IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 29, 2019, accepted December 19, 2019, date of publication December 26, 2019,
date of current version January 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2962513

Improved Kiwifruit Detection Using Pre-Trained VGG16 With RGB and NIR Information Fusion

ZHIHAO LIU[1], JINGZHU WU[2], LONGSHENG FU[1,3,4,5], YAQOOB MAJEED[5], YALI FENG[6], RUI LI[1], AND YONGJIE CUI[1]

[1]College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling 712100, China
[2]Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing Technology and Business University, Beijing 100048, China
[3]Key Laboratory of Agricultural Internet of Things, Ministry of Agriculture and Rural Affairs, Yangling 712100, China
[4]Shaanxi Key Laboratory of Agricultural Information Perception and Intelligent Service, Yangling 712100, China
[5]Center for Precision and Automated Agricultural Systems, Washington State University, Prosser, WA 99350, USA
[6]College of Engineering, Shanxi Agricultural University, Jinzhong 030801, China

Corresponding author: Longsheng Fu (fulsh@nwafu.edu.cn)

This work was supported in part by the Key Research and Development Program in Shaanxi Province of China under Grant 2018TSCXL-NY-05-04 and Grant 2019ZDLNY02-04, in part by the China Postdoctoral Science Foundation funded project under grant 2019M663832, in part by the National Natural Science Foundation of China under Grant 31971805, and in part by the Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing Technology and Business University under Grant BTBD-2019KF03.

ABSTRACT This study presents a novel method to apply the RGB-D (Red Green Blue–Depth) sensors and fuse aligned RGB and NIR images with deep convolutional neural networks (CNN) for fruit detection. It aims to build a more accurate, faster, and more reliable fruit detection system, which is a vital element for fruit yield estimation and automated harvesting. Recent work in deep neural networks has led to the development of a state-of-the-art object detector termed Faster Region-based CNN (Faster R-CNN). A common Faster R-CNN network VGG16 was adopted through transfer learning, for the task of kiwifruit detection using imagery obtained from two modalities: RGB (red, green, blue) and Near-Infrared (NIR) images. Kinect v2 was used to take a bottom view of the kiwifruit canopy's NIR and RGB images. The NIR (1 channel) and RGB images (3 channels) were aligned and arranged side by side into a 6-channel image. The input layer of the VGG16 was modified to receive the 6-channel image. Two different fusion methods were used to extract features: Image-Fusion (fusion of the RGB and NIR images on input layer) and Feature-Fusion (fusion of feature maps of two VGG16 networks where the RGB and NIR images were input respectively). The improved networks were trained end-to-end using back-propagation and stochastic gradient descent techniques and compared to original VGG16 networks with RGB and NIR image input only. Results showed that the average precision (APs) of the original VGG16 with RGB and NIR image input only were 88.4% and 89.2% respectively, the 6-channel VGG16 using the Feature-Fusion method reached 90.5%, while that using the Image-Fusion method reached the highest AP of 90.7% and the fastest detection speed of 0.134 s/image. The results indicated that the proposed kiwifruit detection approach shows a potential for better fruit detection.

INDEX TERMS Fruit detection, image alignment, information fusion, multi-modality faster R-CNN, RGB-D sensor.

I. INTRODUCTION

China is the largest country producing kiwifruits worldwide, with a yield of approximately 2.4 million tons in 2016 from a cultivated area of 197,048 ha [1]. Within China, Shaanxi Province has the most significant production, accounting for approximately 70% and 33% of the Chinese and global productions, respectively [2]. Harvesting kiwifruits in this region relies mainly on manual picking, which is labor-intensive [3]. Therefore, the introduction of robotic harvesting is highly desirable.

Kiwifruits are commercially grown on sturdy support structures, such as T-bar and pergola systems [4]. T-bar trellis is common in China because of its low cost. It consists of a 1.7 m high post and an approximately 1.7 m wide cross arm, which may vary slightly in width according to the shape and size of the orchard. Wires run on the top of cross arms connecting them in the middle and on both edges of the cross arms. The upper stems of the kiwifruit are tied to the top wires to keep the egg-sized kiwifruits hanging downward,

The associate editor coordinating the review of this manuscript and approving it for publication was Peng Liu.

which makes them visible and accessible for manual picking [5]. This canopy structure also provides relatively simpler and structured workspace for mechanized or automated field operations such as robotic picking [3], [6], compared to other fruit trees such as apples [7].

Fast and effective detection of fruit in the field under natural environment is the first step for robotic kiwifruit harvesting [8]. Cui *et al.* [9] captured side-viewed kiwifruit RGB (Red, Green, Blue) images and used Otsu threshold in 0.9R-G component for fruit segmentation, and detected each fruit using Canny operator and elliptical Hough transformation, which reached 89.1% detection rate. Mu *et al.* [10] employed L*a*b* color space to extract side-viewed kiwifruits fruit edges with Canny operator and elliptical Hough transformation, which detected 88.5% kiwifruits and spent 3.98 s on each fruit. Fu *et al.* [3] segmented bottom-viewed kiwifruit images using Otsu threshold in 1.1R-G color component and detected each fruit using minimal bounding rectangle and elliptical Hough transform, which achieved 88.3% correct detection and spent 1.51 s on each fruit. Fu *et al.* [11] developed a kiwifruit detection system at night using artificial lighting by identifying the fruit calyx, which reached a success rate of 94.3% and took 0.5 s in average to recognize a fruit. Fu *et al.* [12] then proposed an image processing algorithm to separate linearly clustered kiwifruits by scanning each detected cluster to find the contact points between the adjacent fruits and drawing a separating line between the two closest contact points, which correctly separated and counted 92.0% of the kiwifruits. These studies primarily used color and shape of calyx and fruit to detect the kiwifruit, which was limited to detect fruit in a single cluster with few fruits and were less effective for multi clusters in the field. To overcome these limitations, a well generalized methodology that invariant and robust to brightness, different viewpoints and highly discriminative feature representations is desired.

Recently, deep learning has made considerable progress in object detection and classification. In particular, convolutional neural networks (CNN) showed superior performance in object detection applications [13]–[15]. There have been some researches using different CNN architectures for fruit detection such as apple [16]–[18], mango [19], strawberry [20] and kiwifruit [21], [22]. Bargoti and Underwood [23] applied VGG16 to detect mangoes and apples in orchards, which achieved an F1-score of 0.9. Häni *et al.* [24] employed U-Net and ResNet50 for apple detection and yield estimation, which reached 97.8% accuracy. Fu*et al.* [21] used LeNet to detect kiwifruits in the orchard, which reached 89.29% detection rate and cost 0.27 s on average to recognize a fruit. Fu *et al.* [22] presented a kiwifruit image detection using ZFNet and achieved 92.3% detection rate and cost 0.005 s on average to detect a fruit. These studies showed good promising for fruit detection in RGB images using CNN.

