Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/issn/15375110

Research Paper

Faster R–CNN–based apple detection in dense-foliage fruiting-wall trees using RGB and depth features for robotic harvesting

Longsheng Fu [a,b,c,d,*], *Yaqoob Majeed* [d], *Xin Zhang* [d], *Manoj Karkee* [d], *Qin Zhang* [d]

[a] *College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling, Shaanxi 712100, China*
[b] *Key Laboratory of Agricultural Internet of Things, Ministry of Agriculture and Rural Affairs, Yangling, Shaanxi 712100, China*
[c] *Shaanxi Key Laboratory of Agricultural Information Perception and Intelligent Service, Yangling, Shaanxi 712100, China*
[d] *Centre for Precision and Automated Agricultural Systems, Washington State University, Prosser, WA, 99350, USA*

ARTICLE INFO

Article history:
Received 1 November 2019
Received in revised form
20 May 2020
Accepted 7 July 2020
Published online 25 July 2020

Keywords:
RGB-D camera
Depth filter
ZFNet
VGG16
Robotic harvesting

Apples in modern orchards with vertical-fruiting-wall trees are comparatively easier to harvest and specifically suitable for robotic picking, where accurate apple detection and obstacle-free access are fundamentally important. However, field images have complex backgrounds because of the presence of nontarget trees and fruit in adjacent rows. An outdoor machine vision system was developed with a low-cost Kinect V2 sensor to improve the accuracy of apple detection by filtering the background objects using depth features. A total of 800 set images were acquired in a commercial fruiting-wall Scifresh apple orchard with dense-foliage canopy. Images were collected in both daytime and nighttime with artificial light. The sensor was kept at 0.5 m to the tree canopies. A depth threshold of 1.2 m was used to remove background. Two Faster R–CNN based architectures ZFNet and VGG16 were employed to detect the Original-RGB and the Foreground-RGB images. Results showed that the highest average precision (*AP*) of 0.893 was achieved for the Foreground-RGB images with VGG16, which cost 0.181 s on average to process a 1920 × 1080 image. *AP* values for the Foreground-RGB images with ZFNet and VGG16 were both higher than those of the Original-RGB images. The results indicated that the use of a depth filter to remove background trees improved fruit detection accuracy by 2.5% and that only a minimal difference was found in processing speed between two image datasets. The proposed technique and results are expected to be applicable for robotic harvesting on fruiting-wall apple orchards.

* *Corresponding author*. College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling, Shaanxi 712100, China.
E-mail addresses: fulsh@nwafu.edu.cn, longsheng.fu@wsu.edu (L. Fu).
https://doi.org/10.1016/j.biosystemseng.2020.07.007
1537-5110/

Nomenclature

3D	3-dimensional
AP	average precision
CNN	convolutional neural networks
Faster R–CNN	Faster Region-based Convolutional Neural Network
FN	false negatives, means the number of missed apple objects
FoV	field of view
FP	false positives, means the number of falsely detected apple objects
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
IoU	Intersection over Union
k	number of anchor boxes in the Faster R–CNN
LED	light-emitting diode
NIR	near-infrared
P	Precision
R	Recall
RGB	a colour space, R for red colour component, G for green colour component, and B for blue colour component
RGB-D	Red, Green, Blue, and Depth
SNAP	simple, narrow, accessible, and productive
ToF	Time-of-Flight
TP	true positives, means the number of correctly detected apple objects
U.S.	United States
V2	version 2
V3	version 3
VGG16	Visual Geometry Group with 16 layers
YOLO	You Look Only Once
ZFNet	Zeller Fergus Net

1. Introduction

Fresh market apples are one of the premium fruit crops in the United States and comprise a leading agricultural commodity in Washington State. Across the United States, farmers grow an estimated wholesale value of approximately $4 billion of apples per year (U.S. Apple Association, 2016). At present, fresh market apples are harvested manually worldwide. In the Pacific Northwest region of the United States, labour cost for fresh market apple harvesting accounted for 20%–30% of all on-farm production cost (He, Zhang, Ye, Karkee, & Zhang, 2019). he intense demand for seasonal labours created a significant risk of having an insufficient supply of farm labour to complete time-sensitive agricultural tasks (Zhang et al., 2018). Therefore, the industry is seeking for technological innovations to reduce the dependence on manual labour and thus assist apple growers in maintaining a competitive position in the global marketplace.

Mechanical harvesting is a widely investigated method for tree fruit crops. Apples in modern orchards with a simple, narrow, accessible, and productive (SNAP) system are comparatively easier to harvest than those in traditional orchards, both manually and mechanically. This type of tree architecture is specifically suitable for mechanical harvesting. One of the solutions is bulk harvesting with shake-and-catch systems that apply vibrations to the tree trunk or branch to harvest the apples (He, Fu, Karkee, & Zhang, 2017, He, Fu, Sun, Karkee, & Zhang, 2017, Peng et al., 2017; Zhang et al., 2018), but this method often results in a damage level that is unacceptable for many commercially important varieties grown for fresh market. The alternative method is selective picking with robotic technologies where accurate apple detection, localisation, and obstacle-free access are fundamentally important (Silwal et al., 2017).

The machine vision system critically influences the efficiency of robotic picking, and researchers have implemented machine vision systems in various ways to automate fruit detection, as detailed in the several review studies (Bac, van Henten, Hemming, & Edan, 2014; Bechar & Vigneault, 2016; Gongal, Amatya, Karkee, Zhang, & Lewis, 2015; Lee et al., 2010; Vougioukas, 2019; Zujevs, Osadcuks, & Ahrendt, 2015). In general, the studies reported achieved fruit detection accuracy ranging from 80% to 95% and stated variable light conditions, fruit clusters, occlusions, and complex backgrounds as the most significant challenges for accurate fruit identification in the orchard environment.

Image segmentation, a primary step in fruit identification, is greatly affected by the background under field conditions (Fu, Tola, Al-Mallahi, Li, & Cui, 2019). Images acquired in the SNAP system often have more complex backgrounds because of the presence of nontarget trees and fruit, potentially reducing the accuracy of fruit detection. Silwal, Gongal, and Karkee (2014) developed an over-the-row platform with a tunnel structure and artificial light to minimise the effects of the nontarget trees and apples in the SNAP system. They reported a 78.9% accuracy for red-apple identification using iterative circular Hough transform and blob analysis on 60 images with 978 apples. This platform was also used to measure apple fruit size in a tree canopy using a colour camera and a Time-of-Flight (ToF) light-based 3-dimensional (3D) camera and reached an accuracy of 84.8% using a dataset of 150 apples that were selected from images containing 25 random trees (Gongal, Karkee, & Amatya, 2018).

Alternatively, the background objects (trees and apples from nontarget orchard rows) in the SNAP system can be removed using depth features if an RGB-D (Red, Green, Blue, and Depth) camera was adopted. Commercially available RGB-D cameras use structured pattern illumination in the near-infrared (NIR) spectrum for the acquisition of depth images with high spatial resolution. Those economically affordable devices combine the advantages of spectral-based and range-based sensors, providing information on colour, distance, and shape. As such, the RGB-D cameras offer additional information that can be used to address the challenges in accurate fruit detection (Gan, Lee, Alchanatis, Ehsani, & Schueller, 2018; Gené-Mola, Vilaplana, et al., 2019; Li, Kou, Cheng, Zheng, & Wang, 2017; Lin, Tang, Zou, Xiong, & Fang, 2019; Lin, Tang, Zou, Xiong, & Li, 2019; Liu et al., 2020; Méndez Perez, Cheein, & Rosell-Polo, 2017; Nguyen et al., 2016; Tao & Zhou, 2017; Yu, Zhang, Yang, & Zhang, 2019). However, to the best of our knowledge, no previous work of object detection has used the depth information to remove the nontarget background objects in the SNAP tree system.

