Computers and Electronics in Agriculture 209 (2023) 107814

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Phenotyping of individual apple tree in modern orchard with novel smartphone-based heterogeneous binocular vision and YOLOv5s

Guanao Zhao [a], Ruizhe Yang [a], Xudong Jing [a], Haosen Zhang [a], Zhenchao Wu [a], Xiaoming Sun [a], Hanhui Jiang [a], Rui Li [a,d], Xiaofeng Wei [a,*], Spyros Fountas [e], Huijun Zhang [f], Longsheng Fu [a,b,c,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] Northwest A&F University Shenzhen Research Institute, Shenzhen, Guangdong 518000, China
[e] Agricultural University of Athens, Athens 11855, Greece
[f] College of Environmental Resources, Chongqing Technology and Business University, Chongqing 400067, China

ARTICLE INFO

Keywords:
Smartphone-based heterogeneous binocular vision
Deep learning
Phenotyping
Android smartphone
Modern apple orchard

ABSTRACT

Phenotyping plays a significant role in the breeding of apple tree. However, existing researches mainly relied on instruments, such as LiDAR, RGB-D camera or UAV (unmanned aerial vehicle) embedded with depth sensor, etc., which requires additional costs for users and also inconvenient. Therefore, a novel method of smartphone-based heterogeneous binocular vision was developed to fulfill low-cost automated phenotyping for apple tree. In this study, a pair of cameras on multi-camera smartphone was selected to obtain heterogeneous binocular camera. After that, a so-called virtual focal method was developed to generate standard binocular images from heterogeneous binocular images of smartphone. A well-known YOLOv5s object detection model was trained on a four-class dataset to detect fruits, grafts, trunks and whole trees. Then, the model was simplified to fit the deployment on smartphone. Finally, five phenotypes (trunk diameter, ground diameter, tree height, fruit vertical diameter, and fruit horizontal diameter) of individual apple tree were obtained by pinhole camera model and standard binocular vision. After evaluation of phenotyping manually and by smartphone, our method shows *MAPE* (mean average percentage error) ranging from 6.00 % to 13.73 % for the five phenotypes. Compared with the existing studies, our method has reached a close or even better phenotyping accuracy with only a smartphone. As more and more smartphones have multi-camera, our method is probably the lowest cost phenotyping method for most of the potential users. Results indicated that the approach could be utilized to phenotyping of apple tree.

1. Introduction

Phenotyping for apple tree plays a positively promoting role in improving yield of apple. However, the most common-use traditional labor-intensive phenotyping is high-cost and inefficient. Besides, accuracy of traditional labor-intensive phenotyping by different workers can result in different results, which are also easily affected by weather limitations (Westling et al., 2021; Wu et al., 2022). To fulfill low-cost and efficient phenotyping for plants like apple tree, methods with various sensors were developed (Bauer et al., 2019).

As an alternative, aerial phenotyping methods by using instruments such as UAV (unmanned aerial vehicle) embedded with RGB-D (red, green, blue-depth) camera, multi-spectral camera or LiDAR (light detection and ranging) were developed as a more automated and efficient substitute of traditional labor-intensive phenotyping. Most of the aerial phenotyping methods basically uses UAV to phenotyping in a large scale (Corte et al., 2020; Hobart et al., 2020; Krisanski et al., 2020; Liu et al., 2020; Machimura et al., 2021; Nasiri et al., 2021; Solvin et al., 2020). Chang et al. (2021) developed a UAV system to collect canopy cover, volume and vegetation indices from RGB and multi-spectral images to estimate tomato yield. Herzig et al. (2021) employed a UAV embedded with RGB and multi-spectral sensor to estimate canopy height, vegetation cover and growth dynamics traits to test and compare the use of conventional and new traits for barley breeding. Besides,

* Corresponding authors at: College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling, Shaanxi 712100, China (L.F.).
E-mail addresses: wxf8412@nwafu.edu.cn (X. Wei), fulsh@nwafu.edu.cn (L. Fu).

https://doi.org/10.1016/j.compag.2023.107814
Received 17 January 2023; Received in revised form 25 March 2023; Accepted 27 March 2023
Available online 12 April 2023

UAV-based phenotyping methods with LiDAR were also studied to fulfill phenotyping for tree height, crown width (Liao et al., 2022; Maesano et al., 2020; ten Harkel et al., 2020). Although aerial methods achieved high-throughput and automated phenotyping for plants, only phenotypes that can be collected from top view are available for these methods. In other words, phenotypes that can only be measured on the ground are ignored in these methods, such as DBH (diameter at breast height, namely, trunk diameter) and fruit size.

Furthermore, some terrestrial phenotyping methods were also developed to measure phenotypes that can only be taken on the ground. These methods are mostly take extra instruments like LiDAR, standard binocular camera or other sensors to accomplish (Berk et al., 2020; Fan et al., 2022; Gené-Mola et al., 2021; Shao et al., 2022; Sun et al., 2022; Li et al., 2020). Fan et al. (2022) designed a phenotyping robot embedded with RGB-D camera to estimate maize diameter. Sun et al. (2022) applied Kinect v2 RGB-D camera to measure grafted diameter of apple tree trunk with *MAE* (mean absolute error) of 3.01 mm and *MAPE* (mean absolute percentage error) of 5.86 %. Chen et al. (2020) designed a high-throughput stereo vision system for field-based plant phenotyping, which achieved a *MAE* of 1.44 mm for sorghum stem diameter estimation. These methods are high-cost and inconvenient since their expensive instruments.

