exhibit a higher reflectance than leaves and trunks at the 905 nm laser wavelength; we therefore conclude that this characteristic is a valuable feature for apple detection. An apple detection algorithm,

suitable for dealing with point clouds obtained with an MTLS, was subsequently developed and tested on three apple trees from a commercial apple orchard. The algorithm is divided into three steps: (1) removal of points corresponding to leaves and trunk and clustering the remaining points with a connected component labelling, (2) identification and splitting of clusters that contain more than one apple, and (3) false positive reduction. In order to predict the number of apples grouped in a cluster, three different methods were proposed: template matching, decision tree, and a combination of both approaches. The best results were achieved by applying a decision tree, resulting in a localization success of 87.5% with a 9.8% false detection rate, an identification success of 82.4% with a 10.4% false detection rate, and an F1-score of 85.8% in the test dataset. These outcomes represent an advance in the fruit detection field, since the results are comparable with those from colour (RGB) camera systems used in past efforts; however, the proposed LiDAR-based has the additional advantages that measurements are not affected by illumination conditions and that the method directly provides 3D fruit location information. An important limitation of this work is the small dataset. A larger dataset could allow the parameters to be learnt automatically (instead of being manually selected), thereby obtaining an algorithm that could better generalize with new data. Future efforts should include an analysis of fruit reflectance under different laser wavelengths, the extension of the dataset to other fruit varieties and species, and the application of machine learning algorithms in larger datasets.

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 (projects AGL2013-48297-C2-2-R and MALEGRA, TEC2016-75976-R) and the Spanish Ministry of Science, Innovation and Universities (project RTI2018-094222-B-I00). The Spanish Ministry of Education is thanked for Mr. J. Gené's pre-doctoral fellowships (FPU15/03355). The work of Jordi Llorens was supported by Spanish Ministry of Economy, Industry and Competitiveness through a postdoctoral position named Juan de la Cierva Incorporación (JDCI-2016-29464_N18003). We would also like to thank CONICYT/FONDECYT for grant 1171431 and CONICYT FB0008. Nufri (especially Santiago Salamero and Oriol Morreres) and Vicens Maquinària Agrícola S.A. are also thanked for their support during the data acquisition.

Appendix A. Parameter values and feature analysis

Table A1 presents the values set for each parameter used in the algorithm. Parameter R_{th} is used in the segmentation step. See more details about these parameters in Section 2.3.1. Parameters with sub-index *kj* refer to the thresholds used in Section 2.3.2 - method 2 and were selected after analysing the graphical representation of cluster features shown in Fig. A1. Parameters with sub-index *FP* correspond to the thresholds used to remove false positives (Section 2.3.3) and were selected after analysing the graphical representation of detection features shown in Fig. A2.

Table A1. Parameter values used to detect apples in the presented dataset. The first five parameters were used during the point cloud segmentation step. Parameters sub-indexed with letter *K* correspond to thresholds used in the apple separation step. Parameters sub-indexed with letters *FP* were used in the false positive removal step.

Symbol	Value	Units
R_{th}	60	%
k	20	Points
α	0	---
$minPts$	15	Points
ε	0.03	m
V_{k1}	1.5·10-4	m^3
P_{k1}	85	Points
$\bar{R}_{k1}$	67.5	%
Ψ_{k1}	0.8	---
P_{k2}	400	Points
Ψ_{k2}	0.6	---
V_{k2}	1.2·10-3	m^3
V_{k3}	1.6·10-3	m^3
Ψ_{FP}	0.46	---
$\bar{R}_{FP}$	65.25	%
V_{FP}	10^{-3}	m^3

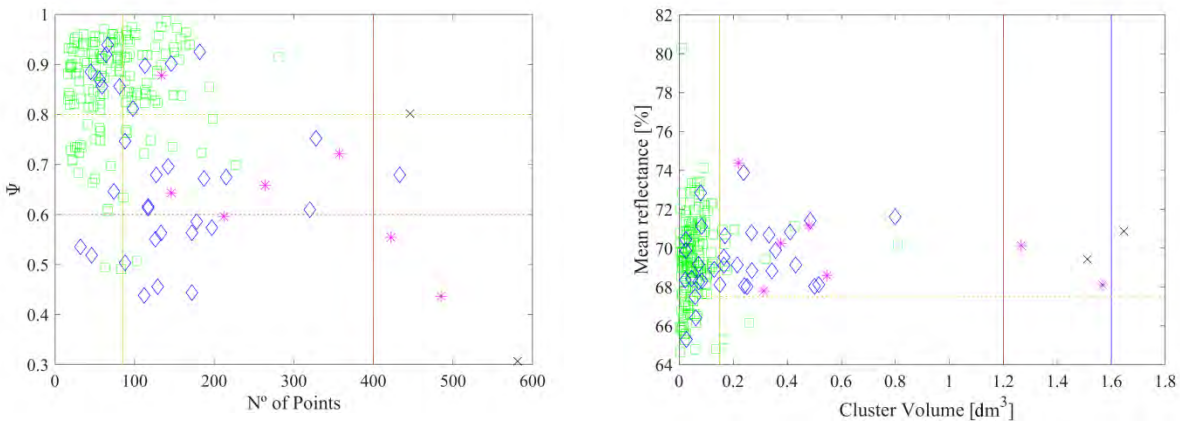


Figure A1. Graphical representation of cluster features. The features analysed are the geometric parameter, Ψ, and the number of points (left), and the mean reflectance and the cluster volume (right). Clusters with one apple are represented in green squares; clusters with two apples are represented in blue diamonds; clusters with three apples in magenta asterisks; and clusters with four apples or more in black crosses. Yellow, red and blue lines correspond to K1, K2 and K3 thresholds, respectively. This analysis was performed on the training data set (Trees 1 and 2) and was used to set the thresholds explained in Section 2.3.2 - method 2.

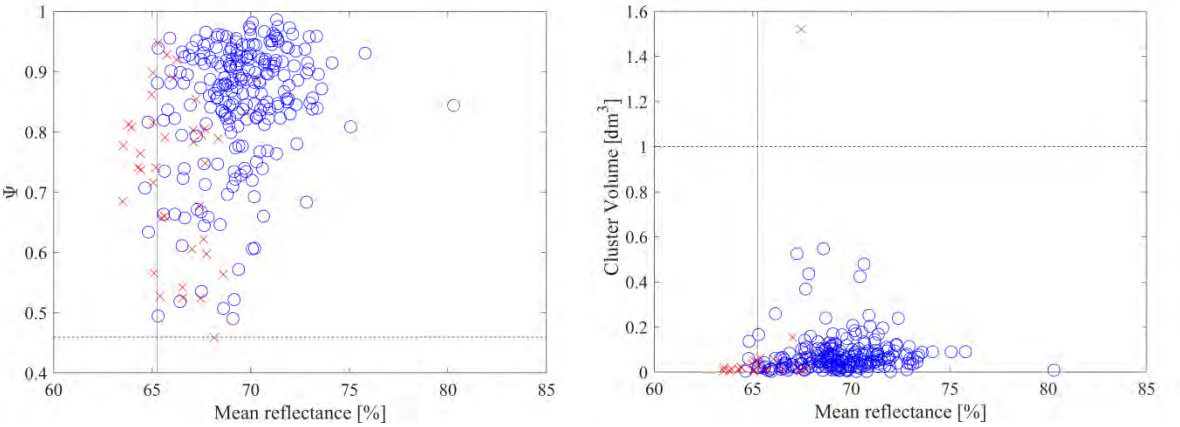


Figure A2. Graphical representation of detection features. The features analysed are the geometric parameter, Ψ, the mean reflectance and the cluster volume. False positives (FP) are represented by red crosses; true positives are represented by blue circles. Horizontal and vertical lines show the thresholds used to remove FP. This analysis was performed on the training data set (Trees 1 and 2) and was used to set the thresholds explained in Section 2.3.2 - method 2.

References

Auat Cheein, F., Torres-Torriti, M., Hopfenblatt, N.B., Prado, Á.J., Calabi, D., 2017. Agricultural service unit motion planning under harvesting scheduling and terrain constraints. *J. F. Robot.* 34, 1531–1542. doi:10.1002/rob.21738

Auat Cheein, F.A., Carelli, R., 2013. Agricultural robotics: Unmanned robotic service units in agricultural tasks. *IEEE Ind. Electron. Mag.* 7, 48–58. doi:10.1109/MIE.2013.2252957

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, 888–911. doi:10.1002/rob.21525

Bargoti, S., Underwood, J., 2017. Deep Fruit Detection in Orchards. 2017 *IEEE Int. Conf. Robot. Autom.* 3626–3633.

Barnea, E., Mairon, R., Ben-Shahar, O., 2016. Colour-agnostic shape-based 3D fruit detection for crop harvesting robots. *Biosyst. Eng.* 146, 57–70. doi:10.1016/j.biosystemseng.2016.01.013

Bechar, A., Vigneault, C., 2017. Agricultural robots for field operations. Part 2: Operations and systems. *Biosyst. Eng.* 153, 110–128. doi:10.1016/j.biosystemseng.2016.11.004

Bechar, A., Vigneault, C., 2016. Agricultural robots for field operations: Concepts and components. *Biosyst. Eng.* 149, 94–111. doi:10.1016/j.biosystemseng.2016.06.014

Brunelli, R., 2009. Template Matching Techniques in Computer Vision - Theory And Practice. *John Wiley & Sons.*

Bulanon, D.M., Burks, T.F., Alchanatis, V., 2009. Image fusion of visible and thermal images for fruit detection. *Biosyst. Eng.* 103, 12–22. doi:10.1016/j.biosystemseng.2009.02.009

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

Cariou, C., Lenain, R., Thuilot, B., Berducat, M., 2009. Automatic guidance of a four-wheel-steering mobile robot for accurate field operations. *J. F. Robot.* 26 (6–7), 504–518. doi:10.1002/rob.20282

Chaivivatrakul, S., Dailey, M.N., 2014. Texture-based fruit detection. *Precis. Agric.* 15, 662–683. doi:10.1007/s11119-014-9361-x

De-An, Z., Jidong, L., Wei, J., Ying, Z., Yu, C., 2011. Design and control of an apple harvesting robot. *Biosyst. Eng.* 110, 112–122. doi:10.1016/j.biosystemseng.2011.07.005

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.* doi:10.1.1.71.1980

Foglia, M.M., Reina, G., 2006. Agricultural robot for radicchio harvesting. *J. F. Robot.* 23 (6–7), 363–377. doi:10.1002/rob.20131

Font, D., Pallejà, T., Tresanchez, M., Runcan, D., Moreno, J., Martínez, D., Teixidó, M., Palacín, J., 2014. A proposal for automatic fruit harvesting by combining a low cost stereovision camera and a robotic arm. *Sensors (Switzerland)* 14, 11557–11579. doi:10.3390/s140711557

Gong, A., Yu, J., He, Y., Qiu, Z., 2013. Erratum to " Citrus yield estimation based on images processed by an Android mobile phone ." *Biosyst. Eng.* 116, 111–112. doi:10.1016/j.biosystemseng.2013.07.004

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

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* 31 (8), 651–666. doi:10.1016/j.patrec.2009.09.011

Jiménez, A.R., Ceres, R., Pons, J.L., 2000. A vision system based on a laser range-finder applied to robotic fruit harvesting. *Mach. Vis. Appl.* 11, 321–329. doi:10.1007/s001380050117

Jiménez, A.R., Jain, A.K., Ceres, R., Pons, J.L., 1999. Automatic fruit recognition: a survey and new results using Range/Attenuation images. *Pattern Recognit.* 32, 1719–1736. doi:10.1016/S0031-3203(98)00170-8

Kaasalainen, S., Jaakkola, A., Kaasalainen, M., Krooks, A., Kukko, A., 2011. Analysis of incidence angle and distance effects on terrestrial laser scanner intensity: Search for correction methods. *Remote Sens.* 3, 2207–2221. doi:10.3390/rs3102207

Kukko, A., Kaasalainen, S., Litkey, P., 2008. Effect of incidence angle on laser scanner intensity and surface data. *Appl. Opt.* 47, 986–992. doi:10.1364/ao.47.000986

Kurtulmus, F., Lee, W.S., Vardar, A., 2014. Immature peach detection in colour images acquired in natural illumination conditions using statistical classifiers and neural network. *Precis. Agric.* 15, 57–79. doi:10.1007/s11119-013-9323-8

Lak, M.B., Minaei, S., Amiriparian, J., Beheshti, B., 2010. Apple fruits recognition under natural luminance using machine vision. *Adv. J. Food Sci. Technol.* 2, 325–327.

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

Linker, R., Cohen, O., Naor, A., 2012. Determination of the number of green apples in RGB images recorded in orchards. *Comput. Electron. Agric.* 81, 45–57. doi:10.1016/j.compag.2011.11.007

Liu, X., Zhao, D., Jia, W., Ruan, C., Tang, S., Shen, T., 2016. A method of segmenting apples at night based on color and position information. *Comput. Electron. Agric.* 122, 118–123. doi:10.1016/j.compag.2016.01.023

Maldonado, W., Barbosa, J.C., 2016. Automatic green fruit counting in orange trees using digital images. Comput. Electron. Agric. 127, 572–581. doi:10.1016/j.compag.2016.07.023

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

Okamoto, H., Lee, W.S., 2009. Green citrus detection using hyperspectral imaging. *Comput. Electron. Agric.* 66, 201–208. doi:10.1016/j.compag.2009.02.004

Payne, A., Walsh, K., Subedi, P., Jarvis, D., 2014. Estimating mango crop yield using image analysis using fruit at “stone hardening” stage and night time imaging. *Comput. Electron. Agric.* 100, 160–167. doi:10.1016/j.compag.2013.11.011

Qureshi, W.S., Payne, A., Walsh, K.B., Linker, R., Cohen, O., Dailey, M.N., 2017. Machine vision for counting fruit on mango tree canopies. *Precis. Agric.* 18, 224–244. doi:10.1007/s11119-016-9458-5

Ray, T.W., 1994. A FAQ on vegetation in remote sensing. California: Div. of Geological and Planetary Sciences California Institute of Technology.

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. *Adv. Agron.* 133, 71–112. doi:10.1016/bs.agron.2015.05.002

Rosell-Polo, J.R., Gregorio, E., Gene, J., Llorens, J., Torrent, X., Arno, J., Escola, A., 2017. Kinect v2 Sensor-based Mobile Terrestrial Laser Scanner for Agricultural Outdoor Applications. *IEEE/ASME Trans. Mechatronics* 22, 2420–2427. doi:10.1109/TMECH.2017.2663436

Rusu, R.B., Marton, Z.C., Blodow, N., Dolha, M., Beetz, M., 2008. Towards 3D Point cloud based object maps for household environments. *Rob. Auton. Syst.* 56, 927–941. doi:10.1016/j.robot.2008.08.005

Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T., McCool, C., 2016. DeepFruits: A Fruit Detection System Using Deep Neural Networks. *Sensors* 16, 1222. doi:10.3390/s16081222

Safren, O., Alchanatis, V., Ostrovsky, V., Levi, O., 2007. Detection of Green Apples in Hyperspectral Images of Apple-Tree Foliage Using machine Vision. *Trans. ASABE* 50, 2303–2313. doi:10.13031/2013.24083

Si, Y., Liu, G., Feng, J., 2015. Location of apples in trees using stereoscopic vision. *Comput. Electron. Agric.* 112, 68–74. doi:10.1016/j.compag.2015.01.010

Siegel, K.R., Ali, M.K., Srinivasiah, A., Nugent, R.A., Narayan, K.M.V., 2014. Do we produce enough fruits and vegetables to meet global health need? *PLoS One* 9 (8), e104059. doi:10.1371/journal.pone.0104059

Stajnko, D., Lakota, M., Hocevar, M., 2004. Estimation of number and diameter of apple fruits in an orchard during the growing season by thermal imaging. *Comput. Electron. Agric.* 42, 31–42. doi:10.1016/S0168-1699(03)00086-3

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

Velodyne, L., 2016. VLP-16 In VLP-16 Manual: User's Manual and Programming Guide; Velodyne LiDAR.

Wachs, J.P., Stern, H.I., Burks, T., Alchanatis, V., 2010. Low and high-level visual feature-based apple detection from multi-modal images. *Precis. Agric.* 11, 717–735. doi:10.1007/s11119-010-9198-x

Wehr, A., Lohr, U., 1999. Airborne laser scanning - An introduction and overview. *ISPRS J. Photogramm. Remote Sens.* 54, 68–82. doi:10.1016/S0924-2716(99)00011-8

Xiang, R., Jiang, H., Ying, Y., 2014. Recognition of clustered tomatoes based on binocular stereo vision. *Comput. Electron. Agric.* 106, 75–90. doi:10.1016/j.compag.2014.05.006

Zhang, B., Huang, W., Wang, C., Gong, L., Zhao, C., Liu, C., Huang, D., 2015. Computer vision recognition of stem and calyx in apples using near-infrared linear-array structured light and 3D reconstruction. *Biosyst. Eng.* 139, 25–34. doi:10.1016/j.biosystemseng.2015.07.011

Zhang, Q., Pierce, F.J., 2016. Agricultural automation: fundamentals and practices. *CRC Press.*

Zhao, C., Lee, W.S., He, D., 2016. Immature green citrus detection based on colour feature and sum of absolute transformed difference (SATD) using colour images in the citrus grove. *Comput. Electron. Agric.* 124, 243–253. doi:10.1016/j.compag.2016.04.009

Zhao, Y., Gong, L., Huang, Y., Liu, C., 2016. A review of key techniques of vision-based control for harvesting robot. *Comput. Electron. Agric.* 127, 311–323. doi:10.1016/j.compag.2016.06.022

Chapter V. P5: Fruit detection, yield prediction and canopy geometric characterization using LiDAR with forced air flow

This chapter was published in *Computers and Electronics in Agriculture* 168 (2020) 105121, https://doi.org/10.1016/j.compag.2019.105121:

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Fruit detection, yield prediction and canopy geometric characterization using LiDAR with forced air flow

Jordi Gené-Mola[a,*], Eduard Gregorio[a], Fernando Auat Cheein[b], Javier Guevara[b], Jordi Llorens[a], Ricardo Sanz-Cortiella[a], Alexandre Escolà[a], Joan R. Rosell-Polo[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 Electronic Engineering, Universidad Técnica Federico Santa María, Valparaíso, Chile

Abstract

Yield monitoring and geometric characterization of crops provide information about orchard variability and vigor, enabling the farmer to make faster and better decisions in tasks such as irrigation, fertilization, pruning, among others. When using LiDAR technology for fruit detection, fruit occlusions are likely to occur leading to an underestimation of the yield. This work is focused on reducing the fruit occlusions for LiDAR-based approaches, tackling the problem from two different approaches: applying forced air flow by means of an air-assisted sprayer, and using multi-view sensing. These approaches are evaluated in fruit detection, yield prediction and geometric crop characterization. Experimental tests were carried out in a commercial Fuji apple (*Malus domestica* Borkh. cv. Fuji) orchard. The system was able to detect and localize more than 80% of the visible fruits, predict the yield with a root mean square error lower than 6% and characterize canopy height, width, cross-section area and leaf area. The forced air flow and multi-view approaches helped to reduce the number of fruit occlusions, locating 6.7 % and 6.5 % more fruits, respectively. Therefore, the proposed system can potentially monitor the yield and characterize the geometry in apple trees. Additionally, combining trials with and without forced air flow

and multi-view sensing presented significant advantages for fruit detection as they helped to reduce the number of fruit occlusions.

Keywords: Apple detection; Fruit counting; Yield prediction; 3D plant modeling; Geometric characterization.

Nomenclature

af	Trial scanning with forced air flow
D	Number of fruit detections
DR	Detection rate
E	Trial scanning from the east side
FoV	Field-of-view
FP	False positive detection
FPR	False positive rate
FPr	False positive removal step
FS	Fruit separation step
GT_{field}	Number of apples manually counted in the orchard
GT_{labels}	Number of apples labelled in the 3D point cloud
H1	Trial scanning with LiDAR sensor at 1.8m height
H2	Trial scanning with LiDAR sensor at 2.5m height
IMU	Inertial measurement unit
K	Number of apples in a cluster of 3D points
LD	Number of labels detected
MD	Multi-detection
MDR	Multi-detection rate
n	Trial scanning without forced air flow application
P	Precision
Pp	Pre-processing
Pt	Number of points that contain a cluster
R	Recall
$\bar{R}$	Mean reflectance of the points of a cluster
rh	Reflectance histogram
RMSE	Root mean square error
RTK-GNSS	Real-time kinematics global navigation satellite system
SVM	Support-vector-machine
T	Total number of fruits in the dataset
TP	True positive
V	Volume of a cluster
W	Trial scanning from the west side
w.r.t.	abbreviation of “with respect to”
λ_n	Normalized eigenvalues
σ_r	Standard deviation reflectance of cluster points
Ψ	Geometric parameter

1. Introduction

Agricultural farms are required to increase production while reducing the environmental impact in a sustainable way (Tilman et al., 2011). Although mechanization and the evolution of agricultural machinery in extensive fields have helped to enhance crop production efficiency (Bechar and Vigneault, 2016), most of the intensive orchards are still being managed in similar ways to traditional farming methods, with labor intensive operations and without addressing in-field spatial variability (Uribeetxebarria et al., 2019). To meet food supply and environmental demands, precision agriculture aims to find new strategies that allow the farmer for a more efficient management of orchards, reducing the amount of inputs while increasing fruit quality and productivity (Vázquez-Arellano et al., 2016).

To confront these challenges, orchard vigor and variability need to be better understood (Kamilaris and Prenafeta-Boldú, 2018). To this end, orchard characterization through information and communication technologies plays a crucial role, as shown in (Colaço et al., 2018a, 2018b; Narvaez et al., 2017). Obtaining an accurate characterization of trees by non-destructive methods at different growth stages provides valuable information that can be used for enhancing precision in orchard management (Andújar et al., 2017; Rosell and Sanz, 2012). This characterization can include phenology monitoring, plant geometric characterization and yield monitoring, among others.

In the last decade, different sensors and methods have been used for the geometric characterization of tree orchards. Due to the unstructured and complex nature of agricultural environments, with variable canopy structures in depth, porosity, training systems, among others (Vázquez-Arellano et al., 2016), the acquisition of 3D information –from depth cameras, structure-from-motion approaches, stereo vision and light detection and ranging (LiDAR) sensors– showed the most promising results in terms of orchard characterization and plant reconstruction, as has been shown in Rosell and Sanz (2012) and Yandún Narváez et al. (2016). Sensors and sensing techniques for the 3D modeling of orchards have been used to estimate parameters such as crop growth, height, shape and leaf area, with applications in pesticide treatments, irrigation,

fertilization, pruning and crop training (Jiménez-Brenes et al., 2017; Narvaez et al., 2017; Pfeiffer et al., 2018; Sanz et al., 2018; Tagarakis et al., 2018).

For yield mapping and monitoring, the most commonly used sensors are RGB cameras (Gongal et al., 2015; Linker, 2017; Maldonado and Barbosa, 2016; Zhao et al., 2016). The main disadvantage of RGB cameras is that they only provide 2D information. Advances in artificial intelligence and computer vision have led to remarkable progress in fruit detection (Bargoti and Underwood, 2017a; Gan et al., 2018; Sa et al., 2016). Nevertheless, fruit detection performance continues to be affected by extrinsic factors that do not depend on the robustness of the algorithm, including the occlusion of fruits by other vegetative organs or lighting conditions (Bac et al., 2014; Gongal et al., 2015; Liu et al., 2016). Very few studies have tackled the problem of occlusions. Bulanon et al. (2009) and Gongal et al. (2018) proposed the use of multi-view imaging systems, while other works have used supportive tools such as an air blower to reduce melon leaf occlusions (Edan et al., 2000) or a mechanism to reduce canopy volume in citrus trees (Lee and Rosa, 2006). Other studies have considered variable lighting conditions, some of which have tried to minimize variable illumination effects by converting images to other color spaces (Payne et al., 2013; Zhou et al., 2012), while others have proposed working with artificial lighting, although this involves the use of tunnel structures around trees or working at night time (Gongal et al., 2016; Linker and Kelman, 2015; Payne et al., 2014).

Other 2D sensors, including thermal, multispectral and hyperspectral cameras, have also shown potential for fruit detection (Bulanon et al., 2008; Okamoto and Lee, 2009; Sa et al., 2016; Safren et al., 2007; Zhang et al., 2015), although their use is not as extended as color cameras due to their cost and the high level of training required for their operation (Linker, 2018). The recent evolution in the fields of sensing and photonics has led to the introduction of the use of 3D sensors. Depth cameras based on stereoscopy, structured light or time-of-flight (ToF) are the most commonly used 3D sensors for fruit detection (Gené-Mola et al., 2019b; Gongal et al., 2015). The main limitation of these sensors is that their performance decreases under direct sunlight (Rosell-Polo et al., 2015).