Regarding image processing techniques used for fruit detection, most of the previous work used traditional handcrafted features to encode the data acquired with different sensors and to estimate the fruit location. More recently, the introduction of deep learning has led to remarkable progress in object recognition. The Faster Region-based Convolutional Neural Network (Faster R–CNN) is one of the most commonly used techniques for small object recognition and thus for fruit detection (Fu et al., 2018; Koirala, Walsh, Wang, & Mccarthy, 2019a, 2019b; Liu et al., 2020; Sa et al., 2016; Stein, Bargoti, & Underwood, 2016; Zhang et al., 2019). Häni, Roy, and Isler (2019) employed a Faster R–CNN–based model for apple detection and achieved a success rate of 90.8% using 103 images with a pixel resolution of 1920 × 1080. Recently, a number of studies on the detection of apples using deep learning techniques have been reported by researchers in Spain (Gené-Mola, Gregorio, et al., 2019; Gené-Mola, Vilaplana, et al., 2019; Gené-Mola, Gregorio, et al., 2020; Gené-Mola, Sanz-Cortiella et al., 2020). Specifically, Gené-Mola, Vilaplana, et al. (2019) used the Faster R–CNN to detect Fuji apples and obtained an F1-score of 0.898 and an average precision (*AP*) of 94.8% on 967 images with 12,839 fruits, demonstrating good detection accuracy. However, those studies were conducted in a tall spindle architecture, which has relatively less dense foliage than the formal training, fruiting-wall architecture of the Scifresh variety being studied in this work.

However, the main drawback of using convolutional neural networks (CNNs) is that they require a large amount of manually labelled data. Lack of substantial datasets is a barrier for exploring emerging sensors that could be useful for fruit detection (Hameed, Chai, & Rassau, 2018; Kamilaris & Prenafeta-Boldú, 2018). Therefore, this paper provides a Scifresh dataset with multimodal images from RGB-D sensors with colour and depth features and corresponding labels with the ground truth-apple locations. The Scifresh has one of the denser-foliage apple canopies because of the high-vigour rootstock being applied, and automatically detecting its fruit is challenging (Gongal et al., 2015).

In this study, an outdoor machine vision system was developed with a low-cost RGB-D camera to improve the accuracy of Scifresh apple detection in the fruiting-wall trees by filtering out the background objects using the depth features. The two most widely used Faster R–CNN models, ZFNet (Zeller Fergus Net) and VGG16 (Visual Geometry Group with 16 layers) architectures, were adapted and implemented. The outline of this paper is as follows: In Section 2, the materials and methods used in the study including the sensor system and detection algorithm, are described. In Section 3, the results are presented and discussed. Lastly, in Section 4, conclusions are drawn, and suggestions are made for further studies.

2. Materials and methods

This study focused on detecting apples in vertical-fruiting-wall tree canopies, which is one of the SNAP systems commonly planted in Washington orchards. Image data were collected during two harvesting seasons (2017 and 2018) in a commercial orchard from a fixed distance of 0.5 m using an RGB-D camera. A series of sensor data were acquired, including aligned RGB and depth information. Trees and apples from nontarget rows in the RGB images (i.e. Original-RGB images) were removed using the depth feature with a distance threshold of 1.2 m (i.e. Foreground-RGB images). Both original-RGB and Foreground-RGB image datasets were selected as the inputs to Faster R–CNN models (i.e. ZFNet and VGG16 architectures) for training and testing. The performances of the two models with two different types of input images were evaluated and compared using the evaluation measures of precision (*P*), recall (*R*), and *AP*. Detailed explanations of these techniques are provided in the following sections.

2.1. *Image acquisition*

Complex background and changes of illumination in the natural orchard environment always make achieving the desired fruit detection accuracy difficult for conventional image processing techniques. However, with depth features, it is possible to estimate the specific space location of fruits and remove most background interferences by limiting the detection range of depth. Therefore, this study used a Microsoft Kinect V2 camera (Microsoft Inc., Redmond, WA) to build an image acquisition system and capture RGB and depth images. The sensor uses the principle of Time-of-Flight (ToF) to acquire depth features with an operating range between 0.5 and 4.5 m, with the average error in depth measurement ranging from 2 to 5 mm within the operating range of 3.5 m in outdoor applications (Vit & Shani, 2018). Kinect V2 sensor was used in this study because it is stable and cost-effective. In addition, using this sensor offers a relatively shorter development time because there is a good research foundation and support (Gené-Mola, Vilaplana, et al., 2019; Jiang, Li, & Paterson, 2016; Majeed et al., 2018; Lin, Tang, Zou, Xiong, & Fang, 2019; Liu et al., 2020).

In this study, all the images were acquired in a commercial apple orchard, as shown in Fig. 1a, where the row spacing was approximately 2.6 m. The Scifresh apple trees were around 4.0 m and trained in seven layers where fruits were mostly distributed on the branches at each layer, as shown in Fig. 1b. Figure 1c showed a self-built platform for image acquisition. Four 850 lumens LED (light-emitting diode) lights (Trilliant 36 Light Emitting Diode Grote, Madison, Indiana) were installed in the platform to create a controlled and uniform lighting environment for nighttime imaging. The Kinect V2 camera was mounted at the centre of four LED lights installed on the platform. The camera was installed such that a vertical distance of approximately 1.3 m from the ground and a range of 0.5 m to the centre of target canopies was maintained. This configuration allowed the capture of apple tree images in the 3rd to 4th layers of the canopies for fruit detection. Lighting was not controlled in this study because of more the variation in the images for deep learning networks, the better generalization capabilities can be achieved with higher accuracy and robustness. And change of sun angle naturally caused the variation in the images will help to develop more robust detection system.

Kinect V2 can acquire RGB images, depth images, and point data using its RGB and depth sensors. The field of view (FoV) of the depth (70.7° × 60.1°) and RGB (84.1° × 53.8°) sensors are different, posing a challenge for co-registering two types of

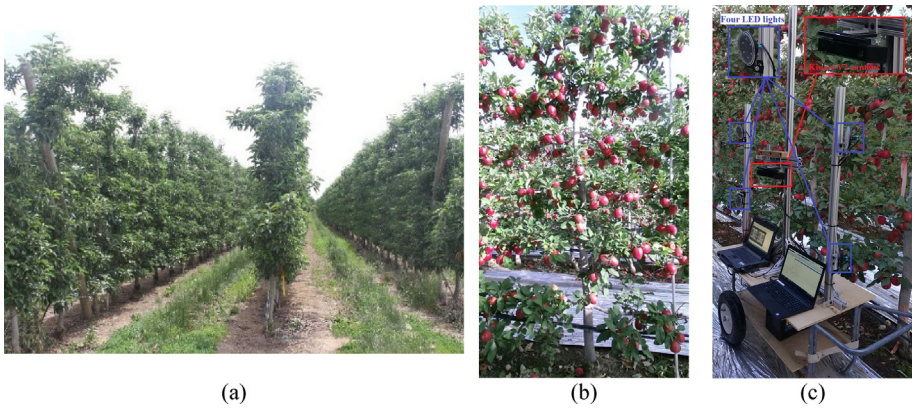

Fig. 1 – (a) Experimental orchard used in this work; (b) Trees of Scifresh apple variety (close to harvest) trained to vertical-fruiting-wall structure; and (c) Self-built platform with mounted Kinect V2 camera and LED lights for image acquisition.

information. However, in the point cloud data, the depth and RGB features were already mapped together by aligning the RGB images (1920 × 1080 pixels) with the depth images (512 × 424 pixels) or reversed, making removing background from the RGB images using the depth features conveniently. As a result, point cloud data with the resolution of 1920 × 1080 pixels was acquired where the direction of the image coordinate system was set as 'colourCentric' to align the depth image with the RGB image. Specific software written in MATLAB 2018a was developed to collect and save data automatically. In total, 392 and 408 pairs of RGB images and point cloud data were acquired from randomly selected apple trees in 2017 and 2018 harvest seasons, respectively.