However, above mentioned studies only used RGB images as input for the CNNs. In uncontrolled field conditions, a single sensor modality can rarely provide the needed information to detect the target fruits under a wide range of varying illumination, partial occlusions, and different appearances. As such, multi-modality sensors may be more beneficial because different sensors can provide complementary information regarding various aspects of the fruits. Sa *et al.* [25] proposed early fusion and late fusion methods of RGB and Near-Infrared (NIR) images based on VGG16 to detect rockmelon, strawberry, apple, avocado, mango, orange, and sweet pepper, of which the highest F1-score reached 0.838. Zhan *et al.* [26] employed RGB and NIR fusion algorithm to distinguish the chestnut quality based on back propagation network, of which the discriminating rate is improved by 3.75% and 6.25%, respectively, compared to using NIR and RGB image separately Abdelsalam and Sayed [27] extracted seven color components from RGB and NIR images of citrus and applied a voting process algorithm to detect citrus defects, of which the accuracy is more than 95%. Zhang *et al.* [28] propose two ship models, the "V" ship head model and the "||" ship body one, and fed these features to a novel ship detection method (SCNN) which is more robust than previous methods. Bai *et al.* [29] present a new eddy detection approach of combining the multilayer features in the neural network with the characteristics of the eddies via deep neural networks to improve eddy detection accuracy, which results in mAP (mean Average Precision) of 90.6%. These studies achieved higher accuracy using multi-modality information fusion. However, they didn't report the impact of information fusion on the operation speed. In addition, the alignment of RGB and NIR images that being fused for CNN was not mentioned. Besides, most of the emerging sensors, such as depth sensors and RGB-D cameras, have not yet been exploited for fruit detection. The primary reason is the lack of substantial datasets [30].

This paper introduces a Hayward-Kiwi RGB-NIR-D dataset, which contains multi-modality aligned images of Hayward kiwifruits in orchards, and presents a novel method to apply the RGB-D sensors and fuse aligned RGB and NIR images with deep learning methods for fruit detection. Faster Region-based CNN (Faster R-CNN) is adapted and implemented for kiwifruit detection using two-modality of aligned RGB and NIR images from the dataset. Two different fusion methods are studied: fusion of the RGB and NIR images on input layer of the Faster R-CNN and fusion of feature maps of two Faster R-CNN networks where the RGB and NIR images were input respectively.

II. MATERIALS AND METHODS

A. HAYWARD-KIWI RGB-NIR-D DATASET

1) IMAGE ACQUISITION

All the kiwifruit images were captured during three harvesting seasons (2016, 2017 and 2018) on the most common cultivar 'Hayward' at Meixian Kiwifruit Experimental Station (34°07'39" N, 107°59'50" E, and 648 m in altitude) of the Northwest A&F University, Shaanxi, China. An image acquisition platform (consisted of an RGB-D camera mounted on

TABLE 1. Specific parameters of the RGB and NIR cameras in Kinect V2.

RGB camera resolution (pixels)	RGB camera field-of-view (FoV)	IR and depth camera resolution (pixels)	IR and depth camera FoV	Working range (m)
1920 × 1080	84.1º × 53.8º	512 × 424	70º × 60º	0.5 - 8

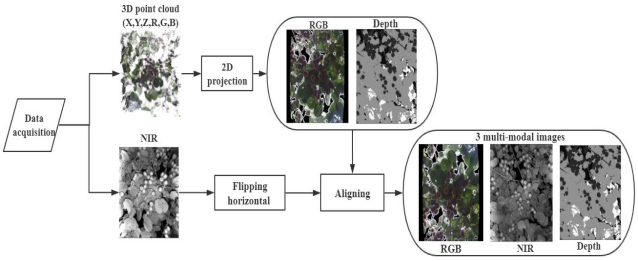


FIGURE 1. Flowchart to prepare Hayward-Kiwi RGB-NIR-D image dataset.

a mobile tripod at the height of 1.0 m) was used to capture images from the bottom of kiwifruit. It was connected to a laptop with Intel Core i7-8750H (4.1 GHz) six-core CPU, a GPU of NVIDIA GTX 1060 6 GB GPU and 8 GB of memory via USB 3.0. The RGB-D camera, Kinect v2 (Microsoft, Redmond, WA, USA) incorporates the RGB camera and a depth sensor that works using NIR according to the ToF (Time of Flight) principle (Song et al. 2017), as shown in Table 1. This camera provides three different types of images: a three-channel RGB image, a single-channel depth image that can be used to generate a three-dimensional (3D) point cloud of the scene, and a single-channel NIR image taken by the depth sensor. The single-channel NIR image was copied and concatenated into a 3-channel NIR image for subsequent experiments. Specific software written in MATLAB 2018a was developed to collect and save images automatically into the laptop. The software generates a 3D point cloud with RGB and depth information for each point based on the depth sensor, and a NIR image for the same scene. Image collection was carried out during all day and night where suitable artificial lighting 30∼50 lux was provided at night, according to Fu *et al.* [3]. This is to ensure the proposed methods can work on different time with varied lighting conditions.

2) IMAGE PREPARATION

The data included a 3D point cloud (with RGB and depth information) and the corresponding NIR image, which is captured from the same scene based on the perspective of the depth sensor. A pre-processing was carried out to align the acquired images: RGB and depth images extracted from the 3D point cloud, and NIR image. Image preparation included two-dimensional (2D) projection of the 3D point cloud, horizontal flipping of the NIR images, and aligning the NIR to the RGB images. Fig. 1 illustrates a flowchart of the image preparation.

After obtaining the 3D point cloud generated by the depth sensor of Kinect v2, a perspective projection onto a plain parallel of kiwifruit bottom was implemented to generate the corresponding projected RGB and depth images. The vertical field-of-view (FoV) of the depth sensor is larger than that of the RGB camera, as shown in Table 1. Due to this reason, a part of the RGB information was not registered with the depth information in the 3D point cloud generated by the depth sensor. Thus, there was no information given at the left and right parts of the RGB images extracted from the 3D point cloud.