LiDAR sensors have been widely applied for geometric characterization (Rosell and Sanz, 2012; Vázquez-arellano et al., 2016). However, their use is marginal for fruit detection and yield monitoring, probably because they are more expensive than other alternative sensors (such as RGB cameras) and do not provide color data. However, in addition to providing 3D information of the scene, one of the advantages of using LiDAR sensors is that the measurements are not affected by lighting conditions. Previous studies using LiDAR sensor data for fruit detection have been carried out in lab conditions and tested with a limited dataset, from 7 to 114 fruits (Feng et al., 2012; Gotou et al., 2003; Jiménez et al., 2000; Tanigaki et al., 2008). In other areas, LiDAR sensors have been placed on robotic platforms to infer the distance between the fruit and the robotic arm, though the actual detection of the fruit was performed using other systems such as RGB imaging (Bulanon and Kataoka, 2010; Ceres et al., 1998; Stein et al., 2016; Yin et al., 2009). More recently, the authors of the present study presented a proof of concept of the usefulness of LiDAR for apple detection in real commercial orchard environments (Gené-Mola et al., 2019a).

In this work, a multi-beam LiDAR sensor is used for remote fruit detection and plant geometrical characterization in a commercial Fuji apple orchard (*Malus domestica* Borkh. cv. Fuji). A fruit detection algorithm based on reflectance thresholding and Support Vector Machine (SVM) was developed. Under the hypothesis that moving the tree foliage and using multi-view sensing will reduce the number of fruit occlusions and will increase the percentage of fruits detected, the system was mounted on an air-assisted sprayer used to generate forced air flow. The rest of the paper is structured as follows: Section 2 presents the experimental setup, the fruit detection algorithm, and the methodology carried out to predict the yield and characterize the canopy; Section 3 evaluates the fruit detection algorithm and the effect of using air action and different sensor positioning in terms of fruit detection accuracy, yield prediction estimation and geometric characterization performance; Section 4 discusses the performance of the system and compares the presented methodology to other works from the state of the art; finally, the conclusions retrieved from this work are presented in Section 5.

2. Methods

2.1. Experimental set up

Data was acquired in a commercial Fuji apple orchard (*Malus domestica* Borkh. cv. Fuji) (Figure 1), located in Agramunt, Catalonia, Spain (E: 336297 m, N: 4623494 m, 312 m a.s.l., UTM 31T - ETRS89). The scanning was carried out 3 weeks before harvesting, at BBCH (Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie) (Meier, 2001) growth stage 85. Trees grown in the selected orchard were 8-years-old and were trained in a tall spindle system with a maximum canopy height of 3.5-4 m, width of 1-1.5 m, and tree spacing of 4x1 m. All tests presented in this paper were carried out on 11 consecutive Fuji apple trees containing a total of 1444 apples (Table 1).

The equipment used for data acquisition was a mobile terrestrial laser scanner (MTLS) system with a multi-beam LiDAR sensor and a real-time kinematics global navigation satellite system (RTK-GNSS). A Puck VLP-16 (Velodyne LIDAR Inc., San José, CA, USA) LiDAR sensor was placed on a vertical plane to scan with a vertical field-of-view (FoV) of 360º, emitting 16 laser beams distributed in a horizontal FoV of 30º. In other words, each laser beam had a unique scanning angle, ranging from +15º to -15º, with a 2º step between the scanning angles. For each scan, the sensor provides a 3D point cloud with calibrated reflectance values (at 905 nm wavelength) of the measured scene, reporting values from 0-100 for diffuse reflectivities from 0% to 100% (Velodyne, 2016). This reflectance calibration was carried out by the sensor manufacturer, and allows getting reflectance values independently of laser power and distance. The LiDAR sensor acquisition frequency rate was set to 10 Hz (10 scans per second), corresponding to a vertical angular resolution of 0.2º. A GPS1200+ (Leica Geosystems AG, Heerbrugg, Switzerland) RTK-GNSS was used, with an absolute error of 0.01/0.02 m (horizontal / vertical), providing positioning measurements with rate of 20 Hz. Each sensor was connected to a rugged laptop, and data were synchronized by acquisition time stamp.

Figure 1. Tested Fuji apple orchard.

The MTLS system was placed on an air-assisted sprayer and was pulled by a tractor at 0.125 m/s forward speed along a linear trajectory parallel to the row of trees. Since the MTLS system did not include an inertial measurement unit (IMU), moving the system at low speed (0.125 m/s) and along a linear trajectory was important to reduce vibrations (in amplitude) and obtain precise point clouds -without misalignments between different scans-. The air-assisted sprayer was used to generate turbulent air with the aim of moving the tree foliage and dis-occlude apples behind the leaves. As the LiDAR sensor was oriented vertically and the scanning plane was orthogonal to the canopy, the LiDAR measuring area was the area under the air flow influence (Figure 2). The data acquired contains measurements from two different LiDAR heights (H_1 and H_2) and two different air conditions (*n* and *af,* where *n* stands for measurements without air flow action; and *af* for measurements with air flow action). The LiDAR sensor position H_1 corresponded to a sensor height of 1.8 m (approximately half of the tree height), while H_2 corresponded to the measurements with a LiDAR height of 2.5 m. In order to generate forced air, the air-assisted sprayer operated at 18π rad s^{-1} (540 rpm of PTO, power take-off angular speed). In these conditions, the sprayer fan generated an air flow speed of 5.5 ± 2.3 m s^{-1}, measured at a distance of 2.4 m from the fan axial center (the approximate horizontal distance between the sensor and the axis/trunk of the measured trees). A total of 4 trials were carried out, corresponding to all possible combinations between sensor heights and air conditions: $H_{1,n}$, $H_{1,af}$, $H_{2,n}$ and $H_{2,af}$. In

order to obtain a complete representation of the canopy, for each trial, the tree row was scanned from the west (*W*) and from the east (*E*) side. Trials merging both scanned sides are denoted as (*E+W*).

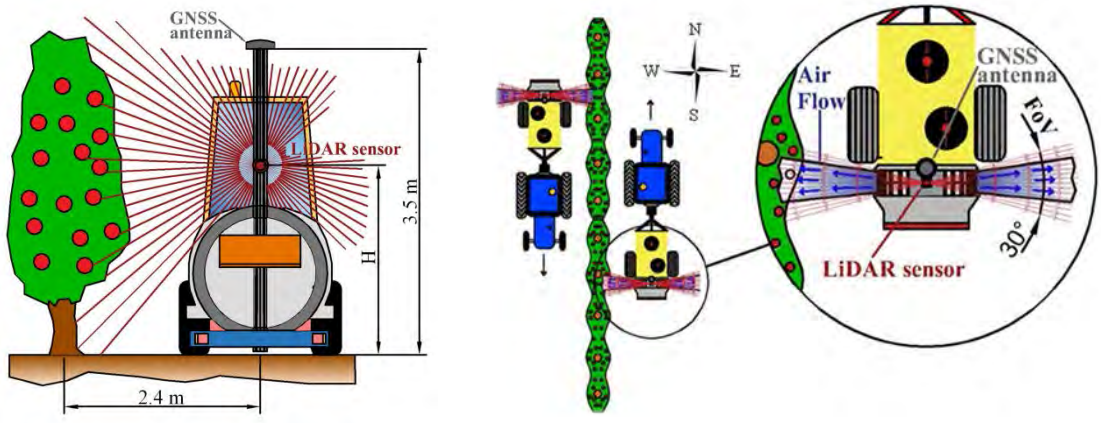


Figure 2. Diagram of the MTLS and the arrangement of its elements. a) Rear view.b) Top view. The zoom-in circle shows the position of the LiDAR sensor with respect to the sprayer fan. Air flow is represented in blue, while red lines illustrate the 16 laser beams distributed along the sensor field-of-view (FoV).

For each scan, a 3D point cloud with coordinates relative to the sensor was provided by the LiDAR system. To generate the point clouds with absolute coordinates, the RTK-GNSS data were used to infer the position and the orientation of the sensor, obtaining a rotation and translation matrix that transformed points in relative coordinates to points in absolute global world coordinates. To generate ground truth of the apples locations (GT_{labels}), the resulting point clouds were manually labelled, placing 3D rectangular bounding boxes around each apple, as shown in Figure 3c. This annotation was carried out using the software CloudCompare (Cloud Compare [GPL software] v2.9 Omnia) and supported by additional RGB images of the tested trees. A total of 1353 fruits were visually identified in the point cloud during ground truth generation, representing the 93.7% of the total amount of fruits manually counted (GT_{field}) in the orchard. Thus, 6.3% of apples were discarded since they were not visible in the 3D point clouds. Table 1 shows the number of fruits per tree manually counted in the orchard GT_{field} compared with the number of labelled apples in the 3D point cloud GT_{labels}. The dataset generated and analyzed during the current study has been made publicly available at *http://www.grap.udl.cat/en/publications/LFuji_air_dataset.html*.

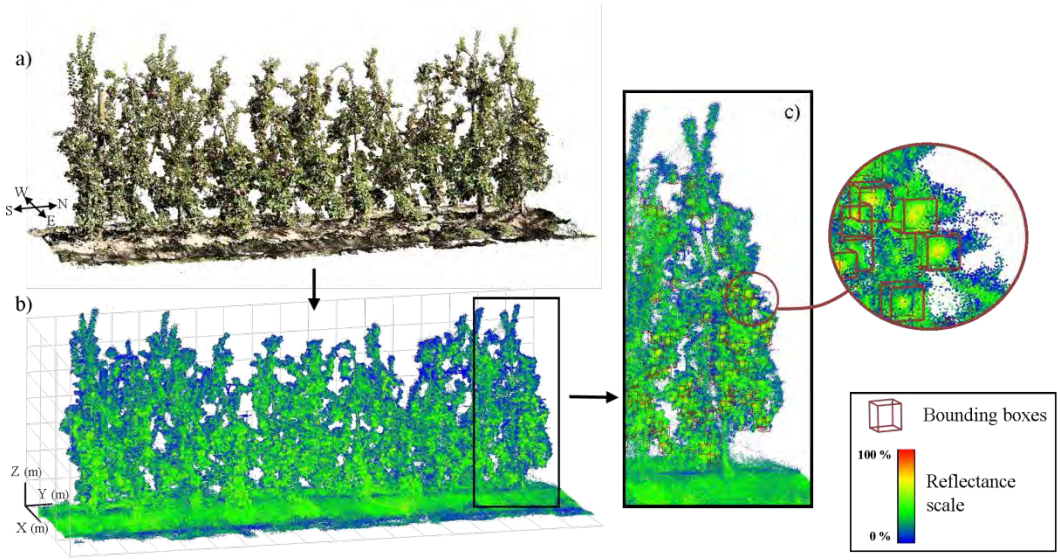


Figure 3. Illustration of the dataset generated for the current study. a) RGB image of the measured Fuji apple trees. b) Point cloud data generated from MTLS measurements. Color scale ranges from 0% (blue) to 100% (red) corresponding to the calibrated reflectance of the measured scene. c) Annotated point cloud with 3D rectangular bounding boxes placed around each apple.

Table 1. Fruit counting ground truth. Comparison between the number of fruits manually counted (GT_{field}) in the orchard and the number of fruits annotated in the 3D point cloud (GT_{labels}).

	GT_{field}	GT_{labels}
Tree 01	139	138
Tree 02	106	100
Tree 03	139	131
Tree 04	137	129
Tree 05	94	85
Tree 06	131	119
Tree 07	119	114
Tree 08	145	137
Tree 09	139	131
Tree 10	136	122
Tree 11	159	147
Total	1444	1353

2.2. Fruit detection algorithm

With the purpose of detecting and locating fruits from the MTLS data, the algorithm presented in Gené-Mola et al. (2019a) was implemented but lightly modified. The main modification included the use of an SVM approach (Cortes and Vapnik, 1995) in order

to avoid using manually set parameters and to automatically train the features that characterize apples. The algorithm consists of 4 main steps: pre-processing, fruit clustering, fruit separation and false positive removal (Figure 4). To combine different scans, such as data from east and west sides, from different sensor heights or different air flow conditions, the 3D point clouds were merged before applying the fruit detection algorithm. The registration of different point clouds was automatic since all scans were georeferenced in absolute world coordinates. Merging different scans in a single point cloud allows for reducing the number of multi-detections, e.g., points from an apple appearing in two different scans (such as from east and west sides) were merged in a unique point cloud, and in consequence, the apple was detected only once.

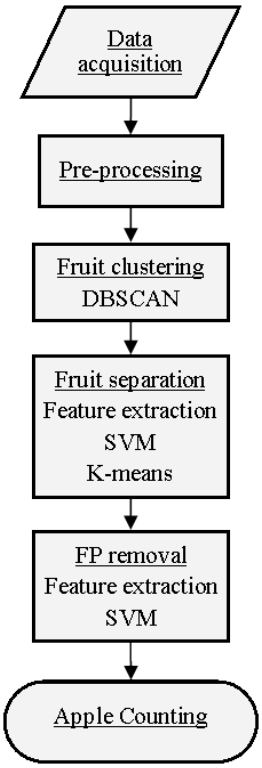


Figure 4. Fruit detection algorithm pipeline.

Pre-processing was based on the fact that apples have a higher IR reflectance than background. In this step, a reflectance threshold was set in order to remove the points that are not likely to belong to an apple. Then, sparse outlier removal (Rusu et al., 2008) was applied to remove noisy points. After removing background points, the remaining points were clustered in groups of connected points by applying a density-based scan algorithm DBSCAN(Ester et al., 1996). These first two steps (pre-processing and fruit

clustering) were implemented following the algorithm presented in Gené-Mola et al. (2019a).

The minimum distance used in DBSCAN to cluster connected points was ε = 0.03 m. Therefore, points belonging to two different fruits closer than ε were grouped in the same cluster. A fruit separation step was applied to discriminate the aforementioned clusters. First, the features of each cluster (volume, number of points, eigenvalues, and reflectance) were extracted. Then, a linear SVM with a penalty factor of *C*=0.35 (Burges, 1998) was used to predict the number of fruits (*K*) that contains each cluster. Clusters that had more than one apple (*K* > *1*) were split into *K* sub-clusters using the K-means algorithm (Jain, 2010).

The last step of the algorithm was a false positive filter. False positives are the detections that have been wrongly classified as apple. These false positives were derived from elements –such as trunks or leaves– that had reflectance values higher than expected (R > 60%), or from clusters that were wrongly split, detecting apples more than once (multi-detections). An SVM was used to classify each cluster as a correct or wrong detection. The SVMs used in the fruit separation and false positive removal steps were fed with 8 cluster features:

- Cluster volume V.
- Number of points Pt that contain a cluster.
- Normalized eigenvalues $\lambda_n = [\lambda_{1n}, \lambda_{2n}, \lambda_{3n}]$. The eigenvalues are obtained by singular value decomposition (SVD), and their value depends on the variance of points (3D data) projected on their principal axes (Jolliffe, 2011). In order to compare different clusters, an eigenvalues normalization is applied, so that the sum of normalized eigenvalues is one.
- Geometrical parameter $\Psi = 27 \cdot \lambda_{1n} \cdot \lambda_{2n} \cdot \lambda_{3n}$ defined as the product of normalized eigenvalues and a normalization factor equal to 27. Since the maximum value of the product of the normalized eigenvalues is equal to $\frac{1}{27}$, achieved when all normalized eigenvalues are equal to $\frac{1}{3}$ (for spherical clusters), the normalization factor of 27 allows the geometrical parameter to be bounded between 0 and 1, being 1 for spherical clusters.

- Reflectance histogram $rh = [rh_1, rh_2, rh_3, rh_4, rh_5] / P$, where $rh_1, rh_2, rh_3, rh_4, rh_5$ are the number of points in the cluster with a reflectance between 60:68, 68:76, 76:84, 84:92 and 92:100, respectively.
- Mean reflectance of cluster points $\bar{R}$.
- Standard deviation reflectance of cluster points σ_R.
- Maximum reflectance of cluster points R_{max}.

All processing presented in this work was implemented using MATLAB® (R2018a, Math Works Inc., Natick, Massachusetts, USA) and is publicly available jointly with the corresponding dataset at *http://www.grap.udl.cat/en/publications/LFuji_air_dataset.html*

2.3. Canopy characterization

The mean canopy height, width, contour and cross-section area were computed following the methodology described in Escolà et al. (2017). The row of trees was splitted into vertical slices of 0.1 m length. For each slice, the maximum canopy height and width were computed as the distance between the two most distant points in the vertical and horizontal directions (denoted as *Z* and *X* axis in Figure 3b, respectively). Then, the mean height and mean width were calculated as the mean of the maximum canopy heights and maximum canopy widths of all slices (110 slices for all 11 trees). The mean canopy contour was obtained similarly, computing the mean canopy width at different tree heights intervals of 0.1 m (40 heights intervals). The area within the mean canopy contour was defined as the cross-section area.

Finally, the leaf area was estimated using the projected tree row surface (PTRS) described in Sanz et al. (2018), which correlates the frontal projected surface and the top projected surface with the linear leaf area (leaf area per meter).

2.4. Performance evaluation

The effect of using forced air action and different sensor positions (multi-view sensing) was evaluated in terms of fruit detection accuracy, yield prediction estimation and geometric characterization performance.

To evaluate the fruit detection accuracy, each detection obtained using the fruit detection algorithm was classified as one of the following groups:

- True positive (TP): Detections that intersect with a ground truth apple label (bounding box annotation) with an overlap higher than 50%. In case of multi-detections, only one true positive was counted.
- False positive (FP): Detections that do not intersect with an annotation with an overlapping higher than 50%.
- Multi-detection (MD): A multi-detection is produced when a single apple is detected *n* times (by different detections). That could happen, for instance, if a single apple detection was wrongly split in *K* detections when applying the fruit separation step. In that case, it is counted one TP and *n - 1* multi-detections MD.

Having the total amount of TP, FP, MD, and the number of labels detected (LD), the fruit detection accuracy is assessed in terms of detection rate (DR), recall (R), precision (P), false positive rate (FPR), multi-detection rate (MDR) and F_1-score, as follows:

$$DR = \frac{LD}{T}, \quad (1)$$

$$R = \frac{TP}{T}, \quad (2)$$

$$P = \frac{TP}{D}, \quad (3)$$

$$FPR = \frac{FP}{D}, \quad (4)$$

$$MDR = \frac{MD}{D}, \quad (5)$$

$$F_1 = 2\frac{R \cdot P}{R+P}, \quad (6)$$

where D is the number of fruit detections and T is the total number of fruits in the dataset.

Since the algorithm used for fruit detection needs to be trained, the test has to be performed using a new dataset, different to the one used for training. To do so, an 11-fold cross-validation was practiced (11 iterations). Each iteration evaluated one tree, while the other trees were used as training set. The final test results presented in

sections 3.1 and 3.2 were obtained aggregating TP, FP and MD from all iterations and computing the metrics previously defined for all the dataset. Section 3.1 reports results after the pre-processing (Pp), fruit separation (FS) and false positive removal (FPr) steps. Different combinations of features used in the FS and FPr steps were evaluated in order to assess the usefulness of using each feature. This evaluation was carried out using data acquired from sensor height H_1, without forced air action and from both row sides ($H_{1,n,(E+W)}$). On the other hand, section 3.2 evaluates, both qualitatively and quantitatively, the fruit detection performance under different conditions and analyzes the effect of using forced air flow and scanning at different sensor heights. Results are reported either with respect to (w.r.t.) GT_{field}, as well as w.r.t. GT_{labels}. These two different approaches allow comparing the present methodology to other fruit detection works evaluated w.r.t. the number of visible fruits (Gongal et al., 2015).

The yield prediction was also evaluated following the 11-fold cross-validation model. First, the fruit detection algorithm was used to automatically count the number of fruits on each tree. Then, 11 iterations were performed in order to assess the yield prediction on each tree. In yield prediction, what is important is not so much the percentage of fruits detected but rather the correlation that exists between the number of detections and the actual number of fruits in the tree (Linker, 2017). Thus, a simple linear regression model $y = a \cdot x + b$ was obtained using the training set. This model related the number of fruits detected D and the actual number of fruits manually counted in the field GT_{field}. Then, the linear model was used to predict the number of fruits of the test set. With that, we ensured that the prediction model was not influenced by the detections of the tested tree. Finally, the prediction error associated to each tree prediction was computed as

$$Error_t = \frac{FruitsPredicted - GT_{field}}{GT_{field}}. \quad (7)$$

In order to have an evaluation of all the dataset, the root mean square error (*RMSE*) was computed:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{N}(Error_t)^2}{N}}, \quad (8)$$

where N is the number of trees evaluated; in this work N=11.

For the geometrical characterization evaluation, the trial $H_{1,n}$ was considered the reference measurement as it was the configuration validated in the original methodologies used to compute the geometrical parameters assessed (Escolà et al., 2017; Sanz et al., 2018). Since training data were not required, this evaluation was carried out on all the dataset at once, comparing the geometrical characterization obtained with standard scanning, $H_{1,n}$, with forced air flow application, $H_{1,(n+af)}$, and with the multi-view approach $H_{(1+2),n}$.

3. Results

This section evaluates the fruit detection algorithm and analyses the effect of air flow and different sensor positioning for fruit detection, yield prediction and geometric characterization. All results are obtained using the data presented in section 2.1.

3.1. Feature assessment

Comparing results using all features, the FS and FPr steps significantly improved the algorithm performance, going from an F_1-score of 0.7449 (after pre-processing) to an F_1-score of 0.7837 and 0.8119 after the FS and FPr steps, respectively (Table 2). Regarding the features, volume was the most useful for FS, presenting an F_1-score improvement of 4% (from 0.7449 to 0.7824), however, it did not contribute in removing FP. The most useful feature for FPr was the reflectance histogram, improving the F_1-score by 2% when it was used. It can be observed that the other tested features also contributed positively to the FS and FPr steps, except for σ_R and R_{max} which slightly penalized the FS although they were useful for FPr.

Table 2. Features assessment using data acquired at sensor height H1 without forced air action ($H_{1,n}$). The evaluation is reported in terms of F_1-score with respect to the number of annotated fruits.

Feature	Pp								
V		✓	✓	✓	✓	✓	✓	✓	✓
rh			✓	✓	✓	✓	✓	✓	✓
Ψ				✓	✓	✓	✓	✓	✓
Pt					✓	✓	✓	✓	✓
λ_n						✓	✓	✓	✓
$\bar{R}$							✓	✓	✓
σ_R								✓	✓
R_{max}									✓
Pp + FS	0.7449	0.7824	0.7834	0.7838	0.7842	0.7843	0.7847	0.7839	0.7837
Pp + FS + FPr	0.7449	0.7824	0.8019	0.8089	0.8090	0.8098	0.8103	0.8110	**0.8119**

3.2. Fruit detection results

For simplicity, in this section results w.r.t. GT_{labels} are presented outside parentheses, together with results w.r.t. GT_{field} inside parentheses. Trials with data from only one tree side (either E or W) presented the lowest detection rates, reporting F_1-score values of 0.537 (0.513) and 0.624 (0.598), at conditions $H_{1,n,E}$ and $H_{1,n,W}$, respectively. The performance significantly improved when the trees were scanned from both sides, presenting an F_1-score of 0.812 (0.784) at conditions $H_{1,n,(E+W)}$. The detection rates dropped when trees were scanned under forced air flow conditions $H_{1,af,(E+W)}$ because the point cloud quality decreased (became "blurred") and, while some apples were dis-occluded by the air flow effect, others were occluded (Figure 5). The forced air flow usefulness was found when combining the data acquired with and without forced air flow $H_{1,(n+af),(E+W)}$. Merging these trials, the percentage of detected fruits increased by more than 7%, achieving a detection rate of 0.894 (0.838), a recall of 0.814 (0.763) and an F_1-score of 0.826 (0.799). Trials at height H_2 presented lower detection rates because the LiDAR was positioned in an upper position and, in consequence, the system failed on detecting fruits at the bottom parts of trees (Figure 6 and Figure 8). Nevertheless, combining trials at different sensor heights $H_{(1+2),n,(E+W)}$ (multi-view approach) increased fruit detection performance similarly to combining different air conditions, achieving an F_1-score of 0.830 (0.802). Finally, combining all trials $H_{(1+2),(n+af),(E+W)}$ (different sensor heights and air flow conditions) increased the detection rate but penalized the recall. This took place due to fact that the point cloud became more blurred, making difficult to split up groups of apples.