Kinect V2 is somewhat sensitive to ambient infrared radiation from the sunlight although it is based on ToF technology. Various studies have been conducted using the Kinect V2 sensor in agricultural field conditions (Liu et al., 2020; Majeed et al., 2020; Nguyen et al., 2016; Tao & Zhou, 2017). We found that significant impact on the performance of this camera is caused when direct sunlight is orthogonal to the direction of viewing in the tree canopies trained to modern, fruiting-wall architectures. The test field consisted of 11-year-old denser-foliage Scifresh apple trees with an average height of 4.0 m. High apple tree canopies with denser foliage blocked the direct sunlight substantially, hence helping the acquisition provide more accurate and reliable RGB-D information. Moreover, images were collected during 7 AM–10 AM and 3 PM–8 PM to avoid the sun being directly above (around noon time) the canopies/sensor. The image acquisition was completed in different environmental conditions including sunny and cloudy days and night with artificial light.

Figure 2a shows an example of an RGB image acquired using Kinect V2. Trees and apples from nontarget rows in the RGB images can be removed using the depth features because the target trees are much closer to the sensor than those in the background. Because a part of depth information is not registered with RGB information in the point cloud, the aligned depth image extracted from the point cloud has no information at the left and right edges of the images, as shown in Fig. 2b. A distance threshold of 1.2 m was then used to remove any objects that were beyond the threshold depth. Figure 2c shows an example image after the background removal where the target tree and apples can be clearly observed. The information areas of both the left and right sides of the Foreground-RGB image were lost because of the different FoV of the colour and depth sensors. In addition, very few pixels in the target area turned to the black colour as the background because of the mismatched depth information in the point cloud, as shown in Fig. 2c.

Ground truth apple targets were manually annotated in the Foreground-RGB images with the resolution of 1920 × 1080 pixels using rectangular annotations (Fig. 3a) and then mapped to the Original-RGB images with the resolution of 1920 × 1080 pixels (Fig. 3b). Finally, a total of 29,839 fruits were labelled in the full set of Foreground-RGB images. The labelled Foreground-RGB images and corresponding Original-RGB images of apples were divided into training (70%, 560 images), validation (15%, 120 images), and test (15%, 120 images) groups. The training images were randomly obtained from the independent and uniform sampling of the whole dataset. All images were mutually exclusive, ensuring the reliability of the later evaluation standards.

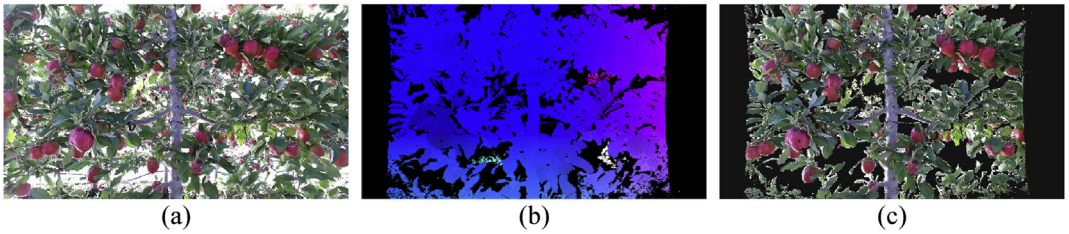

Fig. 2 – (a) Original-RGB image in the sunny day; (b) Aligned depth image extracted from the point cloud; and (c) Corresponding Foreground-RGB image after background removal using the depth threshold of 1.2 m.

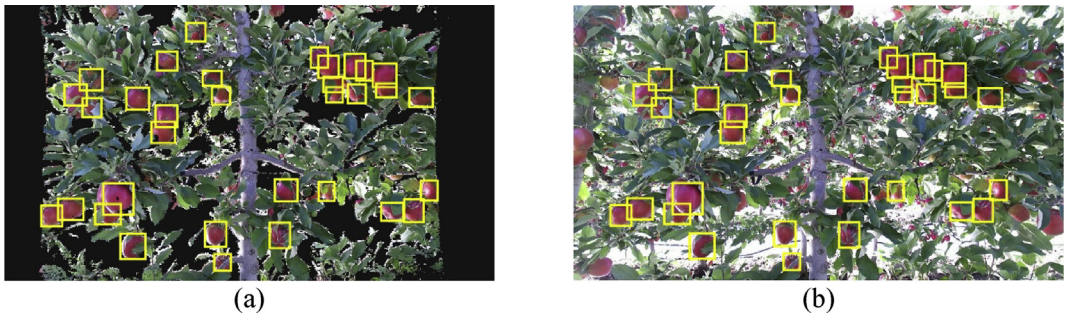

Fig. 3 – (a) Ground truth apple targets were manually annotated in the Foreground-RGB image using the rectangular annotations and (b) then mapped to the Original-RGB image.

2.2. *Faster R–CNN*

The Faster R–CNN method merges region proposals and objects classification and detection into one unified deep object detection network (Ren, He, Girshick, & Sun, 2017). Two networks (Region Proposal Network and Faster R–CNN) are concatenated as one network that can be trained and tested through an end-to-end process (Bargoti & Underwood, 2017). The Faster R–CNN method can depend on different CNN architectures such as ZFNet (Zeiler & Fergus, 2014) and VGG16 (Simonyan & Zisserman, 2015),as shown in Table 1. To verify our method is invariant to different object proposal methods, experiments were conducted on the Faster R–CNN based on both the ZFNet and VGG16 architectures.

ZFNet was the winner of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2013. The input to ZFNet was RGB images that were quantised to 224 × 224 pixels. The image was processed by the network resulting in the probability that the image belongs to each category. ZFNet consists of five shareable convolutional layers, max-pooling layers, dropout layers, and three fully connected layers. It uses a 7 × 7 size filter and a decreased stride value in the first layer. The last layer of ZFNet was the softmax layer, and this was used to convert the score into the probability that the image belongs to each category.

VGG16 won second place in the 2014 ILSVRC and performed well in multiple transfer learning tasks. It has 13 shareable convolutional layers and three fully connected layers. The input to the first convolutional layer had a fixed size of 224 × 224 pixels, which is same as ZFNet. In VGG16, the image is passed through a stack of convolutional layers where the filters are used with a very small receptive field of 3 × 3. The convolution stride is fixed to 1 pixel, and the spatial padding of convolutional layer input is such that the spatial resolution was preserved after convolution. Spatial pooling is performed by five max-pooling layers, following some of the convolutional layers. Max-pooling is performed over a 2 × 2 pixels window, with a stride of two. Three fully connected layers follow a stack of convolutional layers, and the final layer is the softmax layer.

Table 1 – Comparison of the VGG16 and ZFNet architectures.