Horizontal flipping of the NIR images and aligning them to the RGB images were performed to align the three images (RGB, depth, and NIR) with the same pixel resolution. A group of multi-modality images was obtained where each pixel has information from 3 modalities: RGB, NIR and depth (Fig. 1). The RGB and corresponding NIR images from multi-modality images of the same group were combined as a 6-channel image (with 3-channel RGB and 3-channel NIR) and fed as an input of the CNN. To have similar mean and variance between channels, the NIR and depth images of all groups were normalized between 0 and 255 as the RGB images. This normalization was desirable to ensure fast convergence of the network. The RGB channels were saved in 8-bit images while NIR and depth images were stored in 64-bits to avoid data precision loss.

In total, there are 1000 NIR and 1000 corresponding aligned RGB images with the resolution of 512 × 424 pixels were collected for this research. The Hayward-Kiwi RGB-NIR-D dataset consisted of 1000 groups of multi-modality images with 512×424 pixels resolution. Each image included around 30 to 50 target kiwifruits. Ground truth kiwifruit targets were manually annotated in the RGB images using the rectangular annotations and then mapped to the NIR and depth images of the same group, as shown in Fig. 2. A total of 39,678 fruit were labelled in the RGB images of all the dataset. All the labelled dataset of multi-modality images for kiwifruit was divided into training (70%) and testing (30%) groups. The training images were randomly obtained from the independent and uniform sampling of the whole dataset.

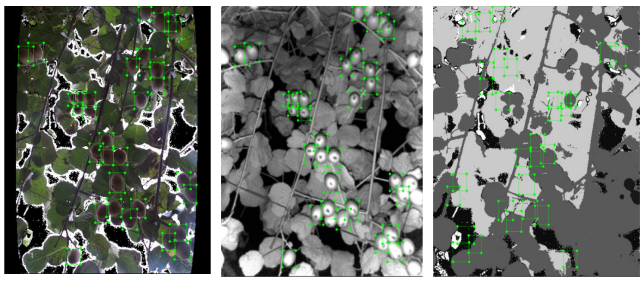


FIGURE 2. Sample of multi-modality images extracted from training dataset and their associated fruit location ground truth (green bounding boxes).

The training and testing images were mutually exclusive, which ensured the reliability of the later evaluation standards. Some examples of the multi-modality kiwifruit images in the training dataset are shown in Fig. 2. In order to further quantization for representation, the depth images saved in 16 bit were mapped to 0∼255 and their contrast was enhanced by histogram equalization. The Hayward-Kiwi RGB-NIR-D dataset with corresponding annotations has been made publicly available at https://github.com/Hayward-kiwi/Hayward-Kiwi-RGB-NIR-D.

B. FRUIT DETECTION USING FASTER R-CNN

The Faster R-CNN architecture merges region proposals, objects classification, and detection into one unified deep object detection network [32], [28]. In Faster R-CNN, two networks (Region Proposal Network and Faster R-CNN) are concatenated as one that can be trained and tested through an end-to-end process [23]. In this paper, a state-of-the-art Faster R-CNN network VGG16 [32], [33], [29] was employed for kiwifruit detection, as shown in Fig. 3. The VGG16 network secured second place in the 2014 ILSVRC (ImageNet Large Scale Visual Recognition Challenge competition). However, it performed better than the first place (GoogleNet) in multiple migration learning tasks [34], [35]. The convolutional layers of original VGG16 network are pre-trained with ImageNet dataset [13] and fine-tuned with the kiwifruits training dataset of RGB and NIR images, which were denoted as RGB-Only and NIR-Only mode respectively. The depth images were not employed in this research because only the highly exposed kiwifruits are not visible in the image, as shown in Fig.2c.

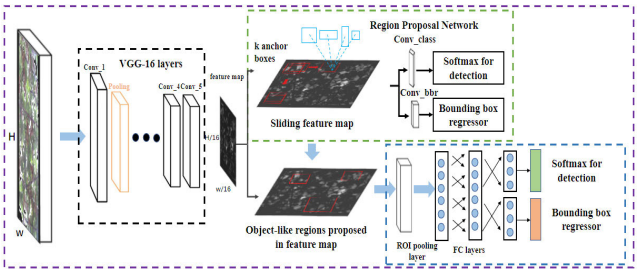


FIGURE 3. Common VGG16 architecture for training kiwifruit RGB images (3 channels input).

C. MULTI-MODALITY FUSION

Two modified networks were developed to receive and fuse the aligned RGB and NIR images. One was VGG16 network that received RGB and NIR images simultaneously. The other was two VGG16 networks that received RGB and NIR images respectively, then being concatenated on the feature map. They were denoted as Image-Fusion and Feature-Fusion modes, respectively.

1) IMAGE-FUSION MODE

The Image-Fusion mode altered the structure of the input layer of the VGG16 network from 3 channels to 6 channels (3 channels for the RGB image and 3 channels for the NIR image). The VGG16 network was modified and adapted to receive RGB and NIR information simultaneously. An overview of this was provided in Fig. 4a, where the R, G, B responses from the first convolutional parameter of the RGB-Only mode were average to initialize the 3 channels for the NIR images.

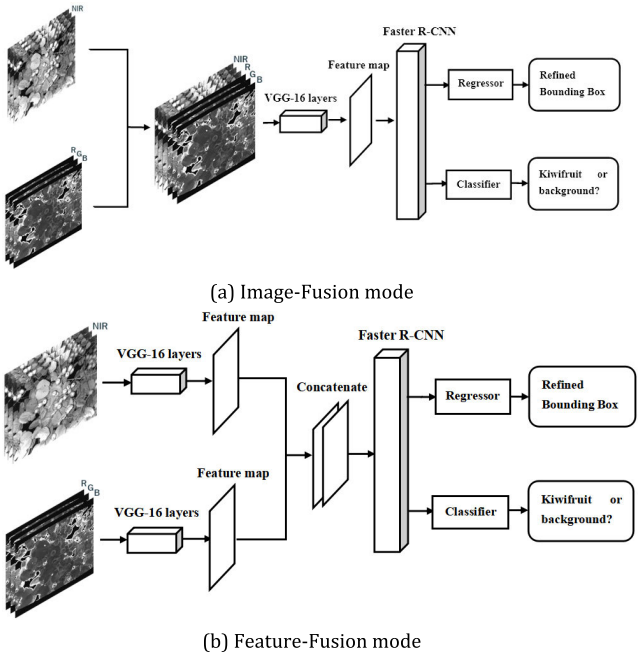


FIGURE 4. Diagrams of the two methods to fuse the aligned RGB and NIR images. (a) Image-Fusion mode and (b) Feature-Fusion mode.