For a visual/qualitative assessment, Figure 5 shows an example of the acquired point cloud and the corresponding fruit detection results. Due to the field-of-view of the RGB camera used to acquire images shown in Figure 5a, only the bottom part (from 0m to 2m height) of tree 4 was illustrated. To help the visualization, the colour scale of Figure 5b represents the calibrated reflectance of the measured scene, illustrating reflectance values from 10% (blue) to 80% (red). From that, it can be observed that apples present a higher reflectance than other tree elements.

Red squares of Figure 5 show an example of an apple that was occluded in trial $H_{1,n}$, and became dis-occluded in $H_{1,af}$, due to the effect of applying forced air flow. The opposite happened in the blue squares example, where an apple that was visible in $H_{1,n}$ became occluded in $H_{1,af}$. These occlusions affected the fruit detection performance, as it can be seen in the detections shown in Figure 5c. To take advantage of the dis-occlusions produced by the air flow effect without being penalized by its occlusions, the data from both trials ($H_{1,n}$ and $H_{1,af}$) were combined, increasing the number of visible apples as it can be observed in the detections shown in $H_{1,(n+af)}$. Comparing the detection rates between $H_{1,n,(E+W)}$ and $H_{1,(n+af),(E+W)}$ in Table 3, more than a 7% of fruits were dis-occluded when combining trials with and without force air flow, presenting an increase of the DR from 0.823 in $H_{1,n,(E+W)}$ to 0.894 in $H_{1,(n+af),(E+W)}$.

Table 3. Fruit detection assessment at different sensor heights and air flow conditions. Results are reported in terms of detection rate (DR), recall (R), precision (P), false detection rate (FDR), multi-detection rate (MDR) and F_1-score. DR, R and F1-score were computed with respect to the number of labelled fruits (L), and to the total amount of fruits manually counted in the field (F). Best achieved results are in bold type.

Trial	*DR*		*R*		*P*	*FDR*	*MDR*	F_1-*score*	
	L	*F*	*L*	*F*				*L*	*F*
***$H_{1,n,(E+W)}$**	**0.823**	**0.771**	**0.758**	**0.710**	**0.875**	**0.104**	**0.021**	**0.812**	**0.784**
$H_{1,n,E}$	0.415	0.389	0.383	0.359	0.898	0.089	0.013	0.537	0.513
$H_{1,n,W}$	0.540	0.506	0.485	0.454	0.875	0.110	0.015	0.624	0.598
$H_{1,af,(E+W)}$	0.768	0.720	0.698	0.654	0.868	0.108	0.024	0.774	0.746
$H_{1,(n+af),(E+W)}$	**0.894**	**0.838**	**0.814**	**0.763**	**0.839**	**0.110**	**0.051**	**0.826**	**0.799**
$H_{2,n,(E+W)}$	0.663	0.621	0.588	0.551	0.841	0.131	0.028	0.692	0.666
$H_{2,n,E}$	0.351	0.329	0.318	0.298	0.909	0.085	0.006	0.471	0.449
$H_{2,n,W}$	0.429	0.402	0.369	0.346	0.798	0.171	0.031	0.505	0.482
$H_{2,af,(E+W)}$	0.573	0.537	0.517	0.484	0.885	0.096	0.019	0.652	0.626
$H_{2,(n+af),(E+W)}$	0.748	0.701	0.656	0.615	0.803	0.143	0.054	0.722	0.696
$H_{(1+2),n,(E+W)}$	**0.892**	**0.836**	**0.802**	**0.751**	**0.860**	**0.095**	**0.045**	**0.830**	**0.802**
$H_{(1+2),n,E}$	0.528	0.495	0.480	0.450	0.884	0.096	0.020	0.622	0.596
$H_{(1+2),n,W}$	0.653	0.612	0.576	0.540	0.853	0.112	0.035	0.688	0.661
$H_{(1+2),af,(E+W)}$	0.868	0.813	0.793	0.743	0.866	0.092	0.042	0.828	0.800
$H_{(1+2),(n+af),(E+W)}$	0.917	0.859	0.789	0.739	0.777	0.101	0.122	0.783	0.758

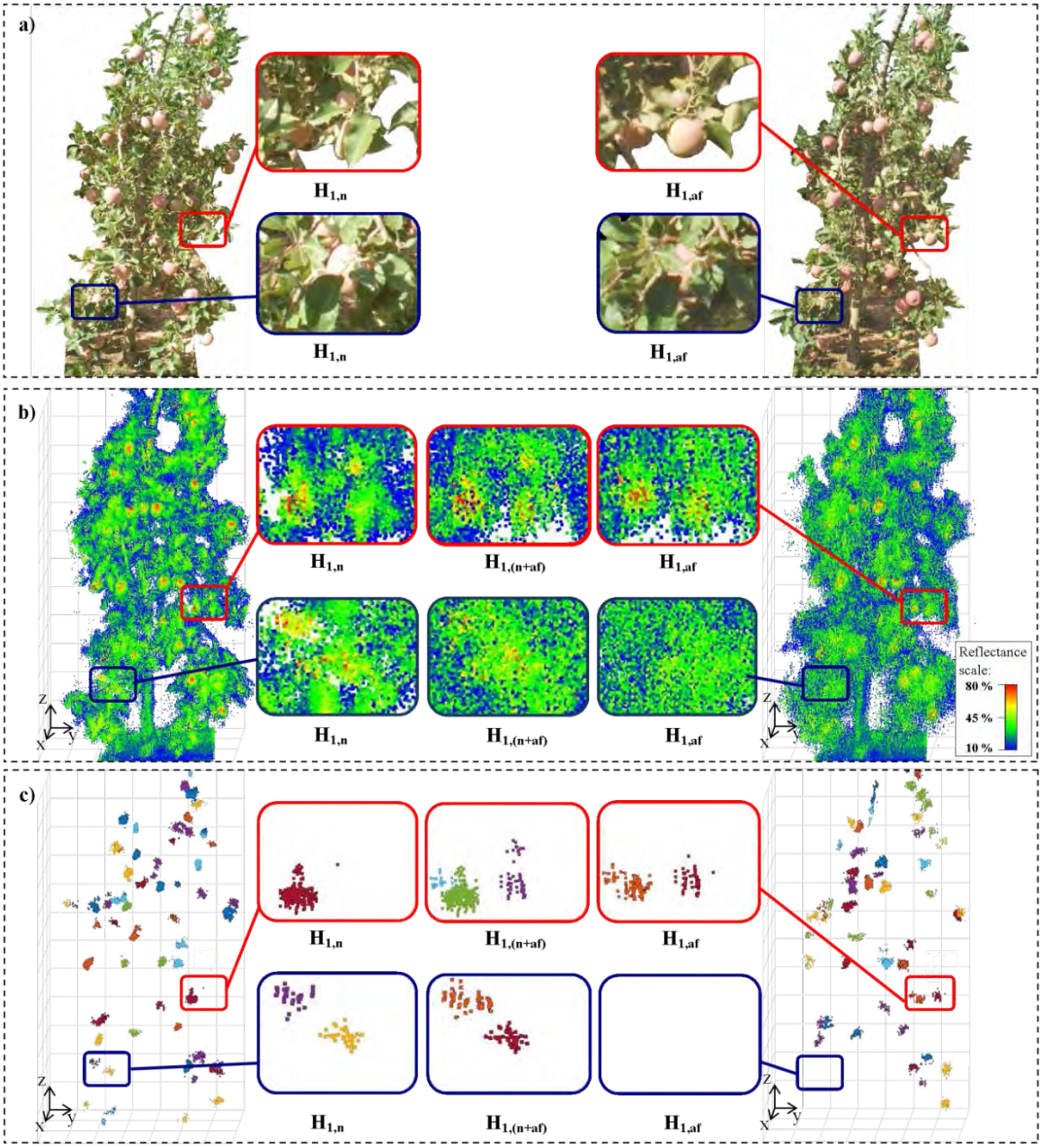


Figure 5. Illustration of the forced air flow effect in fruit detection. Left side corresponds to a trial without forced air ($H_{1,n}$) while right side corresponds to a measurement with forced air flow action ($H_{1,af}$). a) RGB images taken from the sensor position. b) Point cloud obtained with the MTLS. c) Fruit detections using the fruit detection algorithm. Squares in red and blue are a zoom-in of two zones where the air flow effect dis-occluded (red) or occluded (blue) some fruits. Squares denoted by $H_{1,(n+af)}$ correspond to the trial that aggregates the data from $H_{1,n}$ and $H_{1,af}$.

3.3. Fruit location results

As well as not being affected by usual field lighting conditions, the LiDAR sensor used has the advantage of providing the relative 3D location of the fruits detected. When integrated in the designed MTLS, the fruits are located in global 3D coordinates. That allows to obtain the spatial distribution of detected fruits in height (along 'y' axis) and in depth (along 'x' axis). Figure 6 shows that that most of the fruits from the dataset (11 trees) were between heights of 0.6 m and 2.1 m. It is also observed that, when scanning at height H_2, the detection rate dropped in lower zones of the tree, while aggregating data from both scanning heights $H_{(1+2)}$ helps to increase the detection rate (Figure 6). The distribution of fruits along the 'x' axis (Figure 7) shows that the east side ($x > 0$) was 10cm wider and had a 2 % more of fruits than the west side ($x < 0$). Due to the limited number of tested trees (11 trees), further tests should be carried out to study the relation between the canopy width and the fruit yield. The fruit distribution also shows that 1233 fruits out of 1353 (91%) were at distances between -0.4 m and 0.4 m from the center of the tree (along the x axis), with a maximum production of 599 (44%) fruits at distance intervals of ±[0.1, 0.3] (Figure 7). It is also observed that the best detection performance was achieved when combining data acquired with and without forced air action (cyan vertical bars). It should be noted that the mean canopy contour was computed from the average of all canopy widths in the data set and not as the maximum width. For this reason, some detected apples in Figure 7 fall outside the plotted mean canopy contour.

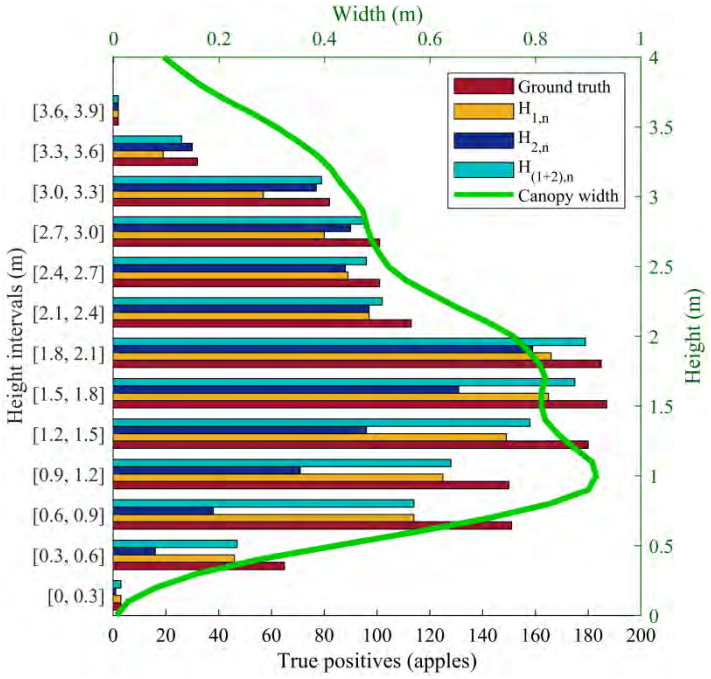


Figure 6. Distribution of fruits in height and comparison with the number of fruits identified when scanning at different sensor height (H_1 and H_2). Data includes information from all dataset (11 trees). Left and bottom axis refers to the horizontal bars, which provide the number of true positives identified by the fruit detection algorithm at different height intervals. Right and top axis (height and width) refers to the mean canopy width illustrated in green.

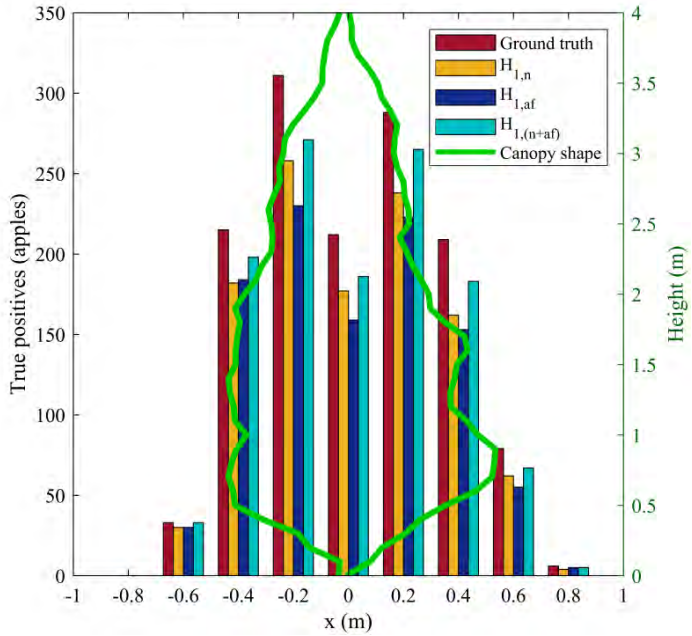


Figure 7. Distribution of fruits in depth (along *x* axis) and comparison with the number of fruits identified in all dataset (11 trees) when scanning at different air flow conditions (*n* and *af*). The mean canopy contour is also illustrated in green. Left and bottom axis refers to the number of true positive identified in each 'x' position, while the right axis refers to the height of the mean canopy contour.

From the distribution histograms presented in Figure 6 and Figure 7, it is straightforward to obtain an evaluation of the fruit detection performance at different tree locations, facilitating an analysis of where the fruit detection system fails or succeeds. Figure 8 illustrates the detection rate (DR) distribution in height and along the *x* axis for the different trials: height sensor positions H_1 and H_2, air flow conditions *n* and *af*, and scanned sides *E* and *W*. Since in H_1 the sensor height was approximately the half of the tree height, the DR of trial *H1* decreased in the lower and upper tree zones (more sharply in the upper zone), approximately at heights under 1m and above 2.5m (Figure 8a). On the other hand, in H_2 the sensor was located in the upper zone, which explains that the DR of trial H_2 decreased in the lower parts of the tree, under 2m height, while reporting DR > 85% for heights above 2.5m (Figure 8a). The trial that combines both sensor heights $H_{(1+2)}$ takes advantage of both views, presenting DR higher than 85% for tree heights above 1m. Regarding the detection performance in depth (along the '*x*' axis), H_2 presented low DR in external zones -further than 0.4m from the center of trees- (Figure 8b). This is because the widest zones were at the bottom of the tree, corresponding to the zone where H_2 had more occlusions. In Figure 8(c-f), the DR was higher in the trials without forced air flow than in the trials with forced air flow. Nevertheless, the DR was improved more than 5% when combining both scanning conditions. This improvement was seen along all tree heights and widths. Finally, Figure 7(g-h) illustrate the DR distribution when processing data from tree row sides separately (*E* or *W*) and together (*E*+*W*). The detection rate dropped more than 28% on the non-scanned side: trial $H_{1,n,E}$ presented low detection rates on the west side ($x < 0$) and trial $H_{1,n,W}$ on the east side ($x > 0$). However, when aggregating the two sides, the detection performance was almost constant in all tree widths because of the reduction of the number of fruit occlusions when scanning from the two sides of the apple trees.

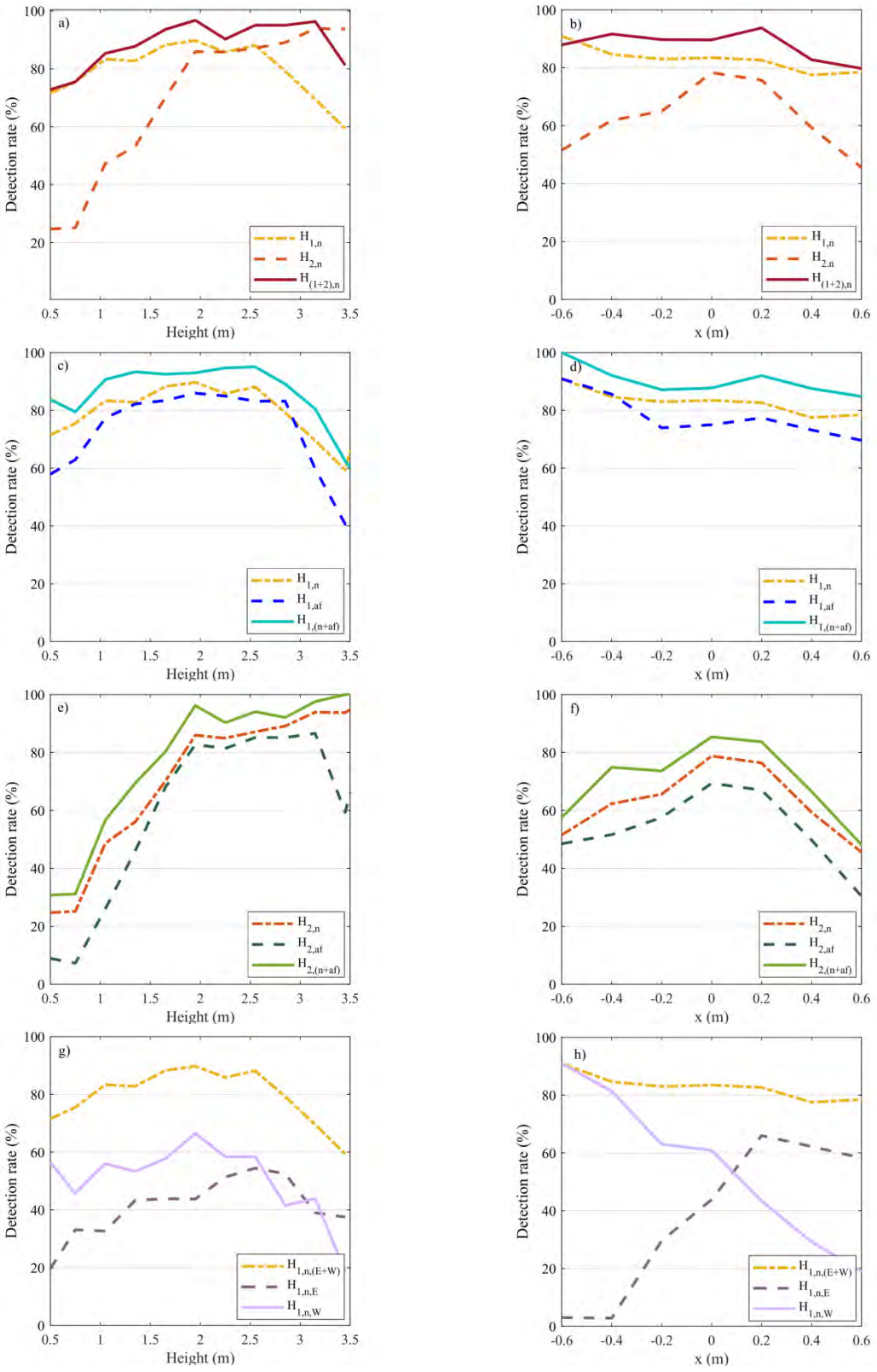


Fig. 2. Detection rate under different air flow conditions (*n* and *af*) and sensor positions (H_1 and H_2) at different tree locations: in height (a, c, e, g) and along the '*x*' axis (b, d, f, h). All plots evaluate data acquired from both row sides (E+W), except g) and h) where the evaluated row side is specified in the corresponding legend.

3.4. Yield prediction results

Table 4 shows yield prediction results for trial $H_{(1+2),n}$, which reported the highest F_1-score, as seen in section 3.2. Results showed higher prediction errors when using data from only one tree side (*E* or *W*), obtaining RMSEs of 19.0% and 12.4% and a coefficient of determination (R^2) of 0.58 and 0.54, when scanning from the east and west sides, respectively. The prediction was significantly improved when using data from both tree sides (*E+W*), presenting errors between -7.7% and 12.0% with an RMSE of 5.7% (Table 4) and R^2=0.87 (Figure 9).

Table 4. Yield predictions results (number of fruits) combining H_1 and H_2 data without forced air flow ($H_{(1+2),n}$). Results are presented for data acquired from one tree side, either east (*E*) or west (*W*), and combining data from both sides (*E+W*).

	Ground Truth	Fruits detected (D)			Fruits predicted (a·D + b)			Prediction error (%)		
	(# fruits)	*E*	*W*	*E+W*	*E*	*W*	*E+W*	*E*	*W*	*E+W*
Tree01	139	71	96	125	133.9	145.7	142.8	-3.7	4.8	2.7
Tree02	106	57	69	98	126.6	119.9	115.5	19.5	13.1	9.0
Tree03	139	64	81	118	128.3	128.2	134.4	-7.7	-7.7	-3.3
Tree04	137	61	89	116	126.0	137.4	132.3	-8.0	0.3	-3.4
Tree05	94	44	59	79	121.0	116.2	92.8	28.8	23.7	-1.3
Tree06	131	66	96	127	130.7	147.5	146.7	-0.2	12.6	12.0
Tree07	119	54	84	103	122.5	133.6	118.9	3.0	12.3	-0.1
Tree08	145	66	77	118	129.3	122.7	133.8	-10.8	-15.4	-7.7
Tree09	139	63	90	126	127.5	138.3	144.0	-8.3	-0.5	3.6
Tree10	136	68	70	114	131.8	114.0	130.1	-3.1	-16.2	-4.4
Tree11	159	120	102	137	237.3	147.0	153.4	49.3	-7.6	-3.5
Total	1444	734	913	1261			RMSE:	**19.0**	**12.4**	**5.7**

When comparing different sensor heights and air flow conditions, the highest performance was achieved by $H_{1,n,(E+W)}$, obtaining an RMSE of 5.4% (Table 5). Although merging different air conditions (*n+af*) or different sensor heights ($H_{(1+2)}$) improved the percentage of fruits detected, neither the air flow effect $H_{1,(n+af),(E+W)}$ nor the multi-view approach $H_{(1+2),n,(E+W)}$ improved yield prediction, presenting similar results to those of trial $H_{1,n,(E+W)}$.

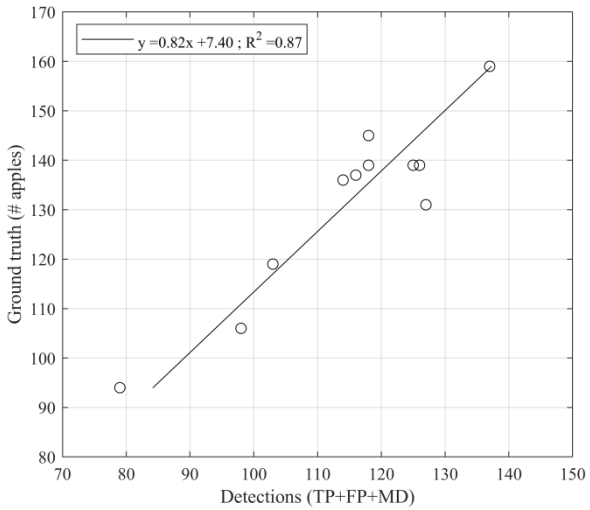


Figure 9. Linear regression between the number of apples detected with $H_{(1+2),n,(E+W)}$ and the actual number of apples per tree (GT_{field}).

Table 5. Yield prediction assessment at different sensor heights (*H1* and *H2*), air conditions (*n* and *af*) and scanned sides (*E* and *W*). Best achieved results are in bold type.

Trial	RMSE (%)
$\boldsymbol{H_{1,n,(E+W)}}$	**5.4**
$H_{1,n,E}$	15.2
$H_{1,n,W}$	15.3
$H_{1,af,(E+W)}$	6.8
$\boldsymbol{H_{1,(n+af),(E+W)}}$	**5.5**
$H_{2,n,(E+W)}$	8.1
$H_{2,n,E}$	11.3
$H_{2,n,W}$	10.6
$H_{2,af,(E+W)}$	12.7
$H_{2,(n+af),(E+W)}$	10.0
$\boldsymbol{*H_{(1+2),n,(E+W)}}$	**5.7**
$*H_{(1+2),n,E}$	19.0
$*H_{(1+2),n,W}$	12.4
$H_{(1+2),af,(E+W)}$	6.7
$H_{(1+2),(n+af),(E+W)}$	8.1

3.5. Geometric characterization results

Regarding height results, it was observed that the forced air flow and multi-view approaches produced very similar results when measuring the mean canopy height. Mean width estimation was neither significantly affected by the multi-view configuration; but a difference higher than 10% was reported when scanning with

forced air flow (Table 6). The mean cross-section area measurement was the parameter which has been affected the most by the scanning conditions, with differences of 22.3% when combining air conditions and 16.3% when combining different sensor heights. These higher deviations are the consequence of an error propagation of height and width measurements. For a qualitative evaluation, Figure 10 illustrates the mean canopy contour obtained in different trials. It can be observed that the multi-view approach (plotted in blue) matched slightly better the reference trial than the trial that combines different air conditions (plotted in red). Finally, concerning the leaf area analysis, both approaches performed similarly, with a 10.6% difference when combining different air conditions and an 8.3% difference with the multi-view approach.