Model	ZFNet	VGG16
Input	224 × 224 pixels	224 × 224 pixels
Convolutional layers	5	13
Max-pooling layers	2	5
Dropout layers	2	2
Fully connected layers	3	3
Softmax layer	1	1

2.3. *Network training*

The training architecture for apple detection is illustrated in Fig. 4 based on Faster R–CNN with ZFNet and VGG16. The Region Proposal Network was implemented as a fully convolutional network that which was optimised through an end-to-end process using backpropagation and stochastic gradient descent. Nonmaximum suppression was applied to the proposal regions to reduce the redundancy. Faster R–CNN framework is capable of multi-class detection, and our work only considered a binary classification problem of apple images acquired in the orchard environment. Therefore, the output layer of the ZFNet and VGG16 were modified to two classes of background and apple regions in this study, and each fruit position was marked with a rectangular box. The fully connected layers more effectively distribute the learned higher-order features to samples (perform high-level reasoning based on leaned features from convolutional layers) (Oquab, Bottou, Laptev, & Sivic, 2014, pp. 1717–1724). Therefore, modifying the fully connected layers of classifying 1000 classes to only 2 classes does not impact the nonlinear mapping of higher-order features.

The random initialisation of weights takes a longer time to converge the model to a stable value or may even fall into a local minimum. Transfer learning is the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned, which can quickly adapt to new tasks in the event of a small dataset (Shin et al., 2016). By using a pretrained model with a large dataset, parameters of underlying structure weight can be shared. In addition, the model can be fine-tuned to overcome the differences between the pretrained and new datasets. In this work, the shared convolutional layers of Faster R–CNN were initialised using a pretrained model for ImageNet classification. All other layers were randomly initialised by drawing weights from a zero-mean Gaussian distribution with a standard deviation of 0.01. A threshold of 70% on Intersection over Union (IoU) was used to decide whether a detected

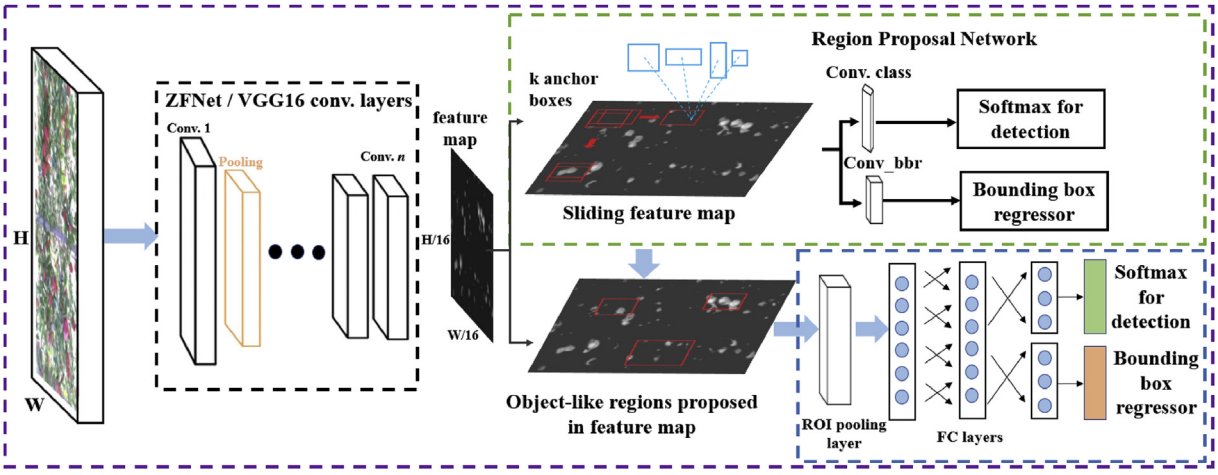

Fig. 4 – Training architecture for apple detection based on Faster R–CNN with ZFNet and VGG16 (*k* is 9 in this study).

instance was true (fruit) or false when comparing the predicted bounding box to ground truth.

The training platform included a computer with Intel Xeon E5-1650 [(3.60 GHz) six-core CPU, a GPU of NVidia TITAN XP 6 GB GPU (3840 CUDA cores) and 16 GB of memory] running on a Windows 7 64-bit system. The software tools included CUDA 8.1, CUDNN 7.5, Python 2.7, and Microsoft Visual Studio 12.0. The experiments were implemented in the Caffe framework (Jia et al., 2014). The Original-RGB images training dataset (560 images) and Foreground-RGB images training dataset (560 images) were used in training the same deep learning networks separately and then validated and tested on corresponding validation and test datasets.

2.4. *Evaluations*

The performances of the proposed methods were evaluated using precision (P), recall (R), *AP*, and detection speed. Among them, the *P* and *R* are defined in Eq. (1) and Eq. (2), respectively.

$$P = TP / (TP + FP) \tag{1}$$

$$R = TP / (TP + FN) \tag{2}$$

where *TP*, *FP*, and *FN* represent the number of correctly detected apple objects (true positives), the number of falsely detected apple objects (false positives), and the number of missed apple objects (false negatives), respectively. *AP* was defined in Eq. (3) as the area under the *P* and *R* curve. It was a standard for measuring the sensitivity of the network to object and an indicator that reflects the global performance of the network.

$$AP = \int_0^1 P_{(R)} dR \tag{3}$$

3. Results and discussion

This section evaluates the proposed fruit detection methodology qualitatively and quantitatively to assess the performance of different Faster R–CNN network architectures on detecting apples in different image datasets.

3.1. *Training assessment and performance of the network*

The number of training images required by deep learning models largely depends on the visual complexity of the images in training and test datasets, the network architecture, image augmentation techniques, and machine learning parameters of the network. Similar to Sa et al. (2016) and Bargoti and Underwood (2017), an experiment was first performed to provide a guideline on the desirable number of images for training deep learning architectures with default network parameters. The training sets of 560 images of two categories (Original-RGB and Foreground-RGB) were randomly selected for subsets of 10, 20, 50, 100, 200, 300, 400, and 560 images, respectively. The process was repeated five times to reflect variations in the dataset, where a subset of images was randomly sampled each time from the training set without replacement.

The *AP* of ZFNet and VGG16 on the validation set plotted against the number of training images from the training set is shown in Fig. 5. Detection performance was improved quickly at the beginning with a small number of training images, reaching approximately 0.6 for apples with only 50 images. There were no significant differences of *AP* in terms of using the Faster R–CNN network or image types when only 10 or 20 images were used. As the number of training images reached 200, performances were all close to convergence, only increasing by 0.02 in the final double increase in the number of training images. Therefore, about 300 images were recommended for training to be used to keep the networks stable.

When the number of training images was over 100, there were observable differences of *AP*s using two types of image datasets on two network architectures, as shown in Fig. 5. Foreground-RGB images on VGG16 achieved the best *AP* from only 50 images used for training, while Original-RGB images on VGG16 obtained the second place from 100 images. Foreground-RGB images on ZFNet yielded an *AP* at least 0.017 better than Original-RGB images on ZFNet from 50 images used for training. Clearly, when the performances were about to converge, VGG16 performed better than ZFNet, and the

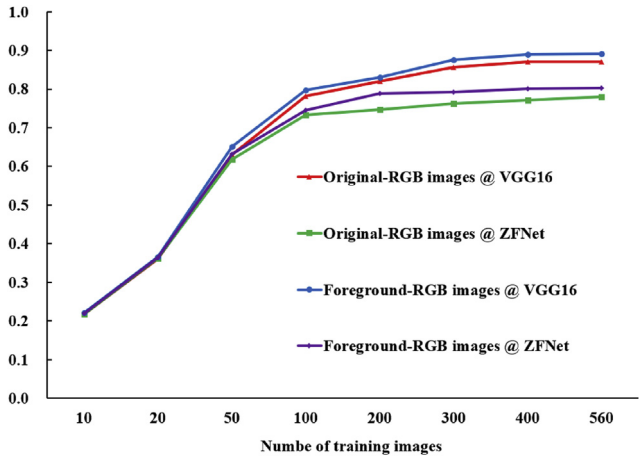

Fig. 5 – Average precision (AP) of ZFNet and VGG16 on the validation set plotted against the number of training images from the training set.

results with Foreground-RGB images were better than the same with Original-RGB images on the same architecture.