2) FEATURE-FUSION MODE

The Feature-Fusion mode inputted the RGB and NIR images separately into two VGG16 networks and then combined them on the feature map. To achieve this, a Concat Layer was added to concatenate the feature maps of RGB and NIR images after convolution, as shown in Fig. 4b. The two VGG16 networks were respectively initialized by the parameter of the RGB-Only and NIR-Only modes and then fine-tuned as the original VGG16 network.

D. NETWORK TRAINING

Training platform included a desktop computer with Intel Xeon E5-1650 (3.60 GHz) six-cores CPU, and a GPU of NVIDIA TITAN XP 12 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 [36]. The test runs on the same platform as the training.

To train the deep learning networks, two sets of information were required, including the images to train with and a corresponding label for each image. Stochastic gradient descent (SGD) was used to train the four different modes (RGB-Only, NIR-Only, Image-Fusion, Feature-Fusion) and the momentum of the network was set to a fixed value

of 0.9 and a weight decay of 0.0005. An optimal learning rate was employed to help the neural network to be quick enough to be trained with significant features. In this work, the learning rate of 0.001 was reduced by changing it to 10% of the current rate per 10,000 iterations.

During training, the RGB-Only and NIR-Only modes were firstly initialized by the parameters of the ImageNet network and updated by training the RGB and NIR training dataset respectively. And then, the parameters of the two VGG16 networks in the Feature-Fusion mode were respectively initialized by the parameters of the RGB-Only and NIR-Only modes trained before and trained by the 6-channel images. Since the network structure of the Image-Fusion mode was similar to that of the RGB-Only mode, the parameters of the Image-Fusion mode were initialized by the trained RGB-Only mode, and the extra 3 channels of first convolution layer were initialized by averaging the R, G, B channels. Finally, the SGD was used to update the initialized parameters of fusion modes and train by the way of training the original VGG16 network.

E. FRUIT CATEGORIES

The kiwifruits in the images taken in the field were not all independent (or singulated) from each other. In this work, fruits were categorized into four groups according to the degree of visibility of fruit area in the image. The first category denoted as occluded fruit, as shown in Fig. 5a, included fruit partially occluded by leaves or branches causing the incomplete outline of the fruit. The second category refers to the fruit with some overlap between each other from the image capture angle, which was denoted as overlapping fruit, as shown by the rectangular boxes in Fig. 5b. The third category referred to the fruits where contours of two or more fruits were adjacent to each other, which was denoted as adjacent fruit, as shown in Fig. 5c. The fourth category was denoted as separated fruit, in which fruit contours were completely independent and separated from each other, as shown in Fig. 5d.

(a) Occluded (b) Overlapped (c) Adjacent (d) Separated

FIGURE 5. Categories of kiwifruit image in the field.

F. EVALUATION

The performance of the networks was evaluated by precision (P), recall (R), average precision (AP), detection rate (DR) and detection speed. The AP is a standard for measuring the sensitivity of the network to object, and an indicator that reflects the global performance of the network.

The DR was defined by the proportion of the detected kiwifruit number in the total number of each kiwifruit category. The kiwifruits in images were categorized into four groups (i.e. occluded, overlapped, adjacent, separated). The DRs of various kiwifruit categories evaluated the performance of the networks on various kiwifruit categories giving comprehensive and accurate measure of performance of networks in kiwifruit detection.

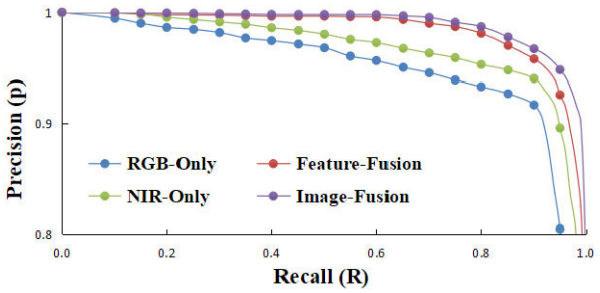


FIGURE 6. Precision-Recall (P-R) curves of the four modes based on VGG16.

III. RESULT AND DISCUSSION

A. EVALUATION OF THE FOUR DIFFERENT MODES

The Precision-Recall (P-R) curves of the four modes were shown in Fig. 6. As expected, the P of the Image-Fusion and Feature-Fusion modes were higher than that of the RGB-Only and NIR-Only modes under the same R condition. While the same results were also presented in the APs, as shown in Table 2. It was noted that the Feature-Fusion mode contains twice as many parameters as the others and requires more resources such as computation time and GPU memory space, because of its two VGG16 networks, as shown Fig. 4b. As such, it cost more time of 0.188 s to process each image under steady state for detection than the other three modes. The initial detection speed will be affected by the CPU cache. The detection speed of the Image-Fusion, RGB-Only, and NIR-Only modes are almost same under steady state (i.e., 0.134, 0.134, 0.135 s/image). It can be observed that the number of parameters of a network is a critical component to increase detection performance, as mentioned by Eitel *et al*. [37]. Regarding the computational efficiency of the neural network, the number of inferred images per second was irrelevant to the number of channels used. This is because the addition of channels only affects the number of operations

TABLE 2. Kiwifruit detection results from the testing dataset using the four modes.

Mode	AP (%)	Detection speed under steady state (s/image)[a]
RGB-Only	88.4	0.134
NIR-Only	89.2	0.134
Image-Fusion	90.7	0.135
Feature-Fusion	90.5	0.188

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

on the first layer, which is insignificant with respect to the whole network.

The NIR-Only mode gave a better performance with the AP of 89.2% than that of the RGB-Only mode (88.4%), as shown in Table 2. Although NIR images have never been reported for kiwifruit detection, the results using the NIR images only are superior of the RGB images. This is because the NIR image obtained by Kinect v2 is enhanced by its own infrared light source, thus effectively avoid the changes in image brightness caused by environmental changes. However, the results are different from Sa *et al*. [25] where the RGB images slightly outperforms the NIR images on sweet pepper since the pre-trained ImageNet parameters are more suitable to operate with RGB inputs. The lower AP of the RGB-Only mode in this study may result from the information loss of the aligned RGB images, as shown in Fig. 2a, whereas the RGB images of the sweet pepper were not reported to be aligned with their NIR images in Sa *et al.* [25].