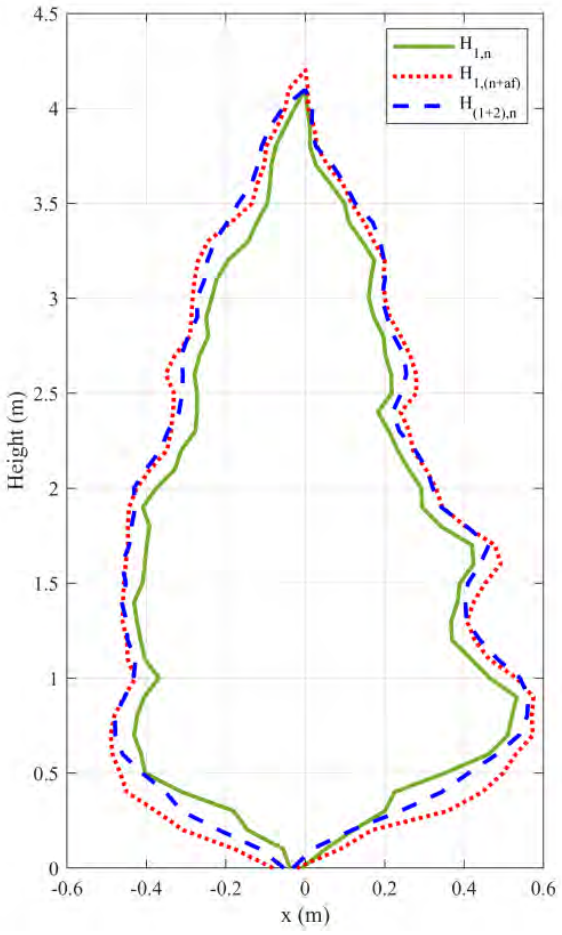


Fig. 3. Illustration of the mean canopy contour obtained at different sensor heights (*H1* and *H2*) and air flow conditions (*n* and *af*). The east side (*E*) corresponds to the positive horizontal distances.

Table 2. Geometric characterization assessment at different sensor heights (*H1* and *H2*) and air conditions (*n* and *af*). Differences with respect to the reference trial *H1_n* are reported within brackets.

Height	Air flow	Mean Height [m]	Mean Width [m]	Mean Cross-Section area [m^2]	Mean Leaf Area [m^2/m]
H1	n	3.64	1.23	2.12	9.77
H1	(n+af)	3.71 (2.1%)	1.36 (10.7%)	2.59 (22.3%)	10.80 (10.6%)
(H1+H2)	n	3.68 (1.3%)	1.27 (3.8%)	2.46 (16.3%)	10.57 (8.3%)

4. Discussion

The fruit detection results showed that trials $H_{1,(n+af)}$, and $H_{(1+2),n}$, located 6.7 % and 6.5 % more fruits than $H_{1,n}$, respectively, presenting $DR > 0.89$ and $R > 0.80$ with respect to the total number of annotated fruits. Although it is difficult to compare systems tested with different datasets, these results are similar (in terms of detection rate) to those obtained in studies based on color images, which have reported accuracies of between 80% and 85% using color features (Gongal et al., 2015) and up to 90% (F_1-score) using deep learning (Bargoti and Underwood, 2017a; Gené-Mola et al., 2019c; Sa et al., 2016). The results are also comparable with those of studies based on vision systems used in orchard harvesting robots, which have reported a mean location success of 80% and a mean identification success of 70% (Bac et al., 2014).

The system used in this study was based on the MTLS described in Escolà et al. (2017). However, there was a big improvement when replacing the 2D LiDAR sensor with the 3D Velodyne VLP-16. Although the original system was not used to detect fruits, having 16 laser beams within a ±15º horizontal FoV, that is 16 different points of view, certainly contributed to improving the fruit detection rates. In future works, the system could be further improved by including an IMU sensor. With that, the system could scan faster -higher forward speed- and along different types of trajectories -not limited to linear-.

Among the main advantages of using the system presented here for fruit detection are that the sensor is not affected by lighting conditions and provides the 3D location of fruits. This allows to know the special fruit distribution in the tree, which can be useful for studying and optimizing agricultural processes (Martin-Gorriz et al., 2014; Widmer

and Krebs, 2001), e.g. comparing different pruning and thinning strategies with the spatial distribution of fruits in the trees and the yield. In Figure 6, it was observed that in general terms the number of fruits increases in wider zones, and in Figure 7 that the number of fruits at the center of the tree was 30% lower than at horizontal distance intervals of ±[0.1, 0.3]. This could be a consequence of canopy porosity (Pfeiffer et al., 2018; Trentacoste et al., 2018), because the center of the tree receives less light than the outer parts, however further analysis is needed. System performance at different heights and widths was analyzed. Figure 8a showed that the multi-view approach $H_{(1+2),n}$ presented similar detection rates at different heights. In contrast, with $H_{1,n}$ system performance decreased when trying to detect fruits from the upper zone of the tree, and with $H_{2,n}$ when detecting fruits from the bottom. As for fruit detection performance along the horizontal axis, $H_{2,n}$ presented low detection rates in the outer parts, whereas $H_{1,n}$ and $H_{(1+2),n}$ presented similar detection rates in all tree widths (Figure 8b).

Knowing the fruit distribution on the tree structure could be valuable for the planning and optimization of harvesting strategies (Bargoti and Underwood, 2017b). For example, depending on the amount of fruits in the top parts of the trees and considering the extra costs involved to pick them (use of ladders or elevation platforms), the farmer could decide not harvest the highest areas. This could also result in improved overall quality of the harvested apples, due to the reduction of fruits damage (in lower and intermediate parts of the tree), caused by metal structures needed to reach the highest points (Młotek et al., 2015).

With respect to the yield prediction results, the presented system was able to estimate the number of fruits on each tree with an RMSE of 5.4% when scanning with standard conditions $H_{1,n}$. Similar results were obtained using the forced air flow and multi-view approaches, with RMSE values of 5.5% and 5.7%, respectively. This means that, although merging air conditions or sensor heights can help to minimize the number of fruit occlusions, these conditions do not provide an advantage for yield prediction, because the correlation between the number of detections and the actual number of fruits in the trees was similar to the one in $H_{1,n}$. The yield prediction errors of ~5.5% reported in trials $H_{1,n}$, $H_{1,(n+af)}$, and $H_{(1+2),n}$, are comparable with other state-of-the-art yield prediction methods, such as those presented by Linker (2018b, 2017), Payne et al.

(2014) and Zhou et al. (2012), which reported yield prediction errors of between 10% and 16%. However, while Linker (2018b, 2017) and Payne et al. (2014) used night-time images to prevent detection errors due to natural lighting, the presented methodology is not affected by lighting conditions.

Another advantage of using a MTLS system compared to other devices or systems used for fruit detection is that it allows simultaneous yield monitoring and geometric characterization. The system was able to measure canopy geometrical parameters at the same time, namely height, width, cross-section area and leaf area. Therefore, spatial maps of different canopy features can be created, observing the relationship between canopy structure and fruit tree productivity. This capability makes the designed system a very interesting tool that can be used to analyze the behavior of different fruit tree varieties in relation to its potential of production (Kühn et al., 2003), the fruit location and, especially, how pruning techniques and training systems could affect production depending on the structural organization (Martin-Gorriz et al., 2014).

5. Conclusions

This work presents an analysis of different methodologies based on the use of a mobile terrestrial laser scanner for remote fruit detection and plant geometrical characterization. In order to minimize fruit occlusions, two different approaches were tested: forced air flow and multi-view sensing. The main contributions of this paper were: (1) A methodology for simultaneous fruit location and canopy geometric characterization; (2) An analysis of the usefulness of forced air flow and multi-view approaches for fruit detection, yield prediction and canopy geometric characterization. The results show that the system was able to detect and locate more than 80% of the total annotated fruits and to predict the yield with an RMSE lower than 6%. These fruit detection results are comparable with those obtained with other state-of-the-art methodologies, with the advantages that the presented system is not affected by lighting conditions and also provides geometric characterization of the tree crop, allowing the comparison between yield and canopy structure. From the scanning conditions analysis, it is concluded that the best configuration for yield prediction and geometric characterization corresponds to mounting the sensor at half the maximum tree height H_1 and scanning without forced air

flow action. However, if the LiDAR-based system is used for fruit detection, combining data acquired with and without forced air flow action $H_{1,(n+af)}$ or using the multi-view $H_{(1+2),n}$ approach are good options to increase the percentage of fruits detected. If the scanning system is used for both fruit detection and geometric characterization, the best option is the multi-view approach, since it increases the fruit detection rate without excessively penalizing geometric characterization. Future works should focus in the analysis of fruit occlusions in different training systems and extending the present study to other fruit varieties.

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). The Spanish Ministry of Education is thanked for Mr. J. Gené's pre-doctoral fellowships (FPU15/03355). The work of Jordi Llorens was supported by the Spanish Ministry of Economy, Industry and Competitiveness through a postdoctoral position named Juan de la Cierva Incorporación (JDCI-2016-29464_N18003). We would also like to thank CONICYT FONDECYT 1171431 and CONICYT FB0008. Nufri (especially Santiago Salamero and Oriol Morreres) and Vicens Maquinària Agrícola S.A. are also thanked for their support during data acquisition, and Ernesto Membrillo and Roberto Maturino for their support in dataset labelling.

References

Andújar, D., Dorado, J., Bengochea-Guevara, J.M., Conesa-Muñoz, J., Fernández-Quintanilla, C., Ribeiro, Á., 2017. Influence of Wind Speed on RGB-D Images in Tree Plantations. *Sensors (Basel)*. 17, 1–12. doi:10.3390/s17040914

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, 888–911. doi:10.1002/rob.21525

Bargoti, S., Underwood, J., 2017a. Deep Fruit Detection in Orchards. *2017 IEEE Int. Conf. Robot. Autom.* 3626–3633. doi:10.1109/ICRA.2017.7989417

Bargoti, S., Underwood, J.P., 2017b. Image Segmentation for Fruit Detection and Yield Estimation in Apple Orchards. *J. F. Robot.* 34, 1039–1060. doi:10.1002/rob.21699

Bechar, A., Vigneault, C., 2016. Agricultural robots for field operations: Concepts and components. *Biosyst. Eng.* 149, 94–111. doi:10.1016/j.biosystemseng.2016.06.014

Bulanon, D.M., Burks, T.F., Alchanatis, V., 2009. Fruit Visibility Analysis for Robotic Citrus

Harvesting. *Trans. Am. Soc. Agric. Biol. Eng.* 52, 277–283. doi:10.13031/2013.25933

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

Bulanon, D.M., Kataoka, T., 2010. A Fruit Detection System and an End Effector for Robotic Harvesting of Fuji Apples. *Agric. Eng. Int. CIGR J.* 12, 203-210.

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* 2, 121–167. doi:10.1023/A:1009715923555

Ceres, R., Pons, J.L., Jimenez, A.R., Martín, J.M., Calderón, L., 1998. Design and implementation of an aided fruit-harvesting robot (Agribot). *Ind. Robot An Int. J.* 25, 337–346. doi:https:// doi.org/10.1108/01439919810232440

Colaço, A.F., Molin, J.P., Rosell-Polo, J.R., Escolà, A., 2018a. Application of light detection and ranging and ultrasonic sensors to high-throughput phenotyping and precision horticulture: Current status and challenges. *Hortic. Res.* 5, 1-11. doi:10.1038/s41438-018-0043-0

Colaço, A.F., Molin, J.P., Rosell-Polo, J.R., Escolà, A., 2018b. Spatial variability in commercial orange groves. Part 2: relating canopy geometry to soil attributes and historical yield. *Precis. Agric.* 20, 805–822. doi:10.1007/s11119-018-9615-0

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. doi:10.1007/BF00994018

Edan, Y., Rogozin, D., Flash, T., Miles, G.E., 2000. Robotic melon harvesting. *IEEE Trans. Robot. Autom.* 16, 831–835. doi:10.1109/70.897793

Escolà, A., Martinez-Casasnovas, J.A., Rufat, J., Arno, J., Arbones, A., Sebe, F., Pascual, M., Gregorio, E., Rosell-Polo, J.R., 2017. Mobile terrestrial laser scanner applications in precision fruticulture/horticulture and tools to extract information from canopy point clouds. *Precis. Agric.* 18, 111–132. doi:10.1007/s11119-016-9474-5

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

Feng, J., Liu, G., Wang, S., Zeng, L., Ren, W., 2012. A Novel 3D Laser Vision System for Robotic Apple Harvesting. *ASABE Annu. Int. Meationg.*

Gan, H., Lee, W.S., Alchanatis, V., Ehsani, R., Schueller, J.K., 2018. Immature green citrus fruit detection using color and thermal images. *Comput. Electron. Agric.* 152, 117–125. doi:10.1016/j.compag.2018.07.011

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.* 25, 104289. 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., Karkee, M., Amatya, S., 2018. Apple fruit size estimation using a 3D machine vision system. *Inf. Process. Agric.* 5, 498–503. doi:10.1016/j.inpa.2018.06.002

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

Gotou, K., Fujiura, T., Nishiura, Y., Ikeda, H., Dohi, M., 2003. 3-D vision system of tomato production robot. *IEEE/ASME Int. Conf. Adv. Intell. Mechatronics, AIM* 2, 1210–1215. doi:10.1109/AIM.2003.1225515

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. *Pattern Recognit.* Lett. 31 (8), 651–666. doi:10.1016/j.patrec.2009.09.011

Jiménez-Brenes, F.M., López-Granados, F., Castro, A.I., Torres-Sánchez, J., Serrano, N., Peña, J.M., 2017. Quantifying pruning impacts on olive tree architecture and annual canopy growth by using UAV-based 3D modelling. *Plant Methods* 13, 1–15. doi:10.1186/s13007-017-0205-3

Jiménez, A.R., Ceres, R., Pons, J.L., 2000. A vision system based on a laser range-finder applied to robotic fruit harvesting. *Mach. Vis. Appl.* 11, 321–329. doi:10.1007/s001380050117

Jolliffe, I., 2011. Principal Component Analysis, in: *Lovric, M. (Ed.), International Encyclopedia of Statistical Science. Springer Berlin Heidelberg, Berlin, Heidelberg*, pp. 1094–1096. doi:10.1007/978-3-642-04898-2_455

Kamilaris, A., Prenafeta-Boldú, F.X., 2018. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* 147, 70–90. doi:10.1016/j.compag.2018.02.016

Kühn, B.F., Pedersen, H.L., Andersen, T.T., 2003. Evaluation of 14 old unsprayed apple varieties. *Biol. Agric. Hortic.* 20, 301–310. doi:10.1080/01448765.2003.9754975

Lee, B.S., Rosa, U.A., 2006. Development of a canopy volume reduction technique for easy assessment and harvesting of Valencia citrus fruits. *Trans. ASABE* 49, 1695–1703. doi:10.13031/2013.22286

Linker, R., 2018. Machine learning based analysis of night-time images for yield prediction in apple orchard. *Biosyst. Eng.* 167, 114–125. doi:10.1016/j.biosystemseng.2018.01.003

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

Linker, R., Kelman, E., 2015. Apple detection in nighttime tree images using the geometry of light patches around highlights. *Comput. Electron. Agric.* 114, 154–162.

doi:10.1016/j.compag.2015.04.005

Liu, X., Zhao, D., Jia, W., Ruan, C., Tang, S., Shen, T., 2016. A method of segmenting apples at night based on color and position information. *Comput. Electron. Agric*. 122, 118–123. doi:10.1016/j.compag.2016.01.023

Maldonado, W., Barbosa, J.C., 2016. Automatic green fruit counting in orange trees using digital images. *Comput. Electron. Agric.* 127, 572–581. doi:10.1016/j.compag.2016.07.023

Martin-Gorriz, B., Castillo, I.P., Torregrosa, A., 2014. Effect of mechanical pruning on the yield and quality of 'Fortune' mandarins. *Spanish J. Agric. Res.* 12, 952–959. doi:http://dx.doi.org/10.5424/sjar/2014124-5795

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

Młotek, M., Kuta, Ł., Stopa, R., Komarnicki, P., 2015. The Effect of Manual Harvesting of Fruit on the Health of Workers and the Quality of the Obtained Produce. *Procedia Manuf.* 3, 1712–1719. doi:10.1016/j.promfg.2015.07.494

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

Okamoto, H., Lee, W.S., 2009. Green citrus detection using hyperspectral imaging. *Comput. Electron. Agric.* 66, 201–208. doi:10.1016/j.compag.2009.02.004

Payne, A., Walsh, K., Subedi, P., Jarvis, D., 2014. Estimating mango crop yield using image analysis using fruit at "stone hardening" stage and night time imaging. *Comput. Electron. Agric.* 100, 160–167. doi:10.1016/j.compag.2013.11.011

Payne, A.B., Walsh, K.B., Subedi, P.P., Jarvis, D., 2013. Estimation of mango crop yield using image analysis - Segmentation method. *Comput. Electron. Agric.* 91, 57–64. doi:10.1016/j.compag.2012.11.009

Pfeiffer, S.A., Guevara, J., Cheein, F.A., Sanz, R., 2018. Mechatronic terrestrial LiDAR for canopy porosity and crown surface estimation. *Comput. Electron. Agric.* 146, 104–113. doi:10.1016/j.compag.2018.01.022

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. *Adv. Agron.* 133, 71–112. doi:10.1016/bs.agron.2015.05.002

Rosell, J.R., Sanz, R., 2012. A review of methods and applications of the geometric characterization of tree crops in agricultural activities. *Comput. Electron. Agric.* 81, 124–141. doi:10.1016/j.compag.2011.09.007

Rusu, R.B., Marton, Z.C., Blodow, N., Dolha, M., Beetz, M., 2008. Towards 3D Point cloud based object maps for household environments. *Rob. Auton. Syst.* 56, 927–941. doi:10.1016/j.robot.2008.08.005

Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T., McCool, C., 2016. DeepFruits: A Fruit Detection System Using Deep Neural Networks. *Sensors* 16, 1222. doi:10.3390/s16081222

Safren, O., Alchanatis, V., Ostrovsky, V., Levi, O., 2007. Detection of Green Apples in Hyperspectral Images of Apple-Tree Foliage Using machine Vision. *Trans. ASABE* 50, 2303–2313. doi:10.13031/2013.24083

Sanz, R., Llorens, J., Escolà, A., Arnó, J., Planas, S., Román, C., Rosell-Polo, J.R., 2018. LIDAR and non-LIDAR-based canopy parameters to estimate the leaf area in fruit trees and vineyard. *Agric. For. Meteorol.* 260–261, 229–239. doi:10.1016/j.agrformet.2018.06.017

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

Tagarakis, A.C., Koundouras, S., Fountas, S., Gemtos, T., 2018. Evaluation of the use of LIDAR laser scanner to map pruning wood in vineyards and its potential for management zones delineation. *Precis. Agric.* 19, 334–347. doi:10.1007/s11119-017-9519-4

Tanigaki, K., Fujiura, T., Akase, A., Imagawa, J., 2008. Cherry-harvesting robot. *Comput. Electron. Agric.* 63, 65–72. doi:10.1016/j.compag.2008.01.018

Tilman, D., Balzer, C., Hill, J., Befort, B.L., 2011. Global food demand and the sustainable intensification of agriculture. *Proc. Natl. Acad. Sci.* 108, 20260–20264. doi:10.1073/pnas.1116437108

Trentacoste, E.R., Calderón, F.J., Puertas, C.M., Banco, A.P., Contreras-Zanessi, O., Galarza, W., Connor, D.J., 2018. Vegetative structure and distribution of oil yield components and fruit characteristics within olive hedgerows (cv. Arbosana) mechanically pruned annually on alternating sides in San Juan, Argentina. *Sci. Hortic. (Amsterdam).* 240, 425–429. doi:10.1016/j.scienta.2018.06.045

Uribeetxebarria, A., Martínez-Casasnovas, J.A., Escolà, A., Rosell-Polo, J.R., Arnó, J., 2019. Stratified sampling in fruit orchards using cluster-based ancillary information maps: a comparative analysis to improve yield and quality estimates. *Precis. Agric.* 20, 179–192. doi:10.1007/s11119-018-9619-9

Vázquez-Arellano, M., Griepentrog, H.W., Reiser, D., Paraforos, D.S., 2016. 3-D Imaging Systems for Agricultural Applications — A Review. *Sensors (Basel)*. 16(5), 618. doi:10.3390/s16050618

Velodyne, L., 2016. VLP-16 In VLP-16 Manual: User's Manual and Programming Guide; Velodyne LiDAR.

Widmer, A., Krebs, C., 2001. Influence of planting density and tree form on yield and fruit quality of "golden delicious" and "royal gala" apples, in: *VII International Symposium on Orchard and Plantation Systems 557*. pp. 235–242. doi:10.17660/ActaHortic.2001.557.30

Yandún Narváez, F.J., Salvo del Pedregal, J., Prieto, P.A., Torres-Torriti, M., Auat Cheein, F.A., 2016. LiDAR and thermal images fusion for ground-based 3D characterisation of fruit trees. *Biosyst. Eng.* 151, 479–494. doi:10.1016/j.biosystemseng.2016.10.012

Yin, H., Chai, Y., Yang, S.X., Mittal, G.S., 2009. Ripe tomato recognition and localization for a tomato harvesting robotic system, in: *SoCPaR 2009 - Soft Computing and Pattern Recognition. IEEE*, pp. 557–562. doi:10.1109/SoCPaR.2009.111

Zhang, B., Huang, W., Wang, C., Gong, L., Zhao, C., Liu, C., Huang, D., 2015. Computer

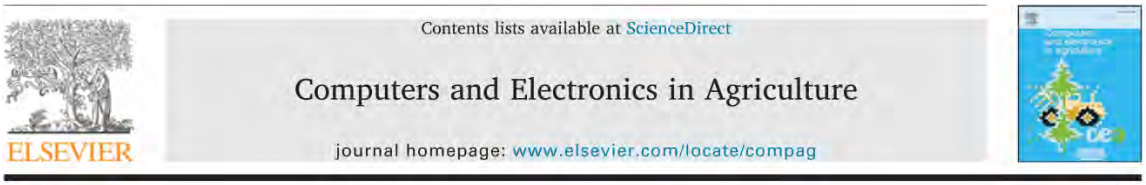
vision recognition of stem and calyx in apples using near-infrared linear-array structured light and 3D reconstruction. *Biosyst. Eng.* 139, 25–34. doi:10.1016/j.biosystemseng.2015.07.011

Zhao, Y., Gong, L., Huang, Y., Liu, C., 2016. A review of key techniques of vision-based control for harvesting robot. *Comput. Electron. Agric.* 127, 311–323. doi:10.1016/j.compag.2016.06.022

Zhou, R., Damerow, L., Sun, Y., Blanke, M.M., 2012. Using colour features of cv. “Gala” apple fruits in an orchard in image processing to predict yield. *Precis. Agric.* 13, 568–580. doi:10.1007/s11119-012-9269-2

Chapter VI. P6: Multi-modal deep learning for Fuji apple detection using RGB-D cameras and their radiometric capabilities

This chapter was published in *Computers and Electronics in Agriculture 162* (2019) 689-698, https://doi.org/10.1016/j.compag.2019.05.016:



Original papers

Multi-modal deep learning for Fuji apple detection using RGB-D cameras and their radiometric capabilities

Jordi Gené-Mola[a], Verónica Vilaplana[b], Joan R. Rosell-Polo[a], Josep-Ramon Morros[b], Javier Ruiz-Hidalgo[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

Fruit detection and localization will be essential for future agronomic management of fruit crops, with applications in yield prediction, yield mapping and automated harvesting. RGB-D cameras are promising sensors for fruit detection given that they provide geometrical information with color data. Some of these sensors work on the principle of time-of-flight (ToF) and, besides color and depth, provide the backscatter signal intensity. However, this radiometric capability has not been exploited for fruit detection applications. This work presents the KFuji RGB-DS database, composed of 967 multi-modal images containing a total of 12,839 Fuji apples. Compilation of the database allowed a study of the usefulness of fusing RGB-D and radiometric information obtained with Kinect v2 for fruit detection. To do so, the signal intensity was range corrected to overcome signal attenuation, obtaining an image that was proportional to the reflectance of the scene. A registration between RGB, depth and intensity images was then carried out. The Faster R-CNN model was adapted for use with five-channel input images: color (RGB), depth (D) and range-corrected intensity

signal (S). Results show an improvement of 4.46% in F1-score when adding depth and range-corrected intensity channels, obtaining an F1-score of 0.898 and an AP of 94.8% when all channels are used. From our experimental results, it can be concluded that the radiometric capabilities of ToF sensors give valuable information for fruit detection.