Figure 6 shows the Precision-Recall (P-R) curves achieved by ZFNet and VGG16 networks in detecting apples on the test datasets of Original-RGB and Foreground-RGB images. As expected, the *P* values of both models on the Foreground-RGB images were higher those that of the Original-RGB images at the same *R* values. Moreover, the VGG16 architecture obtained a higher *P* on both image types than ZFNet at the same R value.

3.2. *Apple detection on Original-RGB images with faster R–CNN*

For the Original-RGB images, ZFNet and VGG16 achieved *AP*s of 0.787 and 0.871, respectively, as shown in Table 2. The *AP* achieved by VGG16 on the Original-RGB images was 10.7% higher than that achieved by ZFNet on the same images. However, VGG16 took 0.182 s to detect apples on an Original-RGB image with a resolution of 1920 × 1080, which was 1.46 times longer than 0.125 s per image taken by ZFNet.

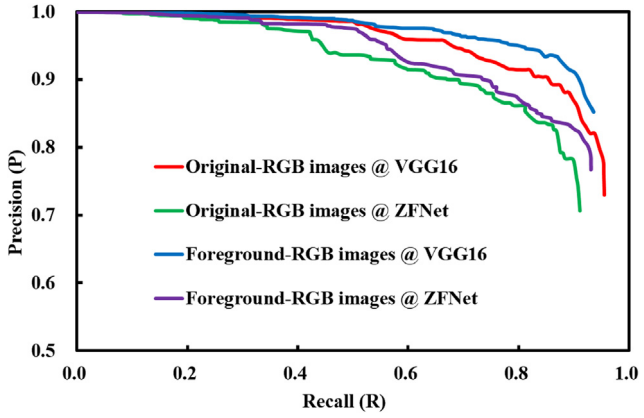

Fig. 6 – Precision-Recall (P–R) curves of ZFNet and VGG16 on the testing set of the Original-RGB and Foreground-RGB images.

Table 2 – Apple-detection results of the Original-RGB and Foreground-RGB images using ZFNet and VGG16.

Image types	AP		Detection speed (second per image)[a]	
	ZFNet	VGG16	ZFNet	VGG16
Original-RGB	0.787	0.871	0.125	0.182
Foreground-RGB	0.805	0.893	0.124	0.181

[a] Detection speed may vary across different hardware settings.

Convolutional layers in both VGG16 and ZFNet are connected in series, where the output of one convolutional layer act as the input to the next layer, allowing the network to extract higher-level features for increased accuracy at the expense of higher trained weights and longer computational time (Tabian, Fu, & Khodaei, 2019). As suggested by past research, a longer computational time with VGG16 was primarily because of the larger number of convolutional layers and parameters than the same with ZFNet that resulted in a large trained weights size of 512 MB, which is more than twice the 225 MB of the ZFNet. The trained weights size in our study are similar to that on mango RGB images detection (A Koirala et al., 2019a; 2019b), which were 533 MB and 230 MB for the VGG16 and ZFNet, respectively.

Although it is difficult to compare methodologies tested with different datasets from different apple varieties, it gives some insights when the results achieved using original RGB images captured from different orchard environments are discussed along with the results achieved in this work. The results from other apple-detection studies using original RGB images were summarised in Table 3.

For the same apple variety, Scifresh planted in the same SNAP canopy trees, Gongal et al. (2016) achieved an accuracy of 78.9% in detecting red apples in 212 images (1697 apples) using an iterative circular Hough transform and blob analysis technique and 1280 × 960 pixel images. The detection speed was not reported. Nguyen et al. (2016) developed an image processing algorithm consisting of distance and colour filtering, clustering segmentation, and cluster separation and tested with 10 Fuji-apple images with 225 fruits. The study reached an accuracy of 81% detection with a computational time of 0.964 s to process a point cloud data of 512 × 424 pixels. Bargoti and Underwood (2017) employed CNN to detect Kanzi and Pink Lady apples in the V-trellis structure and achieved a mean detection probability of 0.85 with an average prediction time of approximately 7 s for 1100 images of 1232 × 1616 pixel resolution. Tian et al. (2019) improved the You Look Only Once (YOLO) V3 for Fuj' apple detection during different growth stages in orchards and obtained an *AP* of 81.7% with an average computational time of 0.304 s for 960 images of 3000 × 3000 pixels resolution. Gené-Mola, Gregorio, et al. (2019) used VGG16 architecture for Fuji apple detection and reported at *AP* of 88.7% with a computational time of 0.074 s for processing a 548 × 373-pixel image. They provided a KFuji RGB-DS dataset (967 images with 12,839 fruits) with corresponding annotations that is publicly available at www.grap.udl.cat/en/publications/datasets.html. All deep-learning-based detection results

Table 3 – Results from other apple-detection studies using original-RGB images.

	Apple variety	Image resolution (pixels)	Number of images	Number of fruits	Main methods	Detection rate (%)	Detection speed (s)
Gongal et al. (2016)	Scifresh	1280 × 960	212	1697	circular Hough transform, blob analysis	78.9	NA[a]
Nguyen et al. (2016)	Fuji	512 × 424	10	225	distance & colour filtering, clustering segmentation, cluster separating	81.0	0.964
Bargoti and Underwood (2017)	Kandi, PinkLady	1232 × 1616	1100	NA	CNN	85.0	~7
Gené-Mola, Gregorio, et al. (2019)	Fuji	548 × 373	967	12,839	VGG16	92.7	0.074
Tian et al. (2019)	Fuji	3000 × 3000	960	NA	YOLOv3-dense	81.7	0.296
This research	Scifresh	1920 × 1080	800	29,839	VGG16	87.1	0.124

[a] NA means data not available.

reported slightly better performance (in terms of detection accuracy and processing time) than the ordinary image processing algorithms. VGG16 in this study for Scifresh did not perform better than achieved previously with Fuji, and this was caused primarily by the difference between foliage densities. Scifresh is one of the densest foliage-apple varieties, whereas Fuji is a mild foliage density variety.

3.3. *Apple detection on Foreground-RGB images with faster R–CNN*

*AP*s for apple detection with Foreground-RGB images achieved by both ZFNet and VGG16 were higher than the Original-RGB images, as shown in Table 1. *AP* for the Foreground-RGB images (0.805) achieved by ZFNet was 2.3% higher than the same with Original-RGB images (0.787), while the same achieved by VGG16 was 2.5%. Because the Scifresh apple cultivar is one of the densest foliage-apple canopies, automated fruit detection for this cultivar is challenging. Therefore, an *AP* of 0.871 achieved by the VGG16 on the Original-RGB images was considered to be high. The percentage of improvements was small (only 2.5%) with Foreground-RGB images over the same with the Original-RGB images. However, this increase suggested a valuable enhancement of the methodology, and the results achieved through incorporating depth-based background removal indicate that it becomes much more challenging to further improve the accuracy as it approaches 1. Similar to Original-RGB images on different architectures, the VGG16 took 0.181 s per image to detect apples on a Foreground-RGB with the resolution of 1920 × 1080 pixels, and this is 1.46 times longer than the ZFNet time (0.124 s per image to detect apples). However, there was only a minimal difference in processing speed between two image datasets on a given network architecture.