The best result, with the highest AP of 90.7% and the fastest detection speed of 0.134 s/image, was obtained by the Image-Fusion mode. The Feature-Fusion mode has a quite close AP (90.5%) to that of the Image-Fusion mode, although double convolutional layers were used to extract kiwifruit image features. There have been some studies working on the relationship between extracted features and detection results [38], [39]. Based on those researches, a hypothesis that the Feature-Fusion mode has two VGG16 networks to learn the features from the RGB and NIR images respectively, which results in a duplicate features learning of some important features belonging to both RGB and NIR images, such as fruit shape and calyx shape. The duplicate features may cause a reduction of features on the concatenate operation of the Feature-Fusion mode. On the other hand, the Image-Fusion mode using one VGG16 network to learn the features of aligned RGB and NIR images simultaneously, which avoid the duplicate learning of the same features of both RGB and NIR images. It could be concluded that the RGB and NIR images are helping each other to reduce the false positives and the modes fusing RGB and NIR can achieve higher accuracy than RGB alone. Sample images of this effect can be found in Fig. 7, where, when comparing results before and after fusing RGB and NIR images, a reduction in the false positives was observed.

B. VALIDATION OF THE FOUR MODES ON THE MODES ON THE ORIGINAL IMAGES

The four trained modes were tested for the detection accuracy of the four different kiwifruit categories with 50 kiwifruit images (a total of 2,518 fruits) that obtained in the field under different lighting conditions. Most kiwifruit in the field images are adjacent to each other, accounted for 60.5% (1524 of 2518), as shown in Table 3. It is consistent with a field survey results of Fu *et al.* [3].

Similar to the APs of the four modes, the total DR of the Image-Fusion mode was higher than the other three modes,

TABLE 3. Detection results of kiwifruit images in different categories using the four modes.

Kiwifruit categories	Number of fruits	DR (%)			
		RGB-Only	NIR-Only	Image-Fusion	Feature-Fusion
Occluded Fruit	346	82.4	82.5	83.6	83.5
Overlapped Fruit	224	85.3	85.3	87.4	87.5
Adjacent Fruit	1,524	88.2	89.5	92.8	91.5
Separated Fruit	424	94.3	94.4	96.7	95.9
Total	2,518	88.1	89.0	91.7	90.8

reached 91.7% for images with various kiwifruit categories, it was 0.9% higher than the Feature-Fusion mode (90.8%), 2.7% higher than the NIR-Only mode (89.0%), and 3.6% higher than the RGB-Only mode (88.1%). Images detected by the Image-Fusion mode has least undetected kiwifruits than other modes, as shown in the first column of Fig. 7. Although the NIR-Only mode has a higher DR than the RGB-Only mode, the position of detection box is less precise than that of the RGB-Only mode, as shown in the third column of Fig. 7.

The Image-Fusion mode achieved higher DR and more precise position, which proved that a single sensor may not provide all the needed information to detect the target fruits under a wide range of variations in field illumination. Varying types of sensors can provide complementary information regarding different aspects of the fruits, as described in Sa *et al.* [25], thus the fusion modes could achieve better detection performance. It was noted that the kiwifruits in the edge of NIR images are hardly detected by the NIR-Only mode, as shown in the third column of Fig. 7. This is because the annotated NIR images were mapped from the corresponding RGB images, of which the kiwifruit in the edges was trained as background.

In terms of various kiwifruit categories, all the four modes obtained the highest DRs (i.e., 94.3% of the RGB-Only mode, 94.4% of the NIR-Only mode, 96.7% of the Image-Fusion mode, and 95.9% of the Image-Fusion mode) on the separated fruits, respectively. It was followed by the adjacent fruits, overlapped fruit, and occluded fruit orderly in each mode, as shown in Table 1. The results are similar to conclusions from other works on fruit detection that the occluded and overlapped fruits have been the most challenge task.

The DRs of the RGB and NIR fusion modes showed the same trend and higher than that only using one image type. The DR of the adjacent fruit improved from 88.2% of the RGB-Only mode to 92.8% of the Image-Fusion mode, and the overlapped fruit improved from 85.3% to 87.4%. As shown by the yellow circles (the undetected adjacent fruit) and the yellow rectangles (the undetected overlapped fruit)