Keywords: RGB-D; Multi-modal faster R-CNN; Convolutional Neural Networks; Fruit detection; Agricultural robotics; Fruit reflectance.

1. Introduction

To meet the food needs of a world's growing population, horticulture must find new ways to increase the production of fruits and vegetables (Siegel et al., 2014). This is a major challenge for agricultural communities, especially in a context of rising farming costs and a shortage of skilled labor. Efficient and sustainable agronomic management is required to reduce economic and environmental costs while increasing orchard productivity.

Improvements in technological fields like robotics and computer science have provided farmers with tools to increase production in an efficient and sustainable way (Underwood et al., 2016). The use of new technologies in precision agriculture has been applied in the optimization of agricultural processes such as water irrigation, agrochemical application, fertilization, pruning and thinning (Auat Cheein and Carelli, 2013; Bargoti and Underwood, 2017b). Farmers can obtain valuable information for optimization of these processes from the detection and quantification of fruit distribution within the canopy.

Advances in sensing and computer vision have facilitated the development of remote fruit detection systems, with applications in yield prediction, yield mapping and automated harvesting. Yield prediction allows farmers to plan the harvest campaign, fruit storage and sales (Bargoti and Underwood, 2017b; Nuske et al., 2014). On many occasions, yield estimation is carried out by manual counting of a few samples, without addressing spatial variability within the orchard. Although simple random sampling (SRS) is a widely used technique for yield estimation, it is necessary to sample a relatively large number of trees for a precise estimation. Though this may sometimes be unfeasible with manual counting, it could be possible by using currently available

computer vision technologies. As for yield mapping, production maps provide useful information for fruit growers. Fruit orchards usually show spatial variability due to soil variations, fertility, water irrigation, among others (Uribeetxebarria et al., 2018). An analysis of yield maps helps farmers to find the reasons for such variability and to determine which areas of lower productivity require special attention. Finally, fruit detection and 3D localization are the first steps in the development of automated harvesting. Hand harvesting is a hard and human-resource intensive labor, which has to find an alternative since the decreasing availability of skilled labor force (Gongal et al., 2015). Despite the latest advances in imaging techniques and computer vision, detecting and localizing fruits within the canopy is still a pending issue that has to face problems derived from the heterogeneity of the environment, such as occlusions with other vegetative organs and variable lighting conditions. Most of the emerging sensors, such depth cameras (RGB-D sensors), have not yet been exploited for fruit detection and localization. The major reason is the lack of substantial datasets (Hameed et al., 2018).

This paper introduces the KFuji RGB-DS database, which contains multi-modal images of Fuji apples in real orchards, and presents a novel study of the usefulness of RGB-D sensors and their radiometric capabilities for fruit detection. The Faster Region-based Convolutional Neural Network (Faster R-CNN) was adapted and implemented for apple detection using multi-modal images obtained with Microsoft's Kinect v2 (Microsoft, Redmond, WA, USA). The multi-modal images were obtained after pre-processing and registering three different modalities: color (RGB), depth (D) and range-corrected IR intensity -proportional to reflectance- (S).

The main contributions of this paper are: (1) provision of the first apple dataset with multi-modal images from RGB-D sensors with color, depth and range-corrected IR intensity data, and the corresponding annotations with the ground truth apple locations; (2) an analysis of the radiometric capabilities of Kinect v2 for fruit detection; (3) an implementation of a high-performance fruit detection system using an adaptation of Faster R-CNN for five-channel input images; (4) a study of the optimal anchor scales and aspect ratios used in the region proposal network (RPN). After this Introduction section, the rest of the paper is structured as follows: section 2 presents related work retrieved from the state of the art; section 3 describes the proposed dataset, explaining

the signal range-correction theoretical basis, the experimental set up for data acquisition, the pre-processing needed to build the 5-channel multi-modal images, and the network implemented for fruit detection; section 4 shows the results and discusses qualitatively and quantitatively the performance of the fruit detector when using each of the modalities provided by the sensor; finally, the conclusions are presented in section 5

2. Related work

Over the years, different sensors and systems have been used for fruit detection and localization (Gongal et al., 2015). The most commonly sensors used are color (or RGB) cameras (Bargoti and Underwood, 2017a; Linker, 2017; Maldonado and Barbosa, 2016; Zhao et al., 2016). However, the drawbacks to these sensors include the fact that they only provide 2D information and their measurements are affected by lighting conditions. Advances in photonics and the exploration of non-visible wavelengths have allowed the introduction of other systems, including thermal, multispectral and hyperspectral cameras. Thermal cameras have been used in fruit detection, differentiating fruits from background by the different thermal inertia of the fruits. Fruits can thus be detected when the ambient temperature is increasing or decreasing (Bulanon et al., 2008; Stajnko et al., 2004). Multispectral and hyperspectral cameras have also been used for fruit detection, allowing the acquisition of data at different bands of the electromagnetic spectrum (Okamoto and Lee, 2009; Sa et al., 2016; Safren et al., 2007; Zhang et al., 2015). However, like RGB cameras, thermal, multispectral and hyperspectral cameras only provide 2D information.

More recently, LiDAR (Light Detection and Ranging) systems have been introduced in agriculture to obtain 3D models of crops (Escolà et al., 2017; Rosell Polo et al., 2009; Sanz et al., 2018). This sensor works according to the time-of-flight principle (ToF), measuring distances to the objects by computing the time required by a laser pulse to complete the round trip between sensor and target. Besides the geometrical information (3D point clouds), this sensor also provides the amount of light backscattered by the scene (related with the reflectance). In this respect, the authors have shown in a recent study (Gené-Mola et al., 2018) that some fruits, like Fuji apples, have higher reflectance

than leaves and trunks, reporting an 85% detection success rate when using the reflectance capabilities of a LiDAR sensor.

Another technology derived from previous ones, and also used in crop monitoring are the RGB-D or depth cameras (Rosell-Polo et al., 2017, 2015). These sensors provide 3D information with color data, allowing the detection and subsequent 3D localization of the fruit. The operating principle can be based on stereo triangulation (Font et al., 2014; Wang et al., 2017) or on a combination of an RGB and a depth sensor, either based on structured light (Nguyen et al., 2016) or on ToF (Barnea et al., 2016; Gongal et al., 2018). Similarly to LiDAR sensors, RGB-D systems based on the ToF principle provide the amount of light backscattered by the scene, which can be related to the reflectance after range correction and sensor calibration (Rodríguez-Gonzálvez et al., 2016). This radiometric capability has been applied to face detection (Chhokra et al., 2018) by using the backscattered IR intensity image (without range correction) as an additional channel (RGB-DI). However, to the best of the authors' knowledge, no previous object detection work has used range-corrected intensity data (proportional to reflectance). The use of this additional information would be of interest in fruit detection, since the reflectance of some fruit varieties is higher than background reflectance (Gené-Mola et al., 2018).

Regarding the processing techniques used for fruit detection, most previous works have used traditional hand-crafted features to encode the data acquired with different sensors and infer fruit location. More recently, the introduction of deep neural networks has led to remarkable progress in object recognition and, therefore, in fruit detection. The object detection network Faster R-CNN (Ren et al., 2017) is the most commonly used for fruit detection (Bargoti and Underwood, 2017a; Gan et al., 2018; Sa et al., 2016; Stein et al., 2016). Although other state-of-the-art networks are computationally more efficient (Liu et al., 2016; Redmon and Farhadi, 2017), real-time inference is not normally a requirement in fruit detection, so Faster R-CNN is often chosen due to its better performance with small objects. The main drawback of using convolutional neural networks is that they require a large amount of labelled data. As pointed out in previous studies (Hameed et al., 2018), the lack of substantial datasets is a barrier for exploring emerging sensors that could be useful for fruit detection.

3. Materials and Methods

3.1 Theoretical basis

As previously introduced, the operation of the Kinect v2 sensor is based on the ToF principle. Thus, the received power coming from an object located at a distance R is given by the elastic LiDAR equation (Höfle and Pfeifer, 2007; Rodríguez-Gonzálvez et al., 2016):

$$P_r = \frac{P_t A \rho}{\pi R^2} \eta_{sys} \eta_{atm} \cos\theta , \quad (1)$$

where P_t is the emitted power, A is the receiving area, ρ is the object reflectance, η_{sys} is the optical efficiency of the instrument, η_{atm} accounts for atmospheric absorption and scattering, and θ is the incidence angle. The η_{atm} is assumed to be equal to unity due to the short working range of the Kinect sensor. The received power P_r is range corrected in order to compare returns coming from different distances:

$$P_r R^2 = K \rho \cos\theta , \quad (2)$$

where K is the system constant that groups the instrument parameters. Rodríguez-Gonzálvez et al. (2016) showed that there is a linear relationship between the digitized intensity E provided by the Kinect v2 and the received power:

$$E = a \cdot P_r + b \quad (3)$$

where a is the gain and b is the offset. From Eq. (2) and Eq. (3) it is found that the range-corrected signal S depends on the reflectance ρ as follows,

$$S = E R^2 \propto \rho \cos\theta \quad (4)$$

where $R^2 = x^2 + y^2 + z^2$, with $[x, y, z]$ the Cartesian coordinates of each point in the 3D cloud with respect to the sensor.

3.2 KFuji RGB-DS dataset

3.2.1 Data acquisition

Data were acquired in a commercial Fuji apple orchard (*Malus domestica* Borkh. cv. Fuji) located in Agramunt, Catalonia, Spain. The images were taken on September 25-28th of 2017, three weeks before harvesting, at BBCH (Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie) phenological growth stage 85 (Meier, 2001).

The data acquisition equipment consisted of two RGB-D cameras mounted on a mobile platform at heights of 1 m and 3 m, respectively (Figure 1) in order to capture data from all the tree height. The RGB-D sensors used were two Microsoft Kinect v2, which incorporate an RGB camera and a depth sensor that works according to the ToF principle. This sensor provides 3 different types of data: a color image, a depth image that can be used to generate a 3D point cloud of the scene, and the received IR backscattered intensity. Specific software written in C# was developed to collect and save data automatically. The software generates a 3D point cloud for each capture, with RGB and backscattered intensity data for each point, and saves it jointly with the raw RGB image. All captures were carried out during the night, using artificial lighting, since performance of the depth sensor drops under direct sunlight exposure (Rosell-Polo et al., 2015). Table 1 summarizes the specifications of the sensor and the platform used for data acquisition.

Figure 1. View of the acquisition equipment showing the Kinect v2 sensors mounted on the mobile platform.

Table 1. Measurement equipment specifications.

RGB-D sensor	Manufacturer and model	Microsoft Kinect v2
	RGB channel resolution (pixels)	1920 x 1080
	RGB channel field-of-view (FOV)	84.1º x 53.8º
	IR and Depth channel resolution (pixels)	512 x 424
	IR and Depth channel FOV	70º x 60º
	Working range (m)	0.5 - 8
Mobile Platform	Developer	GRAP-UdL-AT research group
	Forward speed (km/h)	0.5 (manually adjustable)
	Sensors height (m)	1 - 3

3.2.2 Data preparation

Once the data were collected, for each capture it was obtained a 3D point cloud (with RGB and backscattered intensity information) and the corresponding raw RGB image. Captures were processed separately. Depending on the application where this methodology is used, apples appearing in the overlapped parts of images should be addressed. For instance, for yield estimation, images should be registered in order to count the apples appearing in the overlapped parts only one time. However, this is not the goal of this work.

A pre-processing was carried out to prepare these data as input data of the convolutional neural network. Data preparation included range-correction of the backscattered signal, 2D projection of the 3D point cloud, and image registration between range-corrected intensity and raw RGB images. Figure 2 illustrates a flowchart of the data preparation steps.

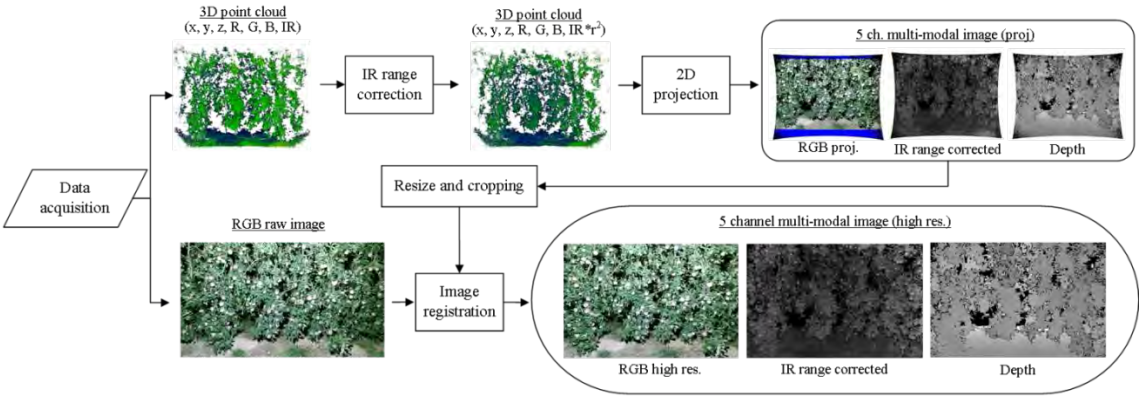


Figure 2. Data preparation diagram. For each frame, the sensor provides a 3D point cloud with RGB and backscattered intensity data, and a raw RGB image. Firstly, the intensity signal is range-corrected. Then, the 3D point cloud is projected onto a 2D plane parallel to the sensor, generating the range-corrected and depth images. Finally, the projected images are resized and cropped in order to register them with the RGB raw image. 5-channel multi-modal images of the scene were obtained, with RGB channels in high resolution.

The range correction of the backscattered signal was performed as described in Section 3.1. After obtaining the 3D point cloud with range-corrected intensity data, a perspective projection onto a plane parallel to the sensor was carried out, generating the corresponding RGB projected, range-corrected intensity and depth images. Since the vertical field-of-view (FOV) of the depth sensor is larger than the vertical FOV of the RGB camera, the top and bottom parts, where no RGB information is given, were cut (blue regions in the RGB proj. image of Figure 2). This step was the responsible of having a different image aspect ratio that the original 512/424. In order to work with an RGB image with higher resolution than that of the IR image, a registration between the RGB raw image and the projected images is required. To do so, projected images were resized (bicubic interpolation) to 1600 x 1080 pixels (px) to achieve the same vertical size as the RGB high resolution image. Finally, an image registration was performed in order to have correspondence between all images, obtaining a 5-channel multi-modal image (with RGB in high resolution) where each pixel has information from 3 modalities: color (RGB), range-corrected intensity and depth (Figure 2). Hereinafter, the RGB image obtained after the point cloud projection is denoted as RGB_p, while the RGB image obtained after registering the raw RGB image is denoted as RGB_{hr}. In order to have similar mean and variance between channels, the range-corrected intensity and depth images were normalized between 0 and 255 -like RGB images-. This normalization is desirable to ensure fast convergence of the network. The RGB channels were saved in 8-bit images while the range-corrected intensity and depth images were stored in 64-bits to avoid data precision loss.

Ground truth fruit locations were manually annotated using the Pychet Labeller toolbox (Bargoti, 2016), labeling a total of 12,839 apples in all the dataset. Due to the large number of fruits per image (more than 100 fruits/image), and taking into account that fruit size (44 ± 6 px in diameter) is relatively small with respect to image size (1600 x 1080 px), each capture was divided into 9 sub-images of 548 x 373 px, with an overlap of 20 px between sub-images to avoid the partially split of fruits at the boundaries in different partitions (Figure 3).

Figure 3. Image sub-division. Each raw image was divided into 9 sub-images to achieve a better relation between apple and image size.

In total, the data set is composed of 967 sub-images, split into training, validation, and test sets as shown in Table 2. Some examples of the multi-modal sub-images used in the training dataset are shown in Figure 4. Due to further quantization for representation, fruits cannot be seen in depth images. The KFuji RGB-DS dataset with corresponding annotations has been made publicly available at *www.grap.udl.cat/en/publications/datasets.html.*

Table 3. Dataset configuration.

Raw image	Sub-image	Fruit size	Training	Validation	Test	No. of fruits (all dataset)
1600x1080 px	548x373 px	44 ± 6 px	619 (64%)	155 (16%)	193 (20%)	12.839

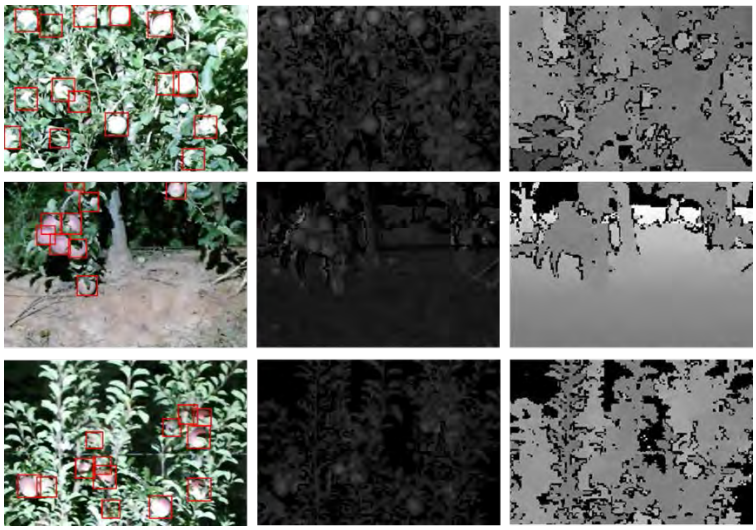


Figure 4. Sample of 3 multi-modal images extracted from training dataset and their associated fruit location ground truth (red bounding boxes). First column corresponds to RGB_{hr}, second column to S and the third column to D channel.

3.3 Experiments

The Faster R-CNN object detection network (Ren et al., 2017) was used in this work as fruit detector. The choice of Faster R-CNN allows a comparison of the performance of our methodology with previous works that also used Faster R-CNN for fruit detection (Bargoti and Underwood, 2017a; Gan et al., 2018; Sa et al., 2016; Stein et al., 2016).

Faster R-CNN was originally developed to detect objects in color images. The original work (Ren et al., 2017) tested the network with PASCAL VOC (Everingham et al., 2010) and COCO (Lin et al., 2014) datasets, reporting a mean average precision (mAP) of 78.8% and 42.7% for the VOC 2007 and COCO test sets, respectively. Faster R-CNN is composed of two modules: (1) a region proposal network (RPN), to identify promising regions of interest (ROIs) that are likely to contain an object; (2) a classification network, which classifies the regions proposed. Both parts share the first convolutional layers, making it a fast object detector. The RPN uses the feature maps produced by the first convolutional layers to produce promising ROIs by means of a series of convolutional and fully connected layers. The RPN output is then used to crop out corresponding regions from the feature maps produced by the first convolutional layers (crop pooling). The regions produced by crop pooling are then passed through a classification network and a regressor to predict the probability of a ROI being apple or background and refine the ROI. Figure 5 illustrates a diagram of the implemented Faster R-CNN network.

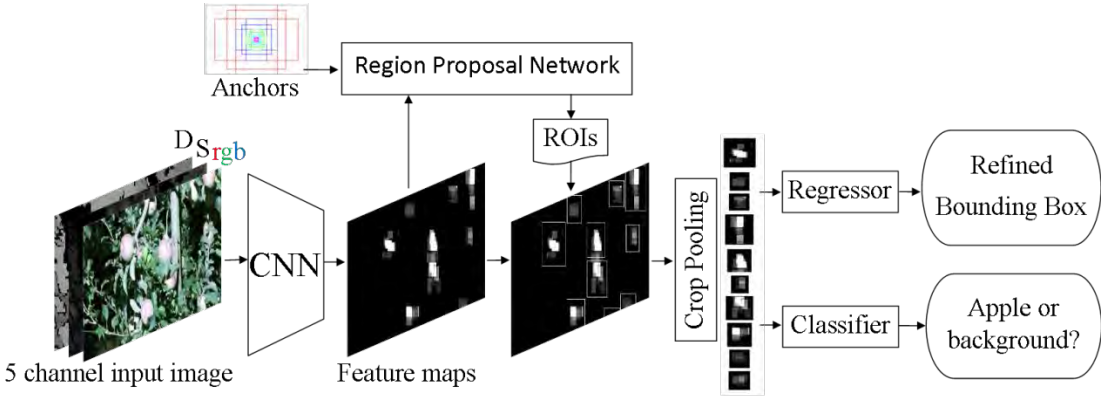


Figure 5. Diagram of the implemented Faster R-CNN. The main modifications to the original Faster R-CNN are the multi-modal input and the anchor scales.

In this work, the first convolutional layers uses the VGG-16 model (Simonyan and Zisserman, 2014) pre-trained with ImageNet dataset (Deng et al., 2009) and fine-tuned with our training dataset. The original implementation of Faster R-CNN was modified to make it suitable for our dataset. The main modifications to the original Faster R-CNN done in this work are: (1) multi-modal input and (2) region proposal adaptation.

Regarding the multi-modal input, since the original implementation of Faster R-CNN uses color images, the input layer was modified to work with 5-channel images. Due to these additional channels, filters from the first convolutional layer increase in depth (from 3 to 5), which implies that more weights must be initialized. Thus, after loading pre-trained weights in the network, additional weights corresponding to channels D and S were randomly initialized.

To generate region proposals, the RPN evaluates different boxes in each position of the image with a stride of 16 pixels. The different types of boxes evaluated are called anchors and are characterized by their scale (box area) and the aspect ratio. The original implementation of Faster R-CNN proposed 3 anchor scales of 8, 16 and 32 -corresponding to box areas of 128^2, 256^2 and 512^2 pixels-, and 3 aspect ratios of 1:1, 1:2 and 2:1. Since the presented dataset has smaller objects than datasets tested in Ren et al. (2017), a study of the optimal anchor scales and aspect ratios was carried out. Besides the anchors proposed in the original paper (8, 16, 32), smaller anchor scales were also tested (2 and 4). The aspect ratios used in this study were the same as those used in the original implementation (1:1, 1:2 and 2:1), however, two different configurations were tested: only considering aspect ratio of 1:1, and combining the three aspect ratios. Figure 6 illustrates the anchors tested in this work compared with the image size. Although using input images of 548 x 373 px, the network resizes the input images to 600 px on the shortest side. For this reason, the shortest side has 600 px instead of 373 px.

To evaluate the performance, average precision (AP), precision, recall and F1-score metrics are reported for the test dataset. Predictions were considered as true positive if the intersection over union (IoU) between prediction and ground truth bounding boxes was greater than 0.5. The network was implemented in PyTorch framework (Paszke et al., 2017) and has been made publicly available at *www.grap.udl.cat/en/publications/datasets.html.*

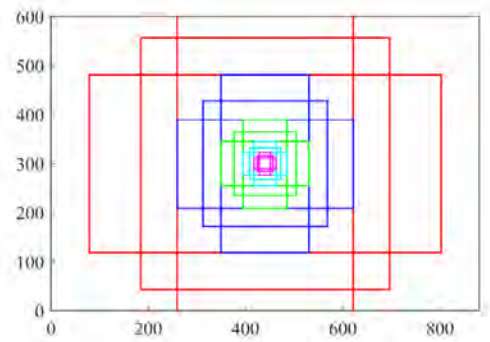


Figure 6. Anchors tested compared with the image size. Five anchor scales were tested: 2 (magenta), 4 (cyan), 8 (green), 16 (blue) and 32 (red). For each anchor, three different aspect ratios were used: 0.5, 1 and 2.

4. Results and discussion

This section presents a qualitative and quantitative evaluation of the proposed fruit detection methodology, assessing the performance when using different image modalities provided by Kinect v2, and studying the optimal anchor scale configurations proposed in the RPN.

To study the usefulness of different image modalities (RGB, S and D), different Faster R-CNN models were trained using each modality separately as well as combinations thereof. This study was carried out using projected color images RGB_p since they have the same resolution as S and D images, to enable a comparison between modalities performed under the same conditions. Nevertheless, results with 5-channel multi-model images using RGB_{hr} are also provided to assess the potential of the sensor for fruit detection.