Nevertheless, the observation of bounding boxes on the resulting images presents more insights into the results, as shown in Fig. 7. For both network architectures, more apple fruits were detected on the Original-RGB image than the Foreground-RGB image. For example, 37 apples were detected on the Original-RGB image using the VGG16 (Fig. 7a), while only 27 apples were detected on the Foreground-RGB images using the same trained system, as shown in Fig. 7b. The reason is that the Original-RGB image has more information than the Foreground-RGB image, as described in Section 2.1. Substantial areas on the left and right edges of the Foreground-RGB image were lost because of the different FoV values of the colour and depth sensors. Because the ground truth data was labelled on the Foreground-RGB image dataset and then mapped to the Original-RGB image dataset, some detected apples in the Original-RGB image were not counted as the *TP* fruits. Besides, some fruits in the Original-RGB image were removed after the distance thresholding process in the Foreground-RGB image, as the yellow rectangles showed in Fig. 7c and d. Therefore, only the visible apples in the Foreground-RGB image were labelled as the ground truth fruit. In comparison to the ground truth fruit, more target fruit was detected on the Foreground-RGB image than the Original-RGB image, agreeing with the results in Table 2.

In contrast, the methodology of using the depth feature to remove background obtaining the Foreground-RGB image

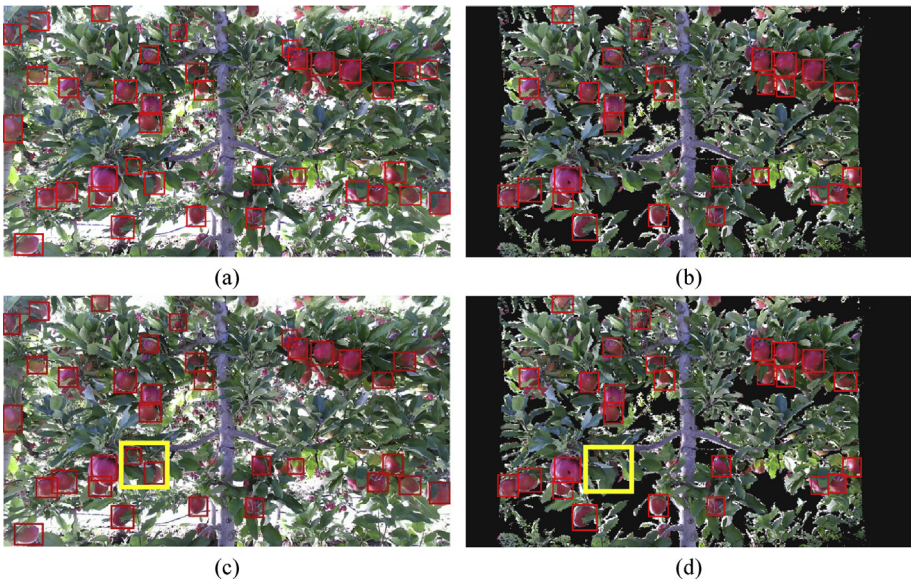

Fig. 7 – Examples of apple detection on Original-RGB image and Foreground-RGB image using the ZFNet and VGG16 trained on Original-RGB image and Foreground-RGB image datasets, respectively. (a) Original-RGB image on VGG16, (b) Foreground-RGB image on VGG16, (c) Original-RGB image on ZFNet, and (d) Foreground-RGB image on ZFNet.

improves the accuracy of fruit detection and thus the yield estimation and robotic harvesting. For the Scifresh apple trees used in this work were trained in the fruiting-wall architectures, fruit detections aiming for yield estimation and robotic harvesting were trained based on images captured from both sides of the canopy using an over-the-row sensing platform (Gongal et al., 2016). They used the location of apples in three-dimensional (3D) space to eliminate duplicate counting of apples that were visible to cameras from both sides of the tree canopy. The error identifying duplicate apples was found to be 21.1%, which was not promising for yield estimation. In our study, the Foreground-RGB image removed some fruits that are partly visible from the camera side, such as the yellow rectangle area in Fig. 7c and d. Those fruits could potentially be detected and then considered for robotic harvesting from the other side.

3.4. *Further studies on apple detection for robotic harvesting*

In this study, the size of the training images was 1920 × 1080 pixels, which was quantised to 224 × 224 pixels by the Faster R–CNN_based ZFNet and VGG16 architectures as the input for training. In this study, we did not consider the information loss that might be caused by the quantisation, and we did not create 224 × 224 patches from the training images. A future work is to create 224 × 224 patches from the training image to increase the size of the training dataset. Besides, the Faster R–CNN is used through transfer learning without optimisation on the depth-filtered images. We are planning to optimise the Faster R–CNN for apple detection on depth-filtered images as a future study. We believe that the improvement demonstrated with depth filtering would remain valid with the optimised network architecture, but this is something we will validate and report on in the future.

Since the Kinect V2 is still sensitive to ambient IR, this work focused on image data collected during morning (7AM–10AM) and afternoon (3PM–8PM) avoiding sun being directly above (around noon time) canopies/sensor. Some other sensors, and sensing systems and techniques that allow collection of images throughout the day and also in the night can be considered. Stereo camera is one of the options that has been studied and employed for long time in fruit detection and robotic harvesting. However, stereovision systems have also suffered from variable lighting conditions leading to comparatively less accurate detection and thus 3D localization. In recent years, development and application of deep learning methods in the fruit detection has achieved improved results, which can improve image correspondence and 3D localization accuracy using stereo-vision systems (Tang et al., 2020). In addition, synchronized lighting-based stereo-systems have been an area of research and development to improve uniformity of lighting. These techniques are expected to provides a good opportunity to improve the field application of stereo-vision camera for robotic fruit harvesting.

Most studies on apple detection identified all the target fruits as one class, while the rest was the background. However, some detected fruits that were partly visible but behind the branches (Fig. 8a), trellis wires (Fig. 8b), and leaves (Fig. 8c) are not easily accessible for robotic harvesting systems being developed around the world. Besides, picking apples occluded by branches or wires is difficult by current linear apple-picking robots. The robotic end-effecter or fruit may be damaged if the robot forcibly picks the apples that are occluded by branches or wires, causing an unpredictable amount of economic losses. Therefore, more than one object classes such as clustered apples, apples blocked by branch/trellis wire, and apples occluded by leaves need to be detected to improve the efficiency of robotic harvesting. For example, detecting the clustered apples class can help in determining

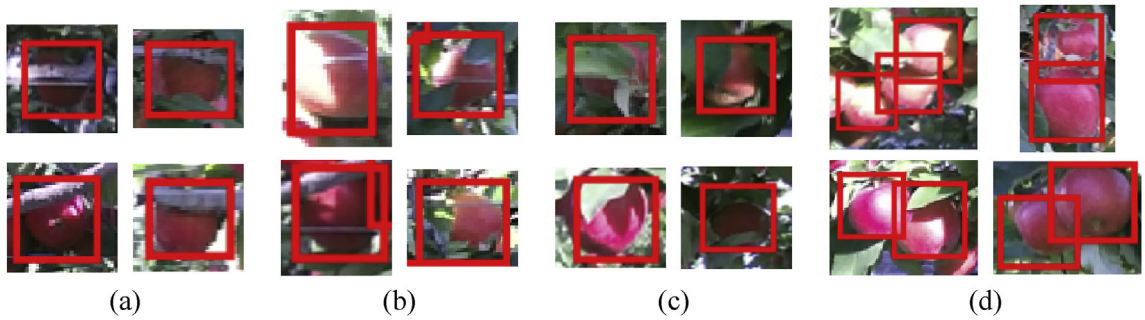

Fig. 8 – Possible classes of detected apples for robotic picking. (a) Fruit blocked by branch; (b) Fruit occluded by trellis wire; (c) Fruit blocked by leaves; and (d) Clustered fruit.

whether to reach the fruit in the front before reaching the one behind, as shown in Fig. 8d, whereas the apples occluded by leaves may be harvested together with leaves. In addition, the apple blocked by branch/trellis wire should be avoided from robotic picking to protect both the robot and fruit. For vertical-fruiting-wall apple trees, fruits that cannot be reached from one side can be reached from the other side. However, in the V-trellis system, fruits on a tree can only be picked from one side; therefore, a human-machine collaboration may be essential in harvesting all fruits.