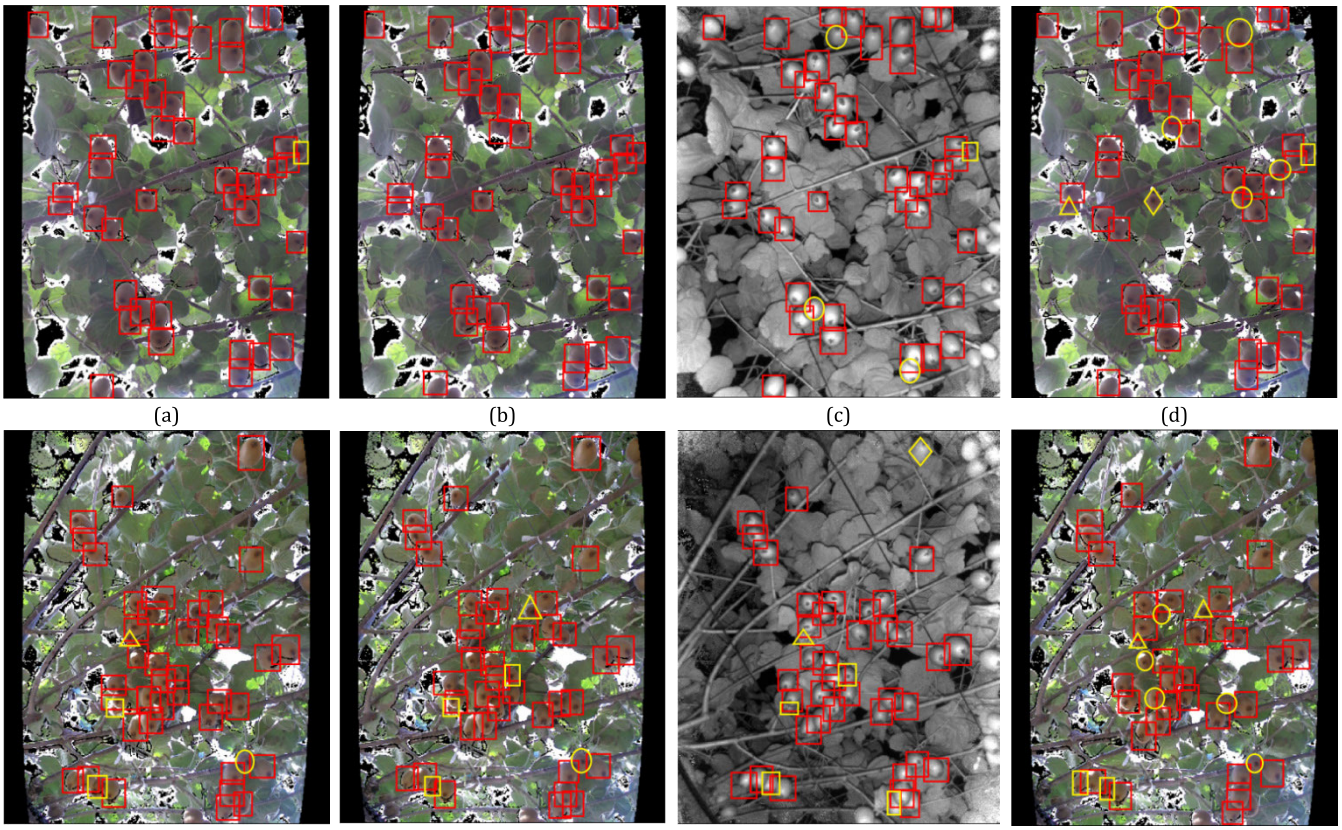


FIGURE 7. Kiwifruit image detection examples of the four different modes. Image-Fusion mode (first column); Feature-Fusion mode (second column); NIR-Only mode (third column); RGB-Only mode (fourth column). Note: The undetected occluded, overlapped, adjacent and separated fruit are respectively marked by four different yellow marks (triangle, rectangle, circle, and diamond).

of Fig. 7h, the detection results of the RGB-Only mode have certain adjacent and overlapped fruit leakage phenomenon, while most fruits in the same position were detected by Image-Fusion mode, as shown in Fig. 7e.

For the two fusion modes, the DRs of the adjacent (91.5%) and separated (95.9%) fruit in the Feature-Fusion mode were lower than the DRs of the Image-Fusion mode, while that of the occluded and overlapped fruit were near to each other. It was noticed that fusion on images performed better than fusion on features for the adjacent and separated fruit, while no differences on the occluded and overlapped fruit.

C. SUMMARY OF KIWIFRUIT DETECTION WORKS

Although it is difficult to compare different works tested with different kiwifruit datasets since there is no public open datasets for fruit detection researches. The performance of two other current state-of-the-art kiwifruit detection using LeNet and ZFNet by Fu *et al.* [21a] and Fu *et al.* [22b] and a kiwifruit segmentation using FCN-8S by Williams *et al.* [6] was summarized and analyzed by image type, camera, image resolution (pixel), detection rate (%), and detection speed (s/image), as shown in Table 4.

The images taken by Kinect v2 for the proposed methods have 512 × 424 pixels resolution which is lower than that of 2352 × 1568 pixels in Fu *et al.* [21a] and Fu *et al.* [22b] and 1900 × 1200 in Williams *et al.* [6], as shown in Table 4. High-resolution images were helping for obtaining higher DRs but resulting slower detection speeds than low-resolution images. The result can be found in Fu *et al.* [22b] and the RGB-Only mode in Table 4. The RGB-Only mode detected on the 512 × 424 image showed a DR of 88.1% which is lower than the 92.3% of Fu *et al.* [22b]. On the other hand, the detection speed of 0.134 s/image was significantly faster than that of the 0.274 s/image of Fu *et al.* [22b]. The FCN-8S used in Williams *et al.* [6] is a pixel-wise detector [40], of which the positioning is more accurate than other region-wise detectors. However, in detecting high resolution images, the pixel-by-pixel classification in Williams *et al.* [6] tends to be slow (3.0 s/image). Besides, Williams *et al.* [6] has a total DR of 81.6% in kiwifruit detection, which is the proportion of segmented kiwifruit calyx number in visible kiwifruit number.

The Image-Fusion and Feature-Fusion modes which fused the RGB and NIR images have achieved better detection performance while maintaining similar detection speed to single modality modes (0.134 s/image). Comparing with Fu *et al.* [22b] that has the highest DR of kiwifruit in current state-of-the-art kiwifruit detection methods, Image-Fusion mode has a DR of 91.7%, meaning that the Image-Fusion has acceptable overall performance of detection speed and detection rate.

TABLE 4. Summary of proposed methods and other deep learning networks for kiwifruit image detection.

	Deep learning networks	Image type	Camera	Image resolution (pixel)	Detection rate (%)	Detection speed (s/image)
Williams *et al.* (2019) [6]	FCN-8S	RGB	Baslar ac1920-40uc	1900×1200	81.6%	3.0
Fu *et al.* (2018a) [21]	LeNet	RGB	Canon S110	2352×1568	89.3	8.1
Fu *et al.* (2018b) [22]	ZFNet	RGB	Canon S110	2352×1568	92.3	0.274
RGB-Only mode	VGG16	RGB	Kinect v2	512×424	88.1	0.134
NIR-Only mode	VGG16	NIR	Kinect v2	512×424	89.0	0.135
Image-Fusion mode	VGG16	RGB, NIR	Kinect v2	512×424	91.7	0.134
Feature-Fusion mode	VGG16	RGB, NIR	Kinect v2	512×424	90.8	0.188

On the other hand, Fu *et al.* [22b] only worked on daytime images, while images from night were not considered. In terms of kiwifruit categories, the DRs of occluded and overlapped kiwifruit detected by the Image-Fusion mode have exceeded the current highest DR, which are 83.4% and 87.5% respectively, as shown in Fu *et al.* [22b]. In terms of that, the Image-Fusion mode has better detection performance in kiwifruit categories that are difficult to be detected.

It is hard to summary the performance of Williams *et al.* [6], because DR and accurate positioning are both important indicators for realizing automated picking of kiwifruit. However, in detecting images of almost the same resolution, the speed of the Image-Fusion mode is about twice than that of Williams *et al.* [6]. Besides, Kinect v2 can obtain location information directly from the depth images. This can reduce computing time comparing with that using stereo vision provided by two RGB cameras. Overall, on the basis of maintaining rapid detection, the Image-Fusion mode can achieve state-of-the-art detection performance.

Rapid and accurate kiwifruit detection is the basis of realizing kiwifruit automated harvesting. Improved methods with RGB and NIR images fused in the input layer achieved a more efficient and accurate detection of kiwifruit. Therefore, this research may provide not only the technique to align RGB and NIR images and implement image fusion for deep learning, but also the knowledge of the Image-Fusion mode performs better.