4.1 Training assessment

The network was trained end-to-end using the loss function proposed in Ren et al. (2017), which is comprised of the sum of a classification loss and a bounding box regression loss. Following Ren et al. (2017), the training loss function considered positive detections if IoU>0.7 and negative if IoU<0.3, while anchors that are neither positive nor negative do not contribute to the loss function. Adam optimizer with a learning rate of 0.0001 was used to update network weights iteratively, performing a total of 309 training (38 validation) iterations per epoch with a batch size of 4 images. Data augmentation was performed with left-right flipping to expand the variability of the training dataset. A validation set was used to evaluate the training after each epoch to check model generalization and identify if it starts to overfit. The number of images

used for training, validation and test were 619, 155 and 193, respectively. Figure 7 shows the loss function for training and validation sets using different image modalities. By comparing models where RGB_p was used, it can be seen how when only using color modality (plotted in cyan) the model starts to overfit earlier than with the addition of S and D channels (plotted in orange and green). Therefore, the use of S and D channels allowed model training during more iterations without overfitting. With respect to training curves, the loss function archived lower values when using only color images. However, the opposite occurred with validation losses, with the best validation loss achieved by combining all modalities. This is a consequence of early overfitting of the RGB_p model, and, from that, it was concluded that the S and D channels helped model generalization, with better results obtained on the validation dataset when using 5-channel multi-modal images. On the other hand, when comparing the performance of using RGB_p or RGB_{hr} images, training and validation loss functions showed an important improvement when RGB_{hr} images were used. This improvement increased when adding S and D channels to RGB_{hr}, although not in the same proportion as multi-modal images with RGB_p. This suggests that if future RGB-D sensors had depth sensors with higher resolution (similar to color cameras), detection performance could be improved even further.

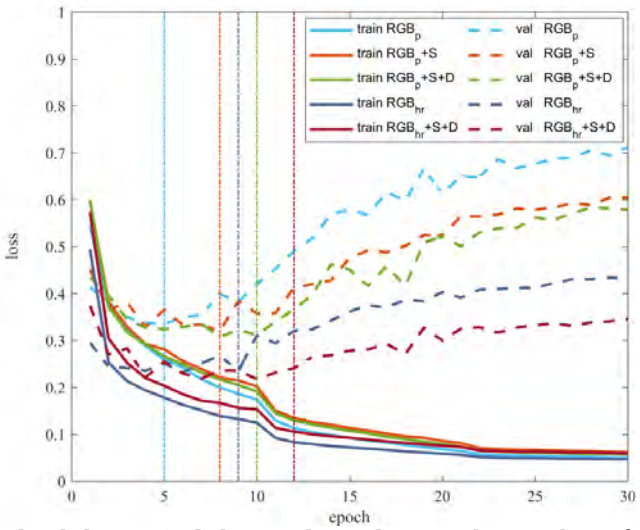


Figure 7. Training and validation (val) losses depending on the number of training epochs. Loss function was computed following Ren et al. (2017). In cyan, the evolution of training and validation losses is plotted using RGB projected images (RGB_p). Orange data refer to projected images with range-corrected signal intensity (S) data, and green data when adding depth (D) information as well. Shown in blue are the results corresponding to use of high resolution RGB (RGB_{hr} -registered row image-). Finally, results of multi-modal data with RGB_{hr}, range-corrected signal and depth images are plotted in red. Vertical lines mark the epoch where each model starts to overfit, which is the epoch in which test results are reported.

4.2 Anchor optimization

This section evaluates the performance of Faster R-CNN with multi-modal images (RGB_{hr}+S+D) depending on the anchor scales used in the RPN. The original paper of Faster R-CNN (Ren et al., 2017) used anchor scales of 8-16-32, but it mentions that the anchor scales used were not specifically chosen for a particular dataset. The present work was evaluated on a very different dataset (with small spherical objects) from the one used in the original Faster R-CNN work. Therefore, the behavior of the network using anchor scales of 2, 4, 8, 16 and 32 and aspect ratios of 1:2, 1:1, 2:2 was analyzed.

Table 3 presents the results obtained when using different anchor scales and aspect ratios in terms of AP. Comparing anchor scales configurations, the worst performances were achieved using anchor scales of 16 and 32, while other configurations performed similarly, with an AP ranging between 93.4% (using anchor scale of 8 and aspect ratios of 1:2, 1:1 and 2:1), and 94.8% (using anchor scales of 4 and aspect ratios of 1:1). Regarding the anchor aspect ratios, best results were obtained only using squared anchors (anchor ratios of 1:1). This responds the fact that fruits are spherical. As for computational efficiency, Table 3 shows that frame rate slightly decreases when combining different anchor scales or aspect ratios. This fact is due to the number of convolutional operations in the RPN increases. However, since the number of object proposals was limited to 100 in all cases (as suggested in Sa et al. 2016) the computational efficiency do not show important differences. From these results, the following sections use anchor scales of 4 and aspect ratios of 1:1, being the configuration that showed the best performance.

Table 3. Fruit detection results on RGB_{hr}+S+D test set using different anchor scales and ratios.

	Anchor aspect ratios			
	[1:2 , 1:1, 2:1]		[1:1]	
Anchor scales	AP (%)	frames/s	AP (%)	frames/s
8-16-32 (Ren et al., 2017)	93.1	12.7	92.8	12.9
2	93.6	12.7	94.0	13.1
4	93.5	12.6	**94.8**	13.6
8	93.4	12.4	93.7	13.3
16	91.3	13.0	86.6	13.2
32	79.9	12.9	86.4	13.0
2-4-8	94.1	12.7	94.4	12.4
4-8-16	94.1	12.6	93.5	12.1
2-4-8-16	93.1	12.8	94.1	12.8

Figure 8 illustrates some fruit detection examples using anchor scales of 4 and aspect ratios of 1:1. Images were selected to show cases where the network succeeds or fails, so that the illustrated examples correspond to the four best (first column), four intermediate (second column) and four worst (third column) scored images from the test dataset. As can be seen, most of the false positives correspond to image regions that are very similar to apples or to real apples that were not labelled because of human errors when labelling. On the other hand, most of the false negatives correspond to highly occluded apples and to apples that were cut by the borders of the image.

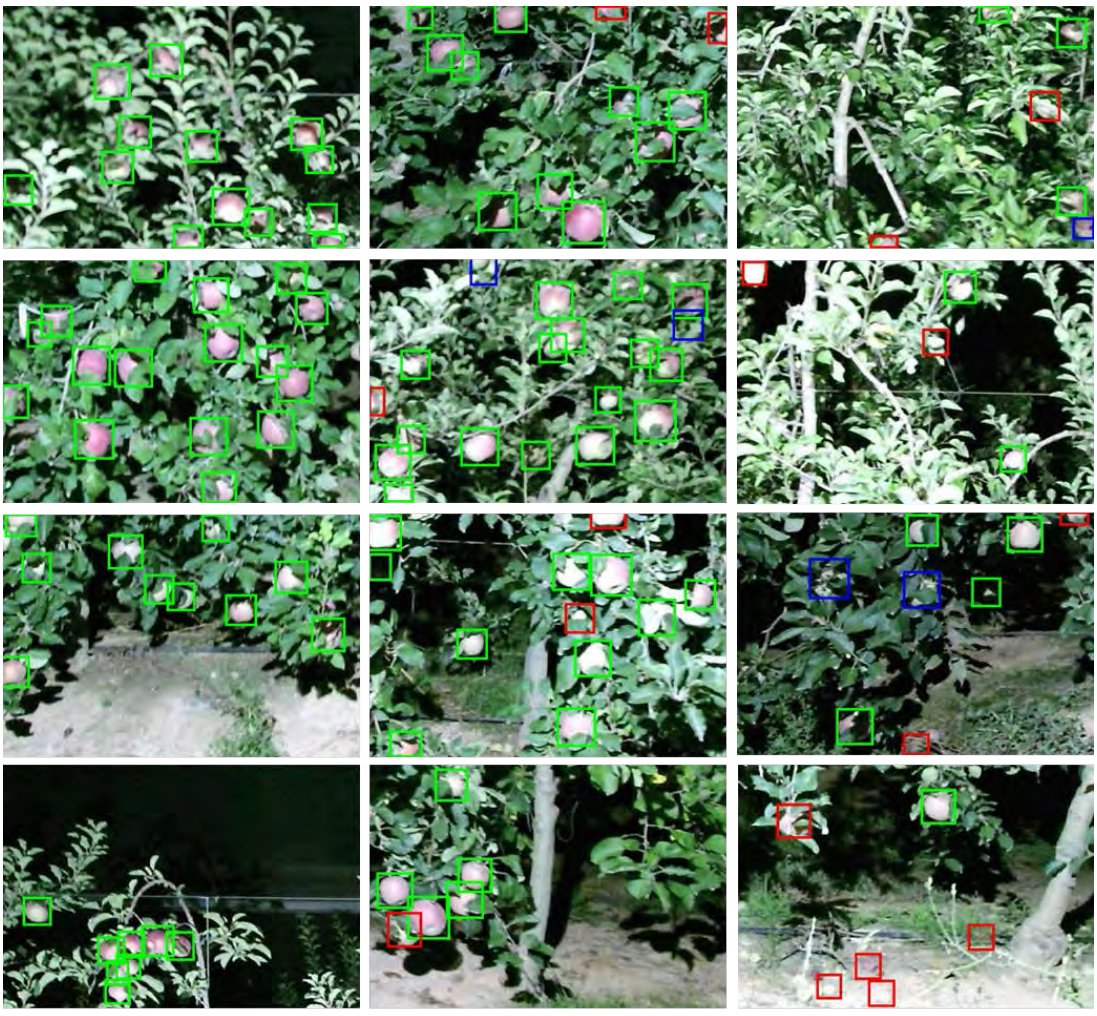


Figure 8. Fruit detection results on RGB_{hr}+S+D test set using anchor scales of 4, 8 and 16. True positive detections are shown in green squares, false positives in red and false negatives in blue. Images are ordered according to their F1-score. Column (a) contains examples of the best detection results, F1-score = 1. Column (b) contains the four intermediate scored images of the test set, corresponding to an F1-score =0.91. Column (c) shows the worst detections, ordered from bottom to top with an F1-score = [0.23, 0.67, 0.67, 0.67].

4.3 Test results from different modalities

Regarding the test set, Table 4 presents fruit detection results obtained from different input image modalities. The performance of Faster R-CNN using each input type was evaluated in terms of Precision (P), Recall (R), F1-score, AP and number of inferred images per second (processing on a GeForce GTX TITAN X GPU). A confidence threshold of 0.85 was selected from Precision and Recall curves (Figure 9). The number of training epoch is also given. This number was chosen from training and validation loss curves, selecting the last epoch that did not present overfit (vertical lines in Figure 7). Figure 10 shows graphically the fruit detection of three selected images from the test set using different input modalities (RGB_p, S, D, RGB_p+S and RGB_p+S+D). True positives are shown in green, false positives in red and false negatives in blue.

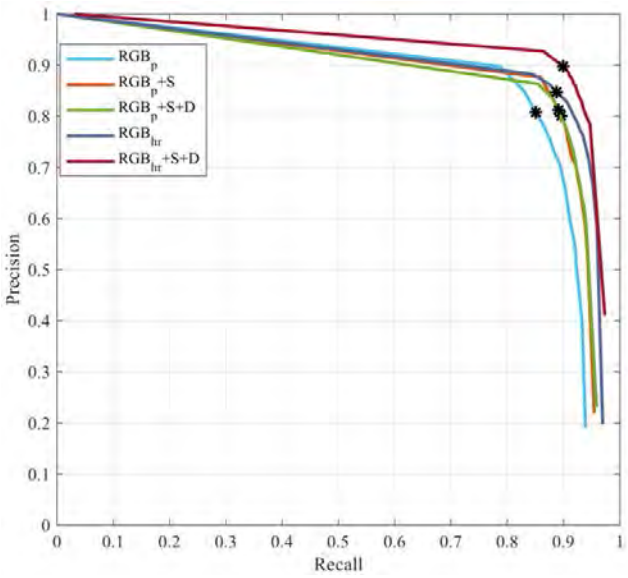


Figure 9. Precision and Recall curves obtained for different image modalities. Black asterisks correspond to the working points with the selected confidence threshold of 0.85.

Table 4. Fruit detection results from test dataset using different image modalities.

Channels	Epoch	P	R	F1-score	AP (%)	frames/s
RGB_p	5	0.808	0.851	0.829	88.7	13.4
S	4	0.848	0.768	0.806	85.9	13.5
D	7	0.699	0.582	0.635	61.3	13.4
RGB_p+S	8	0.887	0.827	0.856	89.8	12.9
RGB_p+D	9	0.802	0.848	0.824	88.0	13.7
S+D	9	0.731	0.821	0.774	84.6	13.4
RGB_p+S+D	10	**0.869**	**0.864**	**0.866**	**91.2**	13.1
RGB_{hr}	9	0.847	0.888	0.867	92.7	12.9
RGB_{hr}+S+D	12	**0.897**	**0.899**	**0.898**	**94.8**	13.6

Comparing results from single modality images (Table 4, rows 1-3), color images gave the best performance with an F1-score of 0.829 and an AP of 88.7%, followed by the range-corrected intensity image with an F1-score of 0.806 and an AP of 85.9%. Note that, although range-corrected intensity images have never been used for fruit detection, the results using this modality are comparable with other state-of-the-art methods. The least valuable modality was the Depth channel which was only able to detect highly exposed (non-occluded) apples, as can be seen in Figure 10. Better results were obtained when combining different modalities (multi-modal images), achieving an F1-score of 0.866 and an AP of 91.2% when all channels were used. The most important benefit of adding S and D was found in the Precision metric, which rose from 0.808 (RGB_p) to 0.869 (RGB_p+S+D), although Recall and AP also improved albeit in smaller percentages. This means that range-corrected intensity and depth images help to reduce false positives. Real examples of this effect can be found in Figure 10, where, when comparing results before and after using S and D channels, a reduction in false positives is observed. Another advantage of using the S channel was found when detecting fruits in shadowed regions, where the RGB image presents a dark non-colored region whereas the S channel shows high intensities. This occurs in Figure 10b, where an apple in a shadowed region was not detected using RGB_p, but was detected using the S channel.

Finally, as was expected, the best performance was achieved using multi-modal images with RGB_{hr}, S and D, reporting an F1-score of 0. 898 and a AP of 94.8%. Regarding the computational efficiency of the neural network, the number of inferred images per second did not present any relation with the number of channels used. This is because the addition of channels only affects the number of operations on the first layer, which is insignificant with respect to the whole network.

Although it is difficult to compare methodologies tested with different datasets, results shows similar performance to other fruit detection works based on neural networks, such as Bargoti and Underwood (2017), Gan et al. (2018) and Sa et al. (2016) which reported F1-scores between 0.838 and 0.929 (using less restrictive IoU thresholds than the present work). However, the use of RGB-D sensors has the advantage that, although detecting fruits in 2D images, it is straightforward to infer the 3D location of each detection.

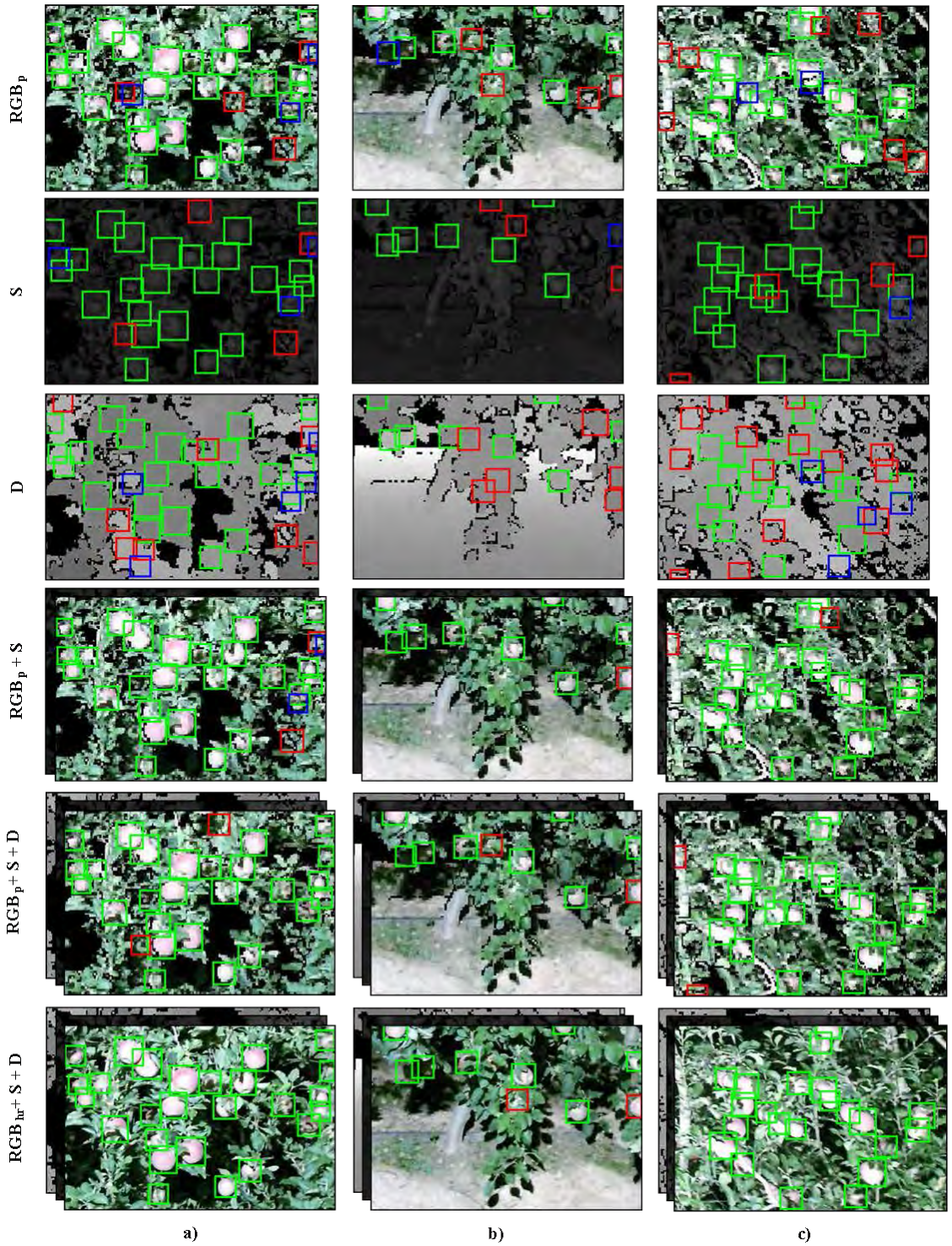


Figure 10. Selected examples of fruit detection results to show the effect of adding range-corrected signal intensity (S) and depth (D) information. For each sample a), b) and c), six different fruit detection results are shown depending on the input data type: RGB_p (first row), S (second row), D (third row), multi-modal RGB_p and S (fourth row), using all modalities RGB_p, S and D (fifth row), and using all modalities with high resolution image RGB_{hr}, S and D (last row). True positive detections are shown in green, false positives in red and false negatives in blue.

The main limitation of the proposed methodology is that the working conditions are restricted to low illuminance levels. However, we expect that future sensors could solve this limitation. For instance, LiDAR-based sensors are already able to build 3D models with reflectance data in natural lighting conditions. In addition, convolutional neural networks have shown good performances with wide enough datasets that contain different illumination conditions (Amara et al., 2017; Chen et al., 2017; Rahnemoonfar and Sheppard, 2017). From that, we expect that if future RGB-D sensors would not be influenced by high illumination levels, the methodology proposed could be used in daylight.

5. Conclusions

This work presents a novel methodology for fruit detection using RGB-D sensors, taking advantage of their radiometric capabilities. Multi-modal images were built using data provided by Microsoft's Kinect v2. To do so, a range correction of the backscattered intensity signal was carried out to overcome signal attenuation (R^{-2} dependence). Then, a registration between different channels was performed, obtaining images with 3 modalities: color (RGB), depth (D), and range-corrected intensity (S). The KFuji RGB-DS dataset and the corresponding annotations have been made publicly available, being the first dataset for fruit detection that contains RGB, depth and range-corrected intensity channels. The Faster R-CNN object detection network was used to evaluate the usefulness of fusing all modalities. Results show an improvement of 4.46% in F1-score when all modalities were used -from 0.829 (RGB_p) to 0.866 (RGB_p-D-S)-. This entails an advance in the field of fruit detection, since the results are comparable to other fruit detection methodologies retrieved from the state of the art, with the additional advantage that, by using RGB-D sensors, it is possible to infer the 3D position of each detection. The optimum anchor scales used in the region proposal network were also analyzed. It is concluded that, for KFuji RGB-DS dataset where fruits are spherical and small with respect to the image size, the optimum configuration were anchor scales of 4 and aspect ratios of 1:1. The main limitation of using RGB-D is that the performance of the depth sensor drops under direct sunlight. Future works will include 3D fruit localization by projecting the 2D fruit detection onto the 3D world

using the depth channel data as well as collecting data of different fruit varieties and at different growth stages.

Acknowledgements

This work was partly funded by the Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya, the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund (ERDF) under Grants 2017 SGR 646, AGL2013-48297-C2-2-R and MALEGRA, TEC2016-75976-R. The Spanish Ministry of Education is thanked for Mr. J. Gené's pre-doctoral fellowships (FPU15/03355). We would also like to thank Nufri and Vicens Maquinària Agrícola S.A. for their support during data acquisition, and Adria Carbó for his assistance in Faster R-CNN implementation.

References

Amara, J., Bouaziz, B., Algergawy, A., 2017. A Deep Learning-based Approach for Banana Leaf Diseases Classification. Btw 79–88.

Auat Cheein, F.A., Carelli, R., 2013. Agricultural robotics: Unmanned robotic service units in agricultural tasks. *IEEE Ind. Electron. Mag.* 7, 48–58. doi:10.1109/MIE.2013.2252957

Bargoti, S., 2016. Pychet Labeller. Available online: https://github.com/acfr/pychetlabeller.

Bargoti, S., Underwood, J., 2017a. Deep Fruit Detection in Orchards. *2017 IEEE Int. Conf. Robot. Autom. 3626–3633.* doi:10.1109/ICRA.2017.7989417

Bargoti, S., Underwood, J.P., 2017b. Image Segmentation for Fruit Detection and Yield Estimation in Apple Orchards. *J. F. Robot.* 34, 1039–1060. doi:10.1002/rob.21699

Barnea, E., Mairon, R., Ben-Shahar, O., 2016. Colour-agnostic shape-based 3D fruit detection for crop harvesting robots. *Biosyst. Eng.* 146, 57–70. doi:10.1016/j.biosystemseng.2016.01.013

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

Chen, S.W., Shivakumar, S.S., Dcunha, S., Das, J., Okon, E., Qu, C., Taylor, C.J., Kumar, V., 2017. Counting Apples and Oranges With Deep Learning: A Data-Driven Approach. *IEEE Robot. Autom. Lett.* 2, 781–788. doi:10.1109/LRA.2017.2651944

Chhokra, P., Chowdhury, A., Goswami, G., Vatsa, M., Singh, R., 2018. Unconstrained Kinect video face database. *Inf. Fusion* 44, 113–125. doi:10.1016/j.inffus.2017.09.002

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database, in: *2009 IEEE Conference on Computer Vision and Pattern Recognition.* doi:10.1109/CVPRW.2009.5206848

Escolà, A., Martinez-Casasnovas, J.A., Rufat, J., Arno, J., Arbones, A., Sebe, F., Pascual, M., Gregorio, E., Rosell-Polo, J.R., 2017. Mobile terrestrial laser scanner applications in

precision fruticulture/horticulture and tools to extract information from canopy point clouds. *Precis. Agric.* 18, 111–132. doi:10.1007/s11119-016-9474-5

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2010. The pascal visual object classes (VOC) challenge. *Int. J. Comput.* Vis. 88, 303–338. doi:10.1007/s11263-009-0275-4

Font, D., Pallejà, T., Tresanchez, M., Runcan, D., Moreno, J., Martínez, D., Teixidó, M., Palacín, J., 2014. A proposal for automatic fruit harvesting by combining a low cost stereovision camera and a robotic arm. *Sensors (Switzerland)* 14, 11557–11579. doi:10.3390/s140711557

Gan, H., Lee, W.S., Alchanatis, V., Ehsani, R., Schueller, J.K., 2018. Immature green citrus fruit detection using color and thermal images. *Comput. Electron. Agric.* 152, 117–125. doi:10.1016/j.compag.2018.07.011

Gené-Mola, J., Gregorio, E., Guevara, J., Auat, F., Escolà, A., Morros, J.-R., Rosell-Polo, J.R., 2018. Fruit Detection Using Mobile Terrestrial Laser Scanning, in: EurAgEng 2018 Conference.