With the development of technology, more and more sensors were embedded in smartphones, which has made low-cost phenotyping based on smartphone possible. For example, smartphones with ToF (Time of Flight) depth sensor were launched in recent years to capture point cloud of scene. However, ToF depth sensor is not a standard specification of smartphones, which is only embedded in a few smartphones. Besides, ToF depth sensors of smartphone are not as good because of low power supply of smartphone, which leads to low quality of point cloud and low working distance. On the contrary, multi-camera is now nearly the common specification for smartphone. Although multi-camera of smartphone usually consist of cameras with different specifications that cannot obtained depth by standard binocular vision, there are still some researches proved that it is possible to obtain depth of the scene by heterogeneous binocular camera (Baek and Kim, 2016; Kim, 2013, 2021; Lin et al., 2021; Sengupta, 1997). According to the potential of heterogeneous binocular camera, it is likely to develop a low-cost and automated phenotyping method with it.

Therefore, a phenotyping method for individual apple tree of modern apple orchard with novel smartphone-based heterogeneous binocular vision and YOLOv5s (You Only Look Once v5s) was developed to reach low-cost and easy-to-use phenotyping (Zhao et al., 2021). A well-known YOLOv5s object detection model was trained and deployed in Android smartphone to detect fruits, trunks, grafts and whole trees in heterogeneous binocular images. After that, virtual focal method was developed to generate standard binocular images from heterogeneous binocular images. Depth of center points of detected phenotypes was calculated by standard binocular images. Finally, pinhole camera model was applied to fulfill apple tree phenotyping.

2. Materials and methods

2.1. Dataset acquisition

This study focuses on phenotyping of individual apple tree in modern apple orchard. Heterogeneous binocular images were captured in two modern apple orchards, including Haisheng modern apple orchard (Fig. 1a, 107°53′58″E, 34°29′07″N, 715 m in altitude, mainly cultivated Gala apple), Baoji City, Shaanxi province, China, and Liangshan modern apple orchard (Fig. 1b, 108°09′17″E, 34°38′19″N, 962 m in altitude, mainly cultivated Mamu apple), Xianyang City, Shaanxi province, China. These two orchards were built with row spacing of 3.50 m and appropriate tree spacing to separate apple trees in image.

The following data of 110 apple trees were collected from the two orchards (55 apple trees for each orchard) to fit the needs of the study: (1) heterogeneous binocular images (Fig. 2), which were captured by heterogeneous binocular camera of three different Android smartphones (Realme GT neo2, Xiaomi 10 and Huawei P30) with maximum imaging resolution. (2) The trunk diameter (measured at 10 cm above the grafting position), tree height (from root of tree to top of the highest branch) and ground diameter of apple tree, were measured by vernier caliper or tape measure. (3) The diameters of 120 apple fruits (60 apple fruits for each orchard) of random 24 apple trees selected from the 110 apple trees were measured by vernier caliper. The location and occlusion of measured apples are totally random. All the images were captured.

The manually measured phenotypes of apples and apple trees were adopted as the ground truth phenotypes. And they were measured multiple times to obtain reliability. Besides, 300 apple tree images were captured and labeled to train YOLOv5s model. All the images in this study were captured in the front view of the tree row by using the system camera APP. When capturing the heterogeneous binocular image of apple trees, Android smartphone was firstly fixed on a tripod. Then, the Android smartphone was set up on an angle to make the apple tree center in the camera preview. Finally, a manual operation record (a set of touches, can be record or replay by using specific APP, such as RepetiTouch and Touch Recorder) was replay to capture heterogeneous binocular image to avoid potential negative effects to the image by manual operation.

Fig. 1. Planting pattern of two apple orchards. (a) Haisheng modern apple orchard, captured in 24th, June 2022 (Cloudy, gentle breeze from northeast). (b) Liangshan modern apple orchard, captured in 09th, August 2022 (Overcast with light rain, gentle breeze from east).

Fig. 2. Heterogeneous binocular images for the same apple tree captured by Realme GT neo2. (a) Image captured with ultra-wide camera. (b) Image captured with main camera.

2.2. Image dataset

The five phenotypes (trunk diameter, ground diameter, tree height, fruit vertical diameter, and fruit horizontal diameter of apple trees) of 300 apple tree images were labeled as four-class ('trunk', 'graft', 'whole tree' for the former three phenotypes, while 'fruit' for the diameters of apple fruits) using different label strategies. The fruit and graft were labeled with their bounding boxes, while apple tree was labeled with bounding box of the whole trunk. Each type of label is depicted in different colors, as shown in Fig. 3. Furthermore, label strategy of trunks depends on their inclination, which are demonstrated in Fig. 4.

A special subdivided label strategy was applied to make trained YOLOv5s model detect trunks of different inclinations with less background inside the detected bounding box. For those vertical trunks with small inclinations, rectangles with aspect ratio of about 2:1 were applied to label it. For inclined trunks, rectangles that contains background as less as possible were applied to label it. All the mentioned trunk labels should have a similar size and aspect ratio.

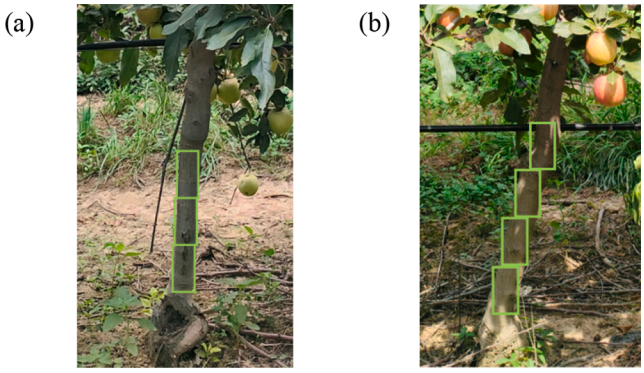

Fig. 4. Labels of trunks with different inclinations (green rectangles in images). (a) Vertical trunk. (b) Inclined trunk.