IV. CONCLUSION

This work presents a novel methodology for kiwifruit image detection using an RGB-D sensor, taking advantage of its extended capabilities. Multi-modality images dataset was built using images provided by Microsoft Kinect v2. A 2D projection of 3D point cloud based on depth sensor was carried out to obtain corresponding RGB images. Then, an alignment between different modalities was performed, obtaining images with 3 modalities: RGB, depth, and NIR. The Hayward-Kiwi RGB-NIR-D dataset and the corresponding annotations is the first dataset for kiwifruit detection that contains aligned RGB, depth, and NIR images which has been made publicly available. Two methods with RGB and NIR images fusion were used to evaluate the fusion modes: Image-Fusion and Feature-Fusion. VGG16 is modified and adapted to receive RGB and NIR information simultaneously in the Image-Fusion mode. The input RGB and NIR images are fed to two VGG16 networks and combine the feature map in the Feature-Fusion mode. Results showed that an improvement of 2.1% AP in the Feature-Fusion mode, and 2.3% AP in the Image-Fusion mode when fusing RGB and NIR images. Besides, the detection speed of the Image-Fusion mode is the fastest, which is 0.134 s/image. The study proposes an improved method for achieving more accurate and faster automated detection of kiwifruit, which provide the knowledge of the Image-Fusion mode performs better. In future research, the depth images in Hayward-Kiwi RGB-NIR-D dataset or other spectral images of kiwifruit will be considered to be fused with RGB and NIR images, thereby further realizing the detection and localization of kiwifruit simultaneously.

ACKNOWLEDGMENT

The authors would like to express their great thanks to Dr. A. Al-Mallahi from the Dalhousie University in Canada for his advice and support to the manuscript writing and English editing.

REFERENCES

[1] D. Brown and R. Hermes, "The food and agriculture organization of the UN and Asian LMEs: A commentary," *Deep Sea Res. II, Top. Stud. Oceanogr.*, vol. 163, pp. 124–126, May 2019.
[2] F. Hu, "Fertilization evaluation of kiwifruit in Guanzhong region of Shaanxi province," *Soils Fertilizers Sci. China*, vol. 54, no. 3, pp. 44–49, 2017.
[3] L. Fu, B. Wang, Y. Cui, S. Su, Y. Gejima, and T. Kobayashi, "Kiwifruit recognition at nighttime using artificial lighting based on machine vision," *Int. J. Agricult. Biol. Eng.*, vol. 8, no. 4, pp. 52–59, Aug. 2015.
[4] H. Huang and A. R. Ferguson, "Review: Kiwifruit in China," *New Zealand J. Crop Horticult. Sci.*, vol. 29, no. 1, pp. 1–14, Mar. 2001.
[5] L. Mu, H. Liu, Y. Cui, L. Fu, and Y. Gejima, "Mechanized technologies for scaffolding cultivation in the kiwifruit industry: A review," *Inf. Process. Agricult.*, vol. 5, no. 4, pp. 401–410, Dec. 2018.

[6] H. A. Williams, M. H. Jones, M. Nejati, M. J. Seabright, J. Bell, N. D. Penhall, J. J. Barnett, M. D. Duke, A. J. Scarfe, H. S. Ahn, J. Lim, and B. A. Macdonald, "Robotic kiwifruit harvesting using machine vision, convolutional neural networks, and robotic arms," *Biosyst. Eng.*, vol. 181, pp. 140–156, May 2019.

[7] A. Silwal, J. R. Davidson, M. Karkee, C. Mo, Q. Zhang, and K. Lewis, "Design, integration, and field evaluation of a robotic apple harvester," *J. Field Robot.*, vol. 34, no. 6, pp. 1140–1159, Sep. 2017.

[8] A. Bechar and C. Vigneault, "Agricultural robots for field operations: Concepts and components," *Biosyst. Eng.*, vol. 149, pp. 94–111, Sep. 2016.

[9] Y. Cui, S. Su, X. Wang, Y. Tian, P. Li, and F. Zhang, "Recognition and feature extraction of kiwifruit in natural environment based on machine vision," *Trans. Chin. Soc. Agricult. Eng.*, vol. 44, no. 5, pp. 247–252, 2013.

[10] J. Mu, J. Chen, G. Sun, F. Liu, Y. Ma, and F. Wang, "Characteristic parameters extraction of kiwifruit based on machine vision," *J. Agricult. Mech. Res.*, vol. 36, no. 6, pp. 138–142, 2014.

[11] L. Fu, S. Sun, V. A. Manuel, S. Li, R. Li, and Y. Cui, "Kiwifruit recognition method at night based on fruit calyx image," *Trans. Chin. Soc. Agricult. Eng.*, vol. 33, no. 2, pp. 199–204, 2017.

[12] L. Fu, E. Tola, A. Al-Mallahi, R. Li, and Y. Cui, "A novel image processing algorithm to separate linearly clustered kiwifruits," *Biosyst. Eng.*, vol. 183, pp. 184–195, Jul. 2019.

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[14] H. H. Aghdam, E. J. Heravi, and D. Puig, "A practical approach for detection and classification of traffic signs using convolutional neural networks," *Robot. Auto. Syst.*, vol. 84, pp. 97–112, Oct. 2016.

[15] L. Zhang, G. Gui, A. M. Khattak, M. Wang, W. Gao, and J. Jia, "Multi–task cascaded convolutional networks based intelligent fruit detection for designing automated robot," *IEEE Access*, vol. 7, pp. 56028–56038, 2019.

[16] X. Liu, D. Zhao, W. Jia, W. Ji, and Y. Sun, "A detection method for apple fruits based on color and shape features," *IEEE Access*, vol. 7, pp. 67923–67933, 2019.

[17] W. Tan, C. Zhao, and H. Wu, "Intelligent alerting for fruit-melon lesion image based on momentum deep learning," *Multimedia Tools Appl.*, vol. 75, no. 24, pp. 16741–16761, Dec. 2016.

[18] Y. Tian, G. Yang, Z. Wang, H. Wang, E. Li, and Z. Liang, "Apple detection during different growth stages in orchards using the improved YOLO–V3 model," *Comput. Electron. Agricult.*, vol. 157, pp. 417–426, Feb. 2019.

[19] M. Stein, S. Bargoti, and J. Underwood, "Image based mango fruit detection, localisation and yield estimation using multiple view geometry," *Sensors*, vol. 16, no. 11, p. 1915, Nov. 2016.

[20] Y. Yu, K. Zhang, L. Yang, and D. Zhang, "Fruit detection for strawberry harvesting robot in non-structural environment based on mask–RCNN," *Comput. Electron. Agricult.*, vol. 163, Aug. 2019, Art. no. 104846.

[21] L. Fu, Y. Feng, T. Elkamil, Z. Liu, R. Li, and Y. Cui, "Image recognition method of multi-cluster kiwifruit in field based on convolutional neural networks," *Trans. Chin. Soc. Agricult. Eng.*, vol. 34, no. 2, pp. 205–211, 2018.