4. Conclusions

This research examined automatically detecting apples (during harvest season) on vertical-fruiting-wall trees for robotic harvesting using RGB and depth images (provided by a Kinect V2 sensor) and Faster R–CNN. In total, 800 images (point cloud) with a pixel resolution of 1920 × 1080 were acquired from a commercial apple orchard near Prosser, WA, in consecutive two years (2017 and 2018). All images were divided into three parts for network training (70%), validation (15%), and testing (15%). Lastly, two groups of images (i.e. Original-RGB images and Foreground-RGB images with depth filtering of unwanted background) were created, and the results with those groups of images were compared using two different pretrained network architectures (i.e. ZFNet and VGG16). The performances of these networks were evaluated using *AP*, and the results showed that higher *AP* overall was achieved with Foreground-RGB images than the same with Original-RGB images for a given network architecture, whereas VGG16 architecture performed better than ZFNet with a given dataset. The highest *AP* of 0.89 was achieved with Foreground-RGB images using VGG16, which took 0.181 s on average to process a test image. This processing speed enables the fruit detection system to run at near real-time. The results indicated that robotic apple harvesting could potentially achieve much better efficiency in terms of picking individual apples by combining depth features to remove unwanted background (including foliage and fruit in the trees next to the desired row of trees). Only one object class (apples) was detected in the study. Therefore, future work could include detecting multiple object classes, such as clustered apples, apples blocked by branches/trellis wires, and apples occluded by leaves to realise a more efficient robotic harvesting.

Funding

This work was supported by the China Postdoctoral Science Foundation funded project (2019M663832); Key Research and Development Program in Shaanxi Province of China (grant number 2018TSCXL-NY-05-04, 2019ZDLNY02-04); National Natural Science Foundation of China (grant number 31971805); International Scientific and Technological Cooperation Foundation of Northwest A&F University (grant number A213021803).

Declaration of Competing Interest

None declared.

Acknowledgements

The authors express their deep gratitude to the Young Faculty Study Abroad Program of the Northwest A&F University Scholarship who sponsored Dr Longsheng Fu in conducting post-doctoral research at the Centre for Precision and Automated Agricultural Systems, Washington State University, and to the Allan Brothers Fruit Company who provided the experimental orchard.

REFERENCES

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. *Journal of Field Robotics, 31*(6), 888–911. https://doi.org/10.1002/rob.21525

Bargoti, S., & Underwood, J. P. (2017). Image segmentation for fruit detection and yield estimation in apple orchards. *Journal of Field Robotics, 34*(6), 1039–1060. https://doi.org/10.1002/rob.21699

Bechar, A., & Vigneault, C. (2016). Agricultural robots for field operations: Concepts and components. *Biosystems Engineering, 149*, 94–111. https://doi.org/10.1016/j.biosystemseng.2016.06.014

Fu, L., Feng, Y., Majeed, Y., Zhang, X., Zhang, J., Karkee, M., et al. (2018). Kiwifruit detection in field images using Faster R-CNN

with ZFNet. *IFAC-PapersOnLine, 51*(17), 45–50. https://doi.org/10.1016/j.ifacol.2018.08.059

Fu, L., Tola, E., Al-Mallahi, A., Li, R., & Cui, Y. (2019). A novel image processing algorithm to separate linearly clustered kiwifruits. *Biosystems Engineering, 183*, 184–195. https://doi.org/10.1016/j.biosystemseng.2019.04.024

Gan, H., Lee, W. S., Alchanatis, V., Ehsani, R., & Schueller, J. K. (2018). Immature green citrus fruit detection using colour and thermal images. *Computers and Electronics in Agriculture, 152*, 117–125. https://doi.org/10.1016/j.compag.2018.07.011

Gené-Mola, J., Gregorio, E., Auat Cheein, F., Guevara, J., Llorens, J., Sanz-Cortiella, R., et al. (2020). Fruit detection, yield prediction and canopy geometric characterization using LiDAR with forced air flow. *Computers and Electronics in Agriculture, 168*, 105121. https://doi.org/10.1016/j.compag.2019.105121

Gené-Mola, J., Gregorio, E., Guevara, J., Auat, F., Sanz-Cortiella, R., Escolà, A., et al. (2019). Fruit detection in an apple orchard using a mobile terrestrial laser scanner. *Biosystems Engineering, 187*, 171–184. https://doi.org/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., et al. (2020). Fruit detection and 3D location using instance segmentation neural networks and structure-from-motion photogrammetry. *Computers and Electronics in Agriculture, 169*, 105165. https://doi.org/10.1016/j.compag.2019.105165

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*, 689–698. https://doi.org/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. *Computers and Electronics in Agriculture, 116*(C), 8–19. https://doi.org/10.1016/j.compag.2015.05.021

Gongal, A., Karkee, M., & Amatya, S. (2018). Apple fruit size estimation using a 3D machine vision system. *Information Processing in Agriculture, 5*(4), 498–503. https://doi.org/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. *Computers and Electronics in Agriculture, 120*, 26–35. https://doi.org/10.1016/j.compag.2015.10.022

Hameed, K., Chai, D., & Rassau, A. (2018). A comprehensive review of fruit and vegetable classification techniques. *Image and Vision Computing, 80*, 24–44. https://doi.org/10.1016/j.imavis.2018.09.016

Häni, N., Roy, P., & Isler, V. (2019). A comparative study of fruit detection and counting methods for yield mapping in apple orchards. *Journal of Field Robotics, 36*. https://doi.org/10.1002/rob.21902. rob.21902.

He, L., Fu, H., Karkee, M., & Zhang, Q. (2017). Effect of fruit location on apple detachment with mechanical shaking. *Biosystems Engineering, 157*, 63–71. https://doi.org/10.1016/j.biosystemseng.2017.02.009

He, L., Fu, H., Sun, D., Karkee, M., & Zhang, Q. (2017). Shake-and-catch harvesting for fresh market apples in trellis-trained trees. *Transactions of the ASABE, 60*(2), 353–360. https://doi.org/10.13031/trans.12067

He, L., Zhang, X., Ye, Y., Karkee, M., & Zhang, Q. (2019). Effect of shaking location and duration on mechanical harvesting of fresh market apples. *Applied Engineering in Agriculture, 35*(2), 175–183. https://doi.org/10.13031/aea.12974

Jiang, Y., Li, C., & Paterson, A. H. (2016). High throughput phenotyping of cotton plant height using depth images under field conditions. *Computers and Electronics in Agriculture, 130*, 57–68. https://doi.org/10.1016/j.compag.2016.09.017

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., et al. (2014). *Caffe: Convolutional architecture for fast feature embedding* (p. arXiv:1408.5093). p. arXiv:1408.5093.