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., Karkee, M., Amatya, S., 2018. Apple fruit size estimation using a 3D machine vision system. *Inf. Process. Agric.* 5, 498–503. doi:10.1016/j.inpa.2018.06.002

Hameed, K., Chai, D., Rassau, A., 2018. A comprehensive review of fruit and vegetable classification techniques. *Image Vis. Comput.* 80, 24–44. doi:10.1016/j.imavis.2018.09.016

Höfle, B., Pfeifer, N., 2007. Correction of laser scanning intensity data: Data and model-driven approaches. *ISPRS J. Photogramm. Remote Sens.* 62, 415–433. doi:10.1016/j.isprsjprs.2007.05.008

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: *European Conference on Computer Vision.* pp. 740–755. 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, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C., 2016. SSD: Single shot multibox detector, in: *European Conference on Computer Vision.* pp. 21–37. doi:10.1007/978-3-319-46448-0_2

Maldonado, W., Barbosa, J.C., 2016. Automatic green fruit counting in orange trees using digital images. *Comput. Electron. Agric.* 127, 572–581. doi:10.1016/j.compag.2016.07.023

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

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. *J. F. Robot.* 31(5), 837–860. doi:10.1002/rob.21541

Okamoto, H., Lee, W.S., 2009. Green citrus detection using hyperspectral imaging. *Comput. Electron. Agric.* 66, 201–208. doi:10.1016/j.compag.2009.02.004

Paszke, A., Chanan, G., Lin, Z., Gross, S., Yang, E., Antiga, L., Devito, Z., 2017. Automatic differentiation in PyTorch. *Adv. Neural Inf. Process. Syst.* 30.

Rahnemoonfar, M., Sheppard, C., 2017. Deep count: Fruit counting based on deep simulated learning. *Sensors (Switzerland)* 17, 1–12. doi:10.3390/s17040905

Redmon, J., Farhadi, A., 2017. YOLO9000: Better, faster, stronger, in: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017.* doi:10.1109/CVPR.2017.690

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

Rodríguez-Gonzálvez, P., Gonzalez-Aguilera, D., González-Jorge, H., Hernández-López, D., 2016. Low-Cost Reflectance-Based Method for the Radiometric Calibration of Kinect 2. *IEEE Sens. J.* 16, 1975–1985. doi:10.1109/JSEN.2015.2508802

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. *Adv. Agron.* 133, 71–112. doi:10.1016/bs.agron.2015.05.002

Rosell-Polo, J.R., Gregorio, E., Gene, J., Llorens, J., Torrent, X., Arno, J., Escola, A., 2017. Kinect v2 sensor-based mobile terrestrial laser scanner for agricultural outdoor applications. *IEEE/ASME Trans. Mechatronics* 22, 2420–2427. doi:10.1109/TMECH.2017.2663436

Rosell Polo, J.R., Sanz, R., Llorens, J., Arnó, J., Escolà, A., Ribes-Dasi, M., Masip, J., Camp, F., Gràcia, F., Solanelles, F., Pallejà, T., Val, L., Planas, S., Gil, E., Palacín, J., 2009. A tractor-mounted scanning LIDAR for the non-destructive measurement of vegetative volume and surface area of tree-row plantations: A comparison with conventional destructive measurements. *Biosyst. Eng.* 102, 128–134. doi:10.1016/j.biosystemseng.2008.10.009

Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T., McCool, C., 2016. DeepFruits: A Fruit Detection System Using Deep Neural Networks. *Sensors* 16, 1222. doi:10.3390/s16081222

Safren, O., Alchanatis, V., Ostrovsky, V., Levi, O., 2007. Detection of Green Apples in Hyperspectral Images of Apple-Tree Foliage Using machine Vision. Trans. ASABE 50, 2303–2313. doi:10.13031/2013.24083

Sanz, R., Llorens, J., Escolà, A., Arnó, J., Planas, S., Román, C., Rosell-Polo, J.R., 2018. LIDAR and non-LIDAR-based canopy parameters to estimate the leaf area in fruit trees and vineyard. *Agric. For. Meteorol.* 260–261, 229–239. doi:10.1016/j.agrformet.2018.06.017

Siegel, K.R., Ali, M.K., Srinivasiah, A., Nugent, R.A., Narayan, K.M.V., 2014. Do we produce enough fruits and vegetables to meet global health need? PLoS One 9 (8), e104059. doi:10.1371/journal.pone.0104059

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv 1409.1556*. doi:10.1016/j.infsof.2008.09.005

Stajnko, D., Lakota, M., Hocevar, M., 2004. Estimation of number and diameter of apple fruits in an orchard during the growing season by thermal imaging. *Comput. Electron. Agric*. 42, 31–42. doi:10.1016/S0168-1699(03)00086-3

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

Underwood, J.P., Hung, C., Whelan, B., Sukkarieh, S., 2016. Mapping almond orchard canopy volume, flowers, fruit and yield using lidar and vision sensors. *Comput. Electron. Agric.* 130, 83–96. doi:10.1016/j.compag.2016.09.014

Uribeetxebarria, A., Daniele, E., Escolà, A., Arnó, J., Martínez-Casasnovas, J.A., 2018. Spatial variability in orchards after land transformation: Consequences for precision agriculture practices. *Sci. Total Environ.* doi:10.1016/j.scitotenv.2018.04.153

Wang, Z., Walsh, K.B., Verma, B., 2017. On-Tree Mango Fruit Size Estimation Using RGB-D Images. *Sensors (Basel).* 17. doi:10.3390/s17122738

Zhang, B., Huang, W., Wang, C., Gong, L., Zhao, C., Liu, C., Huang, D., 2015. Computer vision recognition of stem and calyx in apples using near-infrared linear-array structured light and 3D reconstruction. *Biosyst. Eng.* 139, 25–34. doi:10.1016/j.biosystemseng.2015.07.011

Zhao, C., Lee, W.S., He, D., 2016. Immature green citrus detection based on colour feature and sum of absolute transformed difference (SATD) using colour images in the citrus grove. *Comput. Electron. Agric.* 124, 243–253. doi:10.1016/j.compag.2016.04.009

Chapter VII. P7: Fruit detection and 3D location using instance segmentation neural networks and structure-from-motion photogrammetry

This chapter was published in *Computers and Electronics in Agriculture* 169 (2020) 105165, https://doi.org/10.1016/j.compag.2019.105165:

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

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 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 (latitude: 41º44'47.07"N; longitude: 1º01'52.23"E). 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 (Figure 1a) from the lower part (soil-trunk) to the upper part of the trees. The separation between two consecutive positions was 22 cm (Figure 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 (Figure 1a). With this configuration, the vertical and horizontal overlapping between neighbouring images was higher than 30% and 90%, respectively (Figure 2). This dataset has been made publicly available at www.grap.udl.cat/en/publications/datasets.html (Fuji-SfM dataset).

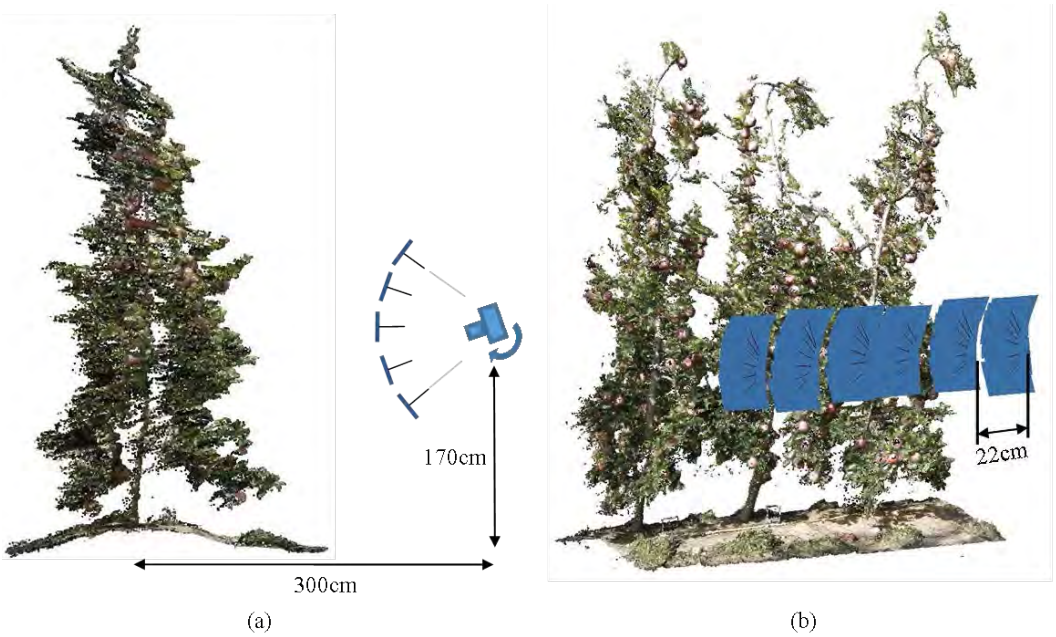


Figure 1. a) Transversal scheme of the layout and distances of the photographic process. b) Isometric view of three scanned trees showing the separation between consecutive photographic positions.

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

2.2 Methodology pipeline

As shown in Figure 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).

Image cropping step, the translation of detections to the original images, and the projection of 2D detections were processed with a 64-bit operating system, with 8GB of RAM and an Intel ® Core(TM) i7-4500U processor (1.80 GHz, boosted to 2.40 GHz). The instance segmentation step (Mask RCNN) was processed in a CPU+GPU machine with a GeForce GTX TITAN X GPU. The 3D model generation (SfM) was tested in the mentioned CPU computer, as well as in a a CPU+GPU machine with a GeForce GTX 1060. Further details of the implementation of these steps are described in the following sub-sections.

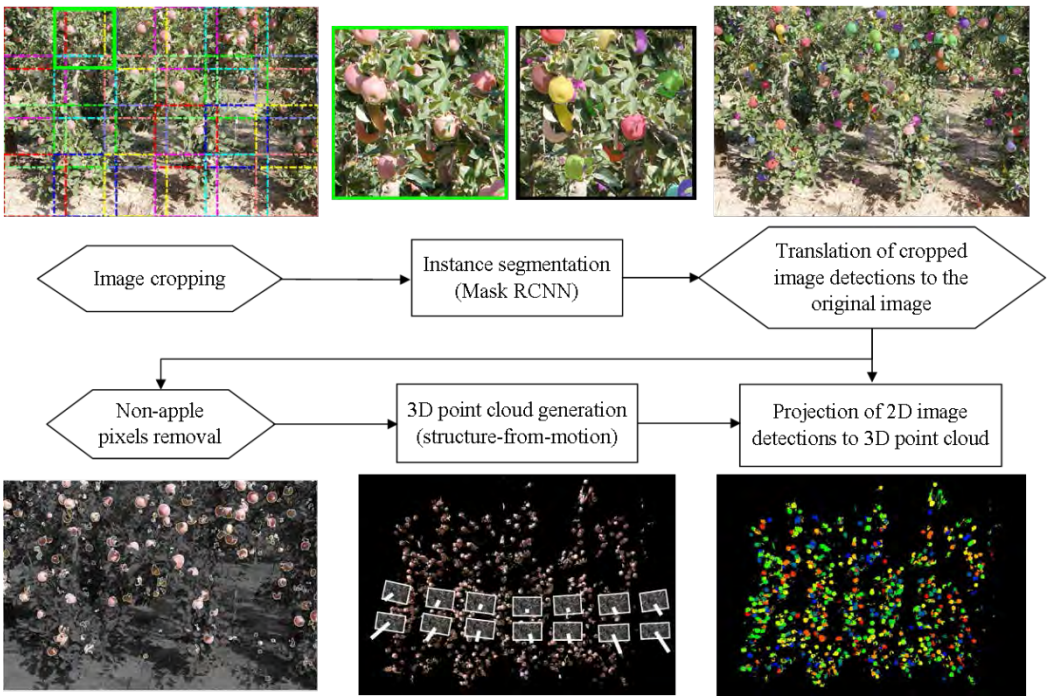


Figure 3. Fruit detection and location methodology flowchart. Hexagons represent data preparation steps while rectangles define data processing steps.

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 Figure 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.

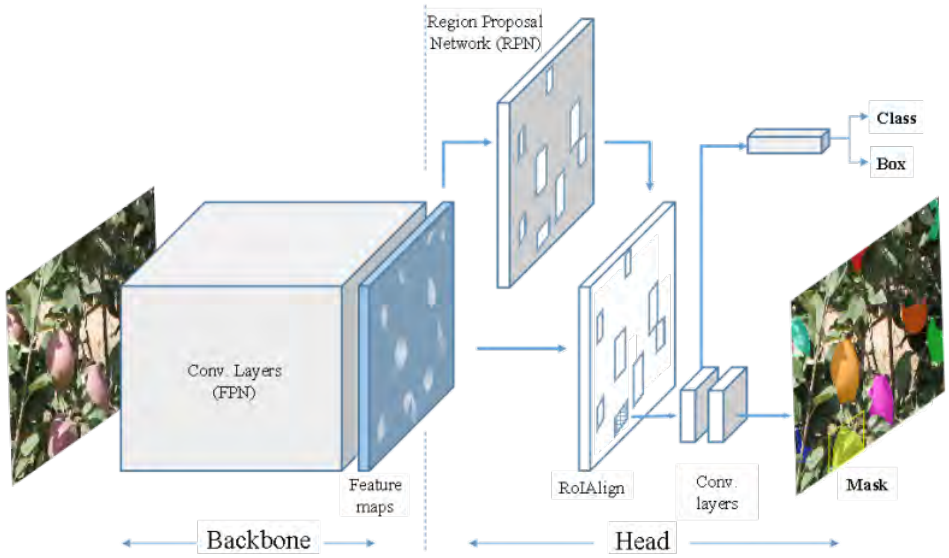


Fig. 4. Diagram of Mask R-CNN architecture.

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 apples that were manually labelled using the VIA annotation software (Dutta and Zisserman, 2019). 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 division s). An overlap between neighbouring sub-images of 213 px in vertical and 192 px in horizontal was applied to avoid the partially split of fruits at the boundaries in different partitions. 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

Instance segmentation results (Section 3.1) were assessed 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).

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 $T_0 = [0\ 0\ 0]$).

Figure 5a represents the 3D point cloud generated using original RGB images. This point cloud was manually annotated, placing rectangular bounding boxes around each apple (Figure 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 Figure 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 Figure 5d) separated by 85 cm were used to scale the resulting 3D point cloud to a real-world scale.

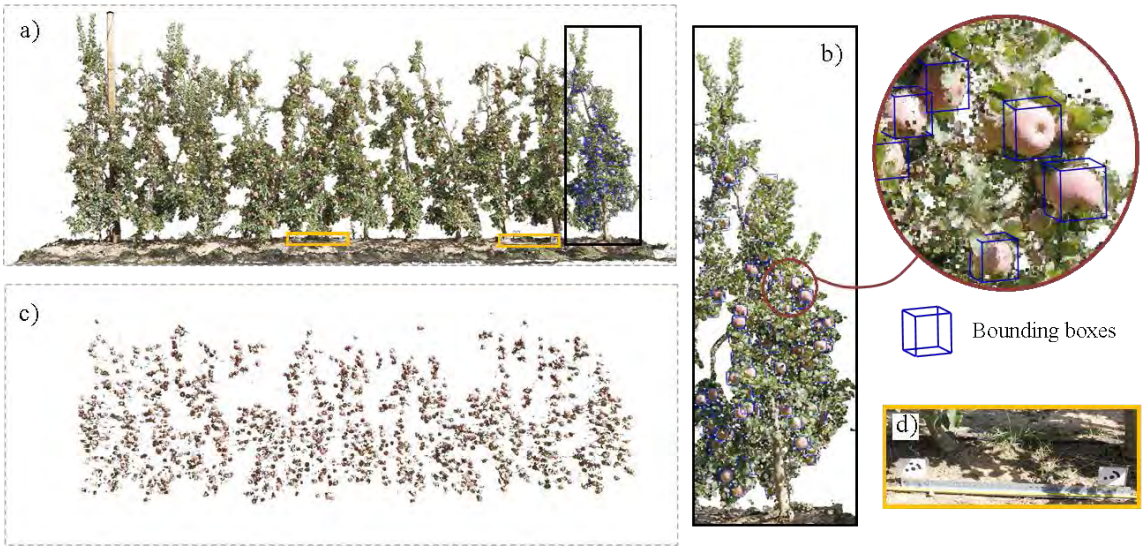


Figure 5. a) Illustration of the 3D point cloud obtained using original RGB images. Yellow rectangles show the positions where reference markers were placed. b) Annotated point cloud with 3D rectangular bounding boxes placed around each apple. c) Apples 3D point cloud obtained using masked images. d)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

Figure 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, Figure 6a shows a small region of the scanned scene and Figure 6b shows the 3D model obtained applying SfM photogrammetry with masked images. In Figure 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 Figure 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 Figure 6c and Figure 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 (Figure 6e), and new detections with no overlap with previous detections or with IoU < 0.5 were projected onto the 3D point cloud (Figure 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, Figure 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 Figure 6g, where the blue detection has been removed.

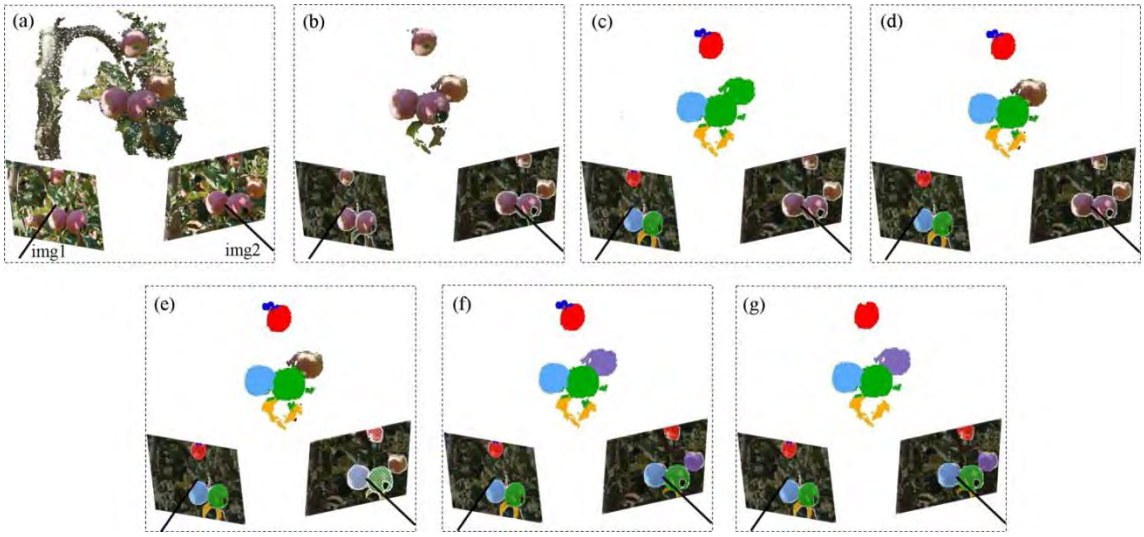


Figure 6. Projection of 2D detections onto 3D point cloud. a) Data acquisition. b) 3D model obtained using structure-from-motion with segmented images. c) Projection of detections from image 1 (img1) onto the 3D point cloud. d) Identification of apples behind detections. e) Identification of apples appearing 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.

3D fruit detection results (Section 3.2) were assessed in terms of detection rate (DR), recall (R), precision (P), false positive rate (FPR), muti-detection rate (MDR), and F1-score, as follows:

$$\mathrm{DR} = \frac{\mathrm{LD}}{\mathrm{T}}, \tag{3}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{T}}, \tag{4}$$

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{D}}, \tag{5}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{D}}, \tag{6}$$

$$\mathrm{MDR} = \frac{\mathrm{MD}}{\mathrm{D}}, \tag{7}$$

$$\mathrm{F1} = 2\frac{\mathrm{R}\cdot\mathrm{P}}{\mathrm{R}+\mathrm{P}}, \tag{8}$$

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. Results

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		

Figure 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 (green rectangles in Figure 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 Figure 7 b-d), or multi-detections (blue rectangles in Figure 7 a,e-f). As for the apples not detected, it can be seen that false negatives (yellow rectangles in Figure 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.

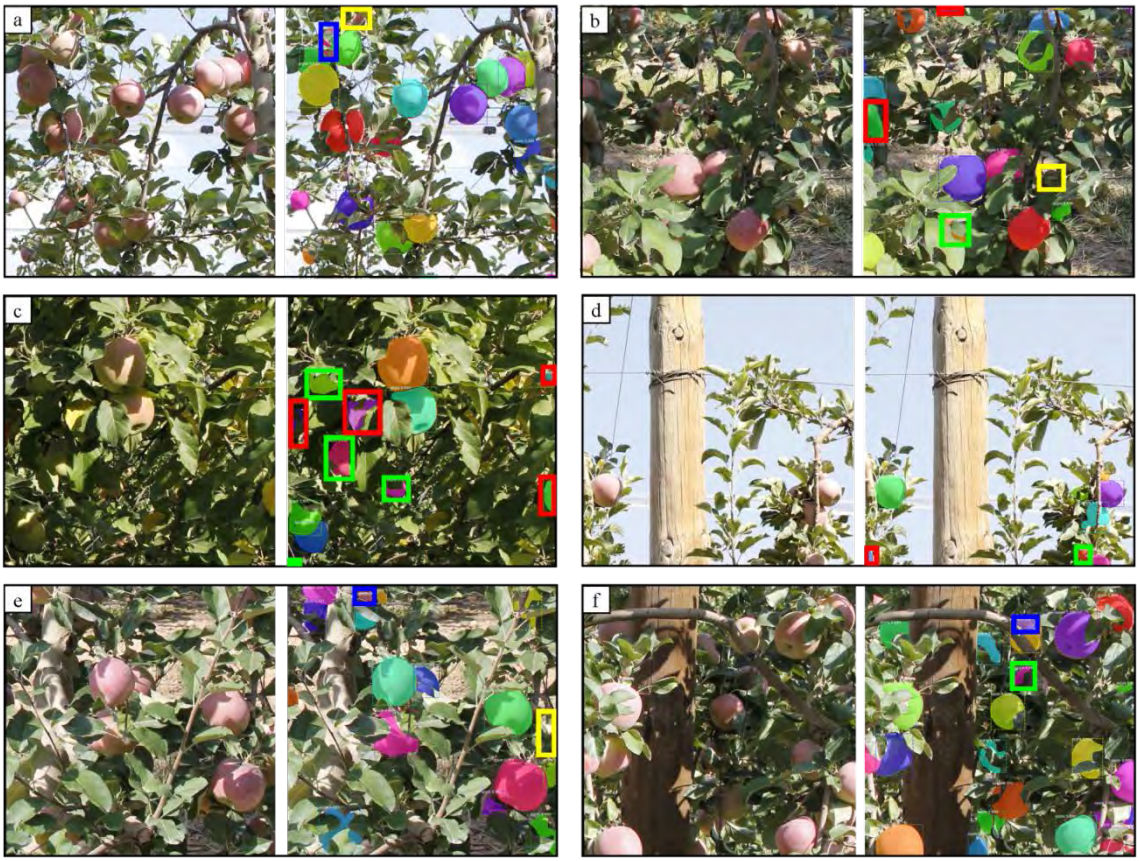


Figure 7. Selected examples of instance segmentation results to show correct detections (colour masks), false positives due to network failures (red rectangles), false positives due to miss-annotated apples (green rectangles), false positives due to multi-detections (blue rectangles), and false negatives (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 Figure 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). Figure 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.

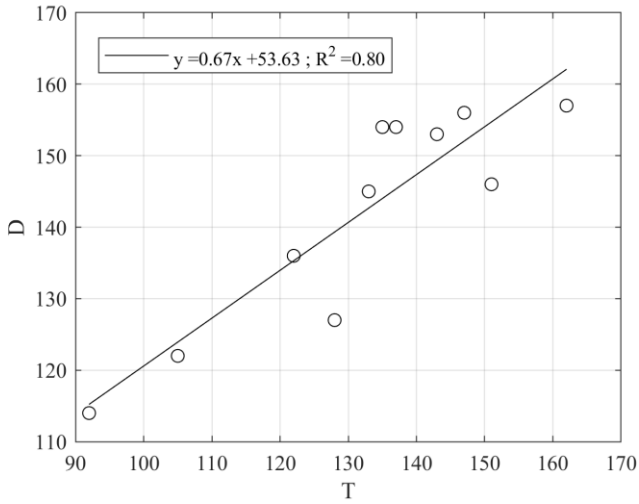


Figure 8. Linear regression between the number of detections (D) and the actual number of fruits per tree (T).