[22] L. Fu, Y. Feng, Y. Majeed, X. Zhang, J. Zhang, M. Karkee, and Q. Zhang, "Kiwifruit detection in field images using faster R–CNN with ZFNet," *IFAC-Papers OnLine*, vol. 51, no. 17, pp. 45–50, 2018.

[23] S. Bargoti and J. P. Underwood, "Image segmentation for fruit detection and yield estimation in apple orchards," *J. Field Robot.*, vol. 34, no. 6, pp. 1039–1060, Sep. 2017.

[24] N. Häni, P. Roy, and V. Isler, "A comparative study of fruit detection and counting methods for yield mapping in apple orchards," *J. Field Robot.*, Aug. 2019.

[25] I. Sa, Z. Ge, F. Dayoub, B. Upcroft, T. Perez, and C. Mccool, "DeepFruits: A fruit detection system using deep neural networks," *Sensors*, vol. 16, no. 8, p. 1222, Aug. 2016.

[26] H. Zhan, X. Li, Z. Zhou, C. Wang, and Y. Gao, "Detection of chestnut defect based on data fusion of near-infrared spectroscopy and machine vision," *Trans. Chin. Soc. Agricult. Eng.*, vol. 27, no. 2, pp. 345–349, 2011.

[27] A. M. Abdelsalam and M. S. Sayed, "Real-time defects detection system for orange citrus fruits using multi-spectral imaging," in *Proc. IEEE 59th Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Oct. 2016.

[28] S. Zhang, R. Wu, K. Xu, J. Wang, and W. Sun, "R-CNN-based ship detection from high resolution remote sensing imagery," *Remote Sens.*, vol. 11, no. 6, p. 631, Mar. 2019.

[29] X. Bai, C. Wang, and C. Li, "A streampath–based RCNN approach to ocean eddy detection," *IEEE Access*, vol. 7, pp. 106336–106345, 2019.

[30] K. Hameed, D. Chai, and A. Rassau, "A comprehensive review of fruit and vegetable classification techniques," *Image Vis. Comput.*, vol. 80, pp. 24–44, Dec. 2018.

[31] W. Song, A. V. Le, S. Yun, S.-W. Jung, and C. S. Won, "Depth completion for Kinect v2 sensor," *Multimedia Tools Appl.*, vol. 76, no. 3, pp. 4357–4380, Feb. 2017.

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

[33] S. Tu, Y. Xue, C. Zheng, Y. Qi, H. Wan, and L. Mao, "Detection of passion fruits and maturity classification using red-green-blue depth images," *Biosyst. Eng.*, vol. 175, pp. 156–167, Nov. 2018.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[35] A. Yang, Y. Xue, H. Huang, N. Huang, X. Tong, X. Zhu, X. Yang, L. Mao, and C. Zheng, "Lactating sow image segmentation based on fully convolutional networks," *Trans. Chin. Soc. Agricult. Eng.*, vol. 33, no. 23, pp. 219–225, 2017.

[36] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2014, pp. 675–678.

[37] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep./Oct. 2015, pp. 681–687.

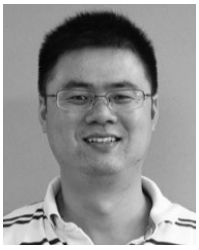
[38] H. Long, Y. Chung, Z. Liu, and S. Bu, "Object detection in aerial images using feature fusion deep networks," *IEEE Access*, vol. 7, pp. 30980–30990, 2019.

[39] X. Xu, Y. Li, G. Wu, and J. Luo, "Multi-modal deep feature learning for RGB–D object detection," *Pattern Recognit.*, vol. 72, pp. 300–313, Dec. 2017.

[40] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 379–387.

ZHIHAO LIU is currently pursuing the master's degree with the College of Mechanical and Electronic Engineering, Northwest A&F University, China. He conducted research on deep learning for classification, object recognition, tracking, and detection for the vision system of agricultural robot. He also studies image / video processing techniques applied on mobile phone.

JINGZHU WU received the Dr.Eng. degree in agricultural electrification and automation from China Agricultural University, Beijing, China, in 2006. From 2006 to 2010, she was a Lecturer with the School of Computer and Information Engineering, Beijing Technology and Business University, where she has been an Associate Professor, since 2010. She was a Visiting Scholar with the Department of Engineering, University of Washington, Seattle, USA, in 2013. She is currently engaged in the research of agricultural products quality detection technology based on spectroscopy and spectral imaging.



LONGSHENG FU received the B.S. degree in electronical and information engineering and the M.S. degree in agricultural electrification and automation from China Agricultural University, Beijing, China, in 2006 and 2008, respectively, and the Ph.D. degree in environmental resources from Hokkaido University, Sapporo, Japan, in 2012. He is currently an Adjunct Faculty with the Center for Precision and Automated Agricultural Systems, Washington State University, Prosser, WA, USA, and an Associate Professor with Northwest A&F University, Yangling, China. He is an Editorial Advisory Board of the journal *Computers and Electronics in Agriculture* and awarded ''Outstanding Contribution in Reviewing'' of the journal *Computers in Industry*, *Biosystems Engineering*, and *Computers and Electronics in Agriculture*.

YAQOOB MAJEED was born in Pakistan, in 1991. He received the B.S. degree in mechatronics and control engineering from the University of Engineering and Technology, Lahore, Pakistan, in 2013. He is currently pursuing the Ph.D. degree with the Department of Biological Systems Engineering, Washington State University, USA. He was working as a Lecturer with the University of Agriculture, Faisalabad, Pakistan. His research interests include agricultural robotics, machine vision, machine learning, deep learning, and artificial intelligence.

YALI FENG received the M.S. degree from the College of Mechanical and Electronic Engineering, Northwest A&F University, China, in 2019. She is currently being engaged in teaching and research with the College of Engineering, Shanxi Agricultural University, China. She conducted research on deep learning for classification, object recognition, and detection for the vision system of agricultural robot.

RUI LI received the B.S. degree from China Agricultural University, Beijing, China, in 2009, the M.S. degree from Hokkaido University, Sapporo, Japan, in 2012, and the Ph.D. degree from Northwest A&F University, Yangling, China, in 2019. She is currently a Technician with Northwest A&F University. She received the Excellent Doctoral Graduate Award.

YONGJIE CUI received the B.S. degree from China Agricultural University, Beijing, China, in 1993, and the M.S. degree from the University of Miyazaki, Japan, in 2003, and the Ph.D. degree from Kagoshima University, Japan, in 2006. He was a Patent Consultant with Miyazaki Technology Licensing Organization, Japan. He is currently a Full Professor with Northwest A&F University, Yangling, China.