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture, 147*, 70–90. https://doi.org/10.1016/J.COMPAG.2018.02.016

Koirala, A., Walsh, K. B., Wang, Z., & Mccarthy, C. (2019a). Deep learning for real-time fruit detection and orchard fruit load estimation: Benchmarking of 'MangoYOLO'. *Precision Agriculture*, 1–29. https://doi.org/10.1007/s11119-019-09642-0

Koirala, A., Walsh, K. B., Wang, Z., & Mccarthy, C. (2019b). Deep learning – method overview and review of use for fruit detection and yield estimation. *Computers and Electronics in Agriculture, 162*, 219–234. https://doi.org/10.1016/j.compag.2019.04.017

Lee, W. S., Alchanatis, V., Yang, C., Hirafuji, M., Moshou, D., & Li, C. (2010). Sensing technologies for precision specialty crop production. *Computers and Electronics in Agriculture, 74*(1), 2–33. https://doi.org/10.1016/j.compag.2010.08.005

Li, R., Kou, X., Cheng, T., Zheng, A., & Wang, S. (2017). Verification of radio frequency pasteurization process for in-shell almonds. *Journal of Food Engineering, 192*, 103–110. https://doi.org/10.1016/j.jfoodeng.2016.08.002

Lin, G., Tang, Y., Zou, X., Xiong, J., & Fang, Y. (2019). Colour-, depth-, and shape-based 3D fruit detection. *Precision Agriculture*, 1–17. https://doi.org/10.1007/s11119-019-09654-w

Lin, G., Tang, Y., Zou, X., Xiong, J., & Li, J. (2019). Guava detection and pose estimation using a low-cost RGB-D sensor in the field. *Sensors, 19*(2), 428. https://doi.org/10.3390/s19020428

Liu, Z., Wu, J., Fu, L., Majeed, Y., Feng, Y., Li, R., et al. (2020). Improved kiwifruit detection using pre-trained VGG16 with RGB and NIR information fusion. *IEEE Access, 8*(1), 2327–2336. https://doi.org/10.1109/ACCESS.2019.2962513

Majeed, Y., Zhang, J., Zhang, X., Fu, L., Karkee, M., & Zhang, Q. (2020). Deep learning based segmentation for automated training of apple trees on trellis wires. *Computers and Electronics in Agriculture, 170*, 105277. https://doi.org/10.1016/j.compag.2020.105277

Majeed, Y., Zhang, J., Zhang, X., Fu, L., Karkee, M., Zhang, Q., et al. (2018). Apple tree trunk and branch segmentation for automatic trellis training using convolutional neural network based semantic segmentation. *IFAC-PapersOnLine, 51*(17), 75–80. https://doi.org/10.1016/J.IFACOL.2018.08.064

Méndez Perez, R., Cheein, F. A., & Rosell-Polo, J. R. (2017). Flexible system of multiple RGB-D sensors for measuring and classifying fruits in agri-food Industry. *Computers and Electronics in Agriculture, 139*(Supplement C), 231–242. https://doi.org/10.1016/j.compag.2017.05.014

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

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR.2014.222

Peng, J., Xie, H., Feng, Y., Fu, L., Sun, S., & Cui, Y. (2017). Simulation study of vibratory harvesting of Chinese winter jujube (Zizyphus jujuba Mill. cv. Dongzao). *Computers and Electronics in Agriculture, 143*, 57–65. https://doi.org/10.1016/j.compag.2017.09.036

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine*

Intelligence, 39(6), 1137–1149. https://doi.org/10.1109/tpami.2016.2577031

Sa, I., Ge, Z. Y., Dayoub, F., Upcroft, B., Perez, T., & McCool, C. (2016). DeepFruits: A fruit detection system using deep neural networks. *Sensors, 16*(8), 23. https://doi.org/10.3390/s16081222

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., et al. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging, 35*(5), 1285–1298. https://doi.org/10.1109/tmi.2016.2528162

Silwal, A., Davidson, J. R., Karkee, M., Mo, C. K., Zhang, Q., & Lewis, K. (2017). Design, integration, and field evaluation of a robotic apple harvester. *Journal of Field Robotics, 34*(6), 1140–1159. https://doi.org/10.1002/rob.21715

Silwal, A., Gongal, A., & Karkee, M. (2014). Apple identification in field environment with over the row machine vision system. *Agricultural Engineering International: CIGR Journal, 16*(4), 66–75.

Simonyan, K., & Zisserman, A. (2015). *Very deep convolutional networks for large-scale visual recognition. 3rd international conference on learning representations*. Retrieved from: http://www.robots.ox.ac.uk/~vgg/research/very_deep/.

Stein, M., Bargoti, S., & Underwood, J. (2016). Image based mango fruit detection, localisation and yield estimation using multiple view Geometry. *Sensors, 16*(11), 1915. https://doi.org/10.3390/s16111915

Tabian, I., Fu, H., & Khodaei, Z. S. (2019). A convolutional neural network for impact detection and characterization of complex composite structures. *Sensors, 19*(22), 4933. https://doi.org/10.3390/s19224933

Tang, Y., Chen, M., Wang, C., Luo, L., Li, J., Lian, G., et al. (2020). Recognition and localization methods for vision-based fruit picking robots : A review. *Frontiers of Plant Science, 11*, 510. https://doi.org/10.3389/fpls.2020.00510

Tao, Y., & Zhou, J. (2017). Automatic apple recognition based on the fusion of colour and 3D feature for robotic fruit picking. *Computers and Electronics in Agriculture, 142*, 388–396. https://doi.org/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. *Computers and Electronics in Agriculture, 157*, 417–426. https://doi.org/10.1016/j.compag.2019.01.012

Vit, A., & Shani, G. (2018). Comparing RGB-D sensors for close range outdoor agricultural phenotyping. *Sensors, 18*(12), 4413. https://doi.org/10.3390/s18124413

Vougioukas, S. G. (2019). Agricultural robotics. *Annual Review of Control, Robotics, and Autonomous Systems, 2*(1). https://doi.org/10.1146/annurev-control-053018-023617, 15.1-15.28.

Yu, Y., Zhang, K., Yang, L., & Zhang, D. (2019). Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. *Computers and Electronics in Agriculture, 163*, 104846. https://doi.org/10.1016/j.compag.2019.06.001

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *European conference on computer vision* (pp. 818–833). https://doi.org/10.1007/978-3-319-10590-1_53

Zhang, L., Gui, G., Khattak, A. M., Wang, M., Gao, W., & Jia, J. (2019). Multi-task cascaded convolutional networks based intelligent fruit detection for designing automated robot. *IEEE Access, 7*, 56028–56038. https://doi.org/10.1109/access.2019.2899940

Zhang, X., He, L., Majeed, Y., Whiting, M. D., Karkee, M., & Zhang, Q. (2018). A precision pruning strategy for improving efficiency of vibratory mechanical harvesting of apples. *Transactions of the ASABE, 61*(5), 1565–1576. https://doi.org/10.13031/trans.12825

Zujevs, A., Osadcuks, V., & Ahrendt, P. (2015). Trends in robotic sensor technologies for fruit harvesting: 2010-2015. *Procedia Computer Science, 77*(Supplement C), 227–233. https://doi.org/10.1016/j.procs.2015.12.378