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 Figure 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.

Figure 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

Regarding the computational cost of the presented methodology, Table 4 includes the inference time of different processing steps implied in the presented methodology. The most computational expensive was the SfM photogrammetry, which required around 500 min to generate the 3D point cloud of the apples contained in the 11 tested trees in a conventional CPU computer. However, this processing time could be significantly reduced by processing this step in a graphic processing unite (GPU). The projection of 2D detections onto the 3D point cloud was also a computational expensive step, which required 260 min to process all images from the dataset. Since the code developed to project 2D detections onto the 3D point cloud was not parallelized, this step could not be processed in the CPU+GPU machine.

Table 4. Computational cost of processing steps implied in the developed methodology. The reported processing time corresponds to the time required to process all the dataset (11 trees, 582 images).

Process	Processing time	
	CPU	CPU+GPU
Instance segmentation (Mask RCNN)	---	35 min
3D point cloud generation (SfM)	500 min	50 min
Projection of 2D detections onto 3D point cloud	260 min	---

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 high amount of images to be processed and 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).

For yield prediction or yield mapping applications, the computational cost of the presented methodology is not a critical issue, as data can be processed offline. However, in the tests carried out in this work, data was acquired manually, being a labour and time consuming task when scanning larger areas. In order to automatize the data acquisition, some authors have used RGB-D sensors integrated on mobile platforms (Milella et al., 2019). Similarly, to optimize the data acquisition of the proposed methodology, future works should study the development of a compact system composed by different cameras mounted on a terrestrial platform.

5. Conclusions

This work proposes the combination of instance 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

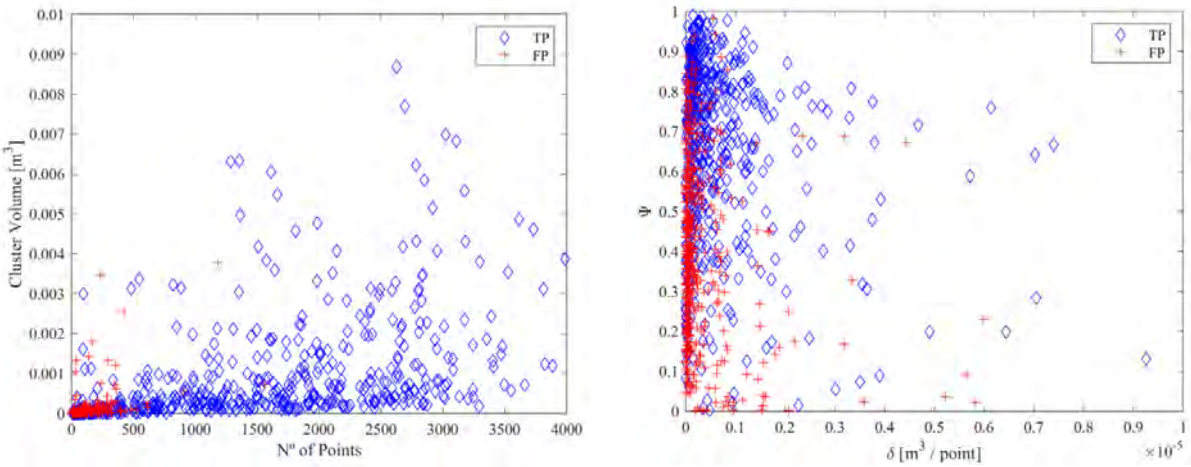


Figure B1 Graphical representation of apple detection features. The features analysed are the volume, number of points, the geometric parameter Ψ, and the detection point density δ. False positives are represented in red crosses; true positives are represented in blue diamonds. This analysis was performed 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, 888–911. 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.* 34, 1039–1060. 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.* 25, 104289. 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.* 2017, 2961–2969. 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)* 14, 20078–20111. 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: *European Conference on Computer Vision.* pp. 740–755. 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

Milella, A., Marani, R., Petitti, A., Reina, G., 2019. In-field high throughput grapevine

phenotyping with a consumer-grade depth camera. *Comput. Electron. Agric.* 156, 293–306. doi:10.1016/j.compag.2018.11.026

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. *J. F. Robot.* 31(5), 837–860. doi:10.1002/rob.21541

Redmon, J., Farhadi, A., 2018. YOLOv3: An Incremental Improvement. Tech Report, arXiv1804.02767.

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. *Adv. Agron.* 133, 71–112. 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.* 176, 172–184. 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

Chapter VIII. General discussion

So far, the fruit detection results obtained with different sensors have been discussed individually in the corresponding chapters. This chapter offers a global comparison and discussion of all the tested sensors and methodologies. Table 1 summarises the main results presented in this thesis. 2D fruit detection results were obtained with respect to the number of annotated fruits (visible fruits) because the actual number of fruits inside the field-of-view was not available. Conversely, 3D fruit detection methods were evaluated with respect to the total amount of fruits manually counted in the field.

Table 1. Comparison between different sensors and methods tested. Abbreviations: thresholding (Th); decision tree (D.Tree); support vector machine (SVM); structure-from-motion (SfM); colour data (RGB); range-corrected intensity data (S); depth data (D); data acquired from a single sensor position and without air-flow ($H_{1,n}$); data acquired combing air-flow conditions ($H_{1,(n+af)}$); data acquired from two different sensor positions ($H_{(1+2),n}$); precision (P); recall (R).

2D fruit detection						
Article	**Sensor**	**Method**	**Data**	**P**	**R**	**F1-score**
P6	RGB-D	FasterRCNN	RGB	0.847	0.888	0.867
P6	RGB-D	FasterRCNN	RGB+S+D	0.897	0.899	0.898
P7	RGB	MaskRCNN	RGB	0.857	0.860	0.858
3D fruit detection						
Article	**Sensor**	**Method**	**Data**	**P**	**R**	**F1-score**
P4	LiDAR	Th+D.Tree	$H_{1,n}$	0.846	0.729	0.783
P5	LiDAR	Th+SVM	$H_{1,n}$	0.875	0.710	0.784
P5	LiDAR	Th+SVM	$H_{1,(n+af)}$	0.839	0.763	0.799
P5	LiDAR	Th+SVM	$H_{(1+2),n}$	0.860	0.751	0.802
P7	RGB	Mask+SfM	SfM point cloud	0.857	0.906	0.881
Yield prediction						
Article	**Sensor**	**Method**	**Data**	**R^2**	**RMSE**	
P5	LiDAR	Th+SVM	$H_{(1+2),n}$	0.87	5.7%	
P7	RGB	Mask+SfM	SfM point cloud	0.80	8.2%	

1. 2D fruit detection

The deep neural network Faster RCNN was used in **Chapter VI** (P6) to detect fruits in 2D multimodal images. Results showed an improvement in fruit detection performance when adding the range-corrected intensity *S* and depth *D* channels to the colour image, reporting an increase of 3.1% in F1-score, from 0.867 to 0.898. Other authors have also

attempted to improve detection performance by combining different data modalities. For instance, Gan et al. (2018) obtained an increase of 2.3% in F1-score when combining colour and thermal images, while Sa et al. (2016) reported an increase of 2.2% in F1-score when combining colour and near-infrared images.

The methodology presented in **Chapter VII** (P7) also uses a CNN for fruit detection in 2D images. In this case, the neural network architecture was the Mask RCNN, which not only provides object detection bounding boxes, but also gives the corresponding segmentation masks. Results showed an F1-score of 0.858, similar to the one reported in P6 with RGB images (F1-score of 0.867). These results are comparable with other state-of-the-art works based on neural networks, which reported F1-score values between 0.73 and 0.97 (Koirala et al., 2019). The comparison between the 2D fruit detection tests presented in P6 and P7 suggests that the methodology of P7 could be further improved by using RGB-S-D multimodal images. That could be done by using an RGB-D_{ToF} sensor. However, the current RGB-D sensors provide data with lower resolution than high definition (HD) colour cameras, which could affect the precision and resolution of the SfM point clouds generated in P7. Additionally, a disadvantage of using RGB-D_{ToF} is that the working conditions are restricted to low illuminance levels. Another approach could be the registration of LiDAR point clouds with HD colour images. That would allow the generation of HD multimodal images and their use based on the hypothesis that it would increase the 2D fruit detection rate without penalising the consistency of the SfM point cloud. However, further research should be carried out to test the viability of this approach.

2. 3D fruit detection and yield prediction

Many fruit detection works from the state-of-the-art are based on 2D sensors and do not confront the 3D location problem (Gongal et al., 2015; Koirala et al., 2019). A relevant contribution of the present thesis is that it combines high performance object detection algorithms with 3D sensor data, allowing the location of detected fruits in the 3D space (3D fruit detection). Knowing the location of the fruits provides valuable information to the farmer for orchard management based on in-field variability, as well as to plan and optimize the harvesting campaign (Bargoti and Underwood, 2017b). Additionally, the use of 3D sensors allows the measurement of geometrical parameters of the canopy at

the same time, as well as the determination of the relationship between these parameters and yield production and the management strategies of irrigation, fertilization, thinning, and pruning, among others (Escolà et al., 2017; Kühn et al., 2003; Martin-Gorriz et al., 2014).

Chapter IV (P4) presented a proof of concept of using LiDAR sensors for 3D fruit detection. The algorithm was based on a reflectance thresholding followed by a decision tree (*D.Tree*). This algorithm was tested in P4 using 3 randomly selected trees. However, the results reported in Table 1 were obtained with the LFuji-air dataset (P1), allowing a comparison with other 3D fruit detection results presented in the table. In **Chapter V** (P5) the algorithm was enhanced by replacing the *D.Tree* with an SVM. The results did not present a significant improvement in terms of F1-score (from 0.783 to 0.784). However, the fact that SVM's are automatically trained based on the features that characterize apples and the reduction of manually set parameters supposed an advance in the algorithm.

Table 2 summarises the strengths and weaknesses of the tested methodologies. An advantage of using LiDAR with respect to other sensors such as depth cameras is that the LiDAR measurements are not affected by the lighting conditions. Nevertheless, the fruit detection performance continues to be affected by the number of fruits occluded by other vegetative organs such as trunks and leaves. Some authors have proposed the use of multi-view sensing to reduce the number of fruit occlusions (Hemming et al., 2014). In this thesis, P5 tackled the occlusions issue by moving tree foliage with a forced air flow and by using multi-view sensing. The results showed that, by combining different air flow conditions, $H_{1,(n+af)}$, the F1-score increased by 1.5%, from 0.784 to 0.799, similar to the multi-view approach $H_{(1,2),n}$, which presented an improvement of 1.8% in F1-score (Table 1).

The use of CNN has demonstrated remarkable progress in 2D image object detection (Koirala et al., 2019). However, CNNs present some limitations when dealing with unstructured data such as LiDAR point clouds (Qi et al., 2017). **Chapter VII** (P7) proposed the combination of Mask R-CNN and SfM photogrammetry to detect fruits in 2D images and locate them in the 3D space. The results outperformed other 3D fruit

detection methodologies tested in this thesis, presenting an F1-score of 0.881 with respect to the total amount of fruits in the 11 tested trees. Comparing 2D results obtained with Mask RCNN and 3D detections obtained after projecting 2D detections into the 3D space, the F1-score metric showed an increase of more than 2%, from 0.858 (2D detections) to 0.881 (3D detections) (Table 1). From that, it is concluded that, due to the multi-view approach on which SfM is based, this technique helped to increase the detection rate, reducing the number of false positives and preventing the double counting of fruits that appeared in two different images.

Table 2. Advantages and disadvantages of the developed/tested methodologies.

Article	Methodology	Data	Advantages	Disadvantages
P4 / P5	MTLS + Th. + SVM	Reflectance	- Not affected by lighting conditions	- Fruit occlusions - Expensive equipment
P5	MTLS + air/Multi-view	Reflectance	- Not affected by lighting conditions - Reduced number of fruit occlusions	- Expensive equipment
P6	RGB-D-S + FasterRCNN	Reflectance + colour	- High detection rates - High inference speed - Low cost	- Fruit occlusions - Limited to low lighting levels
P7	MaskRCNN + SfM	Colour	- High detection rates - Low number of false positives - Reduced number of fruit occlusions - Low cost - High 3D data precision	- High computational time

Although the combination of Mask RCNN with SfM (P7) presented higher fruit detection rates than the use of LiDAR (P4 and P5), LiDAR detections showed a stronger correlation with the actual number of fruits per tree, reporting a coefficient of determination (R^2) of 0.87 and a root mean square error (RMSE) of 5.7%, while the SfM approach showed an R^2 of 0.80 and an RMSE of 8.2%. These prediction errors are comparable with other state-of-the-art yield prediction methods, such as those presented in Linker (2018, 2017), Payne et al. (2014), and Zhou et al. (2012). From the perspective of a qualitative assessment of 3D data, the point cloud obtained with SfM presented less noisy points and a higher precision compared to RGB-D and LiDAR

point clouds. This suggests that, in future works, the methodology presented in P7 could potentially be used for fruit size measuring, which, combined with the good correlation between the number of fruit detections and the number of total fruits in the tree, would allow the computation of fruit load in weight.

References

Bargoti, S., Underwood, J.P., 2017. Image segmentation for fruit detection and yield estimation in apple orchards. *J. F. Robot.* 34, 1039–1060. doi:10.1002/rob.21699

Escolà, A., Martinez-Casasnovas, J.A., Rufat, J., Arno, J., Arbones, A., Sebe, F., Pascual, M., Gregorio, E., Rosell-Polo, J.R., 2017. Mobile terrestrial laser scanner applications in precision fruticulture/horticulture and tools to extract information from canopy point clouds. *Precis. Agric.* 18, 111–132. doi:10.1007/s11119-016-9474-5

Gan, H., Lee, W.S., Alchanatis, V., Ehsani, R., Schueller, J.K., 2018. Immature green citrus fruit detection using color and thermal images. *Comput. Electron. Agric.* 152, 117–125. doi:10.1016/j.compag.2018.07.011

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

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

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

Kühn, B.F., Pedersen, H.L., Andersen, T.T., 2003. Evaluation of 14 old unsprayed apple varieties. *Biol. Agric. Hortic.* 20, 301–310. doi:10.1080/01448765.2003.9754975

Linker, R., 2018. Machine learning based analysis of night-time images for yield prediction in apple orchard. *Biosyst. Eng.* 167, 114–125. doi:10.1016/j.biosystemseng.2018.01.003

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

Martin-Gorriz, B., Castillo, I.P., Torregrosa, A., 2014. Effect of mechanical pruning on the yield and quality of 'Fortune' mandarins. *Spanish J. Agric. Res.* 12, 952–959. doi:http://dx.doi.org/10.5424/sjar/2014124-5795

Payne, A., Walsh, K., Subedi, P., Jarvis, D., 2014. Estimating mango crop yield using image analysis using fruit at "stone hardening" stage and night time imaging. *Comput. Electron. Agric.* 100, 160–167. doi:10.1016/j.compag.2013.11.011

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017. PointNet: Deep learning on point sets for 3D classification and segmentation, in: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. doi:10.1109/CVPR.2017.16

Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T., McCool, C., 2016. DeepFruits: A Fruit Detection System Using Deep Neural Networks. *Sensors* 16, 1222. doi:10.3390/s16081222

Zhou, R., Damerow, L., Sun, Y., Blanke, M.M., 2012. Using colour features of cv. “Gala” apple fruits in an orchard in image processing to predict yield. *Precis. Agric.* 13, 568–580. doi:10.1007/s11119-012-9269-2

Chapter IX. Conclusions

This PhD thesis has contributed to the development of new methodologies for fruit detection based on the combination of different photon-based sensors (LiDAR, RGB and RGB-D) and computer vision techniques.

First, an MTLS was used to analyse the reflectance of apples, trunks, branches and leaves. This analysis showed that apples present higher reflectance values than other tree elements at the 905nm laser wavelength, concluding that reflectance is a valuable feature for fruit detection (**Chapter IV**). On this basis, an algorithm based on reflectance and geometric features was developed to detect fruits in LiDAR point clouds (**Chapter IV** and **V**). The algorithm consisted of four different steps: (1) reflectance thresholding, (2) connected component labelling, (3) identification and splitting of cluster points with more than one apple, and (4) false positive reduction. Three different classification methods were tested in steps (3) and (4), including template matching, decision tree and SVM. From that, it was concluded that the best fruit detection performance can be achieved either with decision trees or with SVM, presenting F1-scores of 0.783 (decision trees) and 0.784 (SVM), respectively. However, the use of SVM presented the additional advantages that the algorithms was automatically trained and the reduced number of manually set parameters.

A significant advantage of using LiDAR sensors with respect to passive sensors such as colour cameras was that the measurements were not affected by illumination conditions. However, the performance of the system was still being affected by extrinsic factors that do not depend on the sensor, such as the number of occluded fruits. To deal with this issue, two different approaches were tested: forced air flow and multi-view sensing (**Chapter V**). From that, it was concluded that combining data acquired with and without forced air flow conditions and the multi-view approach are good options to improve fruit detectability, reporting an increase in F1-score of more than 1.5% in fruit detection. However, these approaches showed no advantage when using the MTLS system for yield prediction.

Similarly to LiDAR sensors, RGB-D cameras based on the ToF principle provide the amount of light backscattered by the scene, which can be related to the reflectance after range correction and sensor calibration. Since apples showed a higher IR reflectance than other tree elements, **Chapter VI** analysed the usefulness of using the backscattered light from RGB-D sensors besides the colour and depth images. To do so, first, the backscattered IR signal was range corrected. Then, a registration of different data was carried out, obtaining images with 3 modalities: colour, depth and range-corrected intensity. Results showed an improvement of more than 3% in F1-score when all modalities were used, with the conclusion that the use of range-corrected intensity from RGB-D sensors helps to increase the percentage of fruits detected. An additional advantage of using RGB-D sensors with respect to colour cameras, is that they can 3D locate the detected fruits by using depth information, while the main disadvantage is that the depth measurement performance decreases under direct sunlight.

To use high performance deep neural networks for object detection in 2D images, without losing 3D spatial information, and to reduce the number of fruit occlusions using a multi-view approach, **Chapter VII** proposed the combination of instance segmentation neural networks and SfM. Results outperformed other methodologies tested in the present thesis, presenting an F1-score of 0.881 in 3D fruit location. Due to the multi-view approach on which SfM is based, this methodology showed a small number of fruit occlusions, reporting a detection rate of 99.1%. However, the algorithm did not succeed in detaching some groups of apples that were detected in a single detection, which responds to the decrease in the recall metric to 0.906. An additional advantage of using SfM was that it helped to reduce the number of false positives, producing an increase in the precision metric from 0.762 (2D image detections) to 0.857 (3D detections). The main disadvantage of this methodology was the computational time required to generate 3D models with SfM, which does not allow real-time data processing. However, the evolution of computing hardware and the development of efficient algorithms may overcome this limitation in the future.

Considering all the conclusions that have been made, future works should include: (1) an analysis of fruit reflectance under different laser wavelengths; (2) the analysis of fruit occlusions in different crop training systems; (3) the projection of 2D fruit detections obtained from RGB-D sensors onto the 3D space using the depth channel data; (4) the extension of this research to other fruit varieties, species and maturity stages; and (5) and the development of a methodology to measure fruit size.

Chapter X. List of contributions

1. Journal papers included in the thesis

- **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., 2019. Fruit detection in an apple orchard using a mobile terrestrial laser scanner. *Biosystems Engineering* 187 (2019), 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., 2019. Multi-modal deep learning for Fuji apple detection using RGB-D cameras and their radiometric capabilities. *Computers and Electronics in Agriculture* 162 (2019), 689–698. doi:10.1016/j.compag.2019.05.016
- **Gené-Mola, J.**, Gregorio, E., Auat Cheein, F., Guevara, J., Llorens, J., Sanz-Cortiellaa, R., Escolà, A., Rosell-Polo, J.R., 2019. Fruit detection, yield prediction and canopy geometric characterization using LiDAR with forced air flow. *Computers and Electronics in Agriculture,* 168 (2020), 105121. doi:10.1016/j.compag.2019.105121
- **Gené-Mola, J.**, Sanz-Cortiella, R., Rosell-Polo, J.R., Morros, J.-R., Ruiz-Hidalgo, J., Vilaplana, V., Gregorio, E., 2020. Fruit detection and 3D location using instance segmentation neural networks and structure-from-motion photogrammetry. *Computers and Electronics in Agriculture* 169 (2020), 105165. doi:10.1016/j.compag.2019.105165
- **Gené-Mola, J.**, Vilaplana, V., Rosell-Polo, J.R., Morros, J.-R., Ruiz-Hidalgo, J., Gregorio, E., 2019. KFuji RGB-DS database: Fuji apple multi-modal images for fruit detection with color, depth and range-corrected IR data. *Data in Brief* 25 (2019), 104289. doi:10.1016/j.dib.2019.104289
- **Gené-Mola, J.**, Gregorio, E., Auat Cheein, F., Guevara, J., Llorens, J., Sanz-Cortiellaa, R., Escolà, A., Rosell-Polo, J.R.. LFuji-air dataset: annotated 3D LiDAR point clouds of Fuji apple trees for fruit detection scanned under different forced air flow conditions. *Data in Brief* (Submitted).
- **Gené-Mola, J.**, Sanz-Cortiella, R., Rosell-Polo, J.R., Morros, J.-R., Ruiz-Hidalgo, J., Vilaplana, V., Gregorio, E.. Fuji-SfM dataset: a collection of annotated images and point clouds for Fuji apple detection and location using structure-from-motion photogrammetry. *Data in Brief* (Submitted).

2. Other journal contributions

- Rosell-Polo, J.R., Gregorio, E., **Gené-Mola, J.**, Llorens, J., Torrent, X., Arno, J., Escola, A., 2017. Kinect v2 Sensor-based Mobile Terrestrial Laser Scanner for Agricultural Outdoor Applications. *IEEE/ASME Transactions on Mechatronics* 18(2), 145-151. doi:10.1109/TMECH.2017.2663436
- Gregorio, E., **Gené-Mola, J.**, Sanz, R., Rocadenbosch, F., Chueca, P., Arnó, J., Solanelles, F., Rosell-Polo, J.R., 2018. Polarization Lidar Detection of Agricultural Aerosol Emissions. *Journal of Sensors* 2018. doi:10.1155/2018/1864106
- Guevara, J., Auat, F., **Gené-Mola, J.**, Rosell-Polo, J.R., Gregorio, E.. Analysing and overcoming the effects of GNSS error on LiDAR based orchard parameters estimation. *Computers and Electronics in Agriculture* (Submitted).
- Guevara, J., **Gené-Mola, J.**, Gregorio, E., Torres-Torriti, M., Reina, G., Auat, F.. Evaluating the performance of 3D scan registration as localization system in urban and agricultural scenarios: an insightful trade-off for autonomous navigation. *Measurement* (Submitted).

3. Conference contributions

- **Gené-Mola, J.**, 2018. Fruit Detection and Localization from RGB-D Sensors, in: *Annual Catalan Meeting on Computer Vision* 2018. Barcelona (ES)
- **Gené-Mola, J.**, Gregorio, E., Guevara, J., Auat, F., Escolà, A., Morros, J.-R., Rosell-Polo, J.R., 2018. Fruit Detection Using Mobile Terrestrial Laser Scanning, in: *EurAgEng 2018 Conference*. Wageningen (NL)
- **Gené-Mola, J.**, Gregorio, E., Llorens, J., Sanz-Cortiella, R, Escolà, A., Rosell-Polo, J.R., 2019. Yield prediction using mobile terrestrial laser scanning. *Poster Proceedings of the 12th European Conference on Precision Agriculture - ECPA 2019*, July 8-11, Montpellier, France (FRA) ISBN 978-2-900792-49-0
- **Gené-Mola, J.**, Vilaplana, V., Rosell-Polo, J.R., Morros, J.-R., Ruiz-Hidalgo, J., Gregorio, E., 2019. Uso de redes neuronales convolucionales para la detección remota de frutos con cámaras RGB-D. *Proceedings of the 10th Iberian Agroengineering Congress.* Huesca (ES). pp. 1081-1087. ISBN: 978-84-16723-79-9. doi:10.26754/c_agroing.2019.com.3425
- Gregorio, E., **Gené-Mola, J.**, 2019. Mecatrónica: adquisición de competencias transversales y especificas en el marco de una asignatura multidisciplinar. *III congreso virtual internacional y V congreso virtual iberoamericano sobre recursos educativos innovadores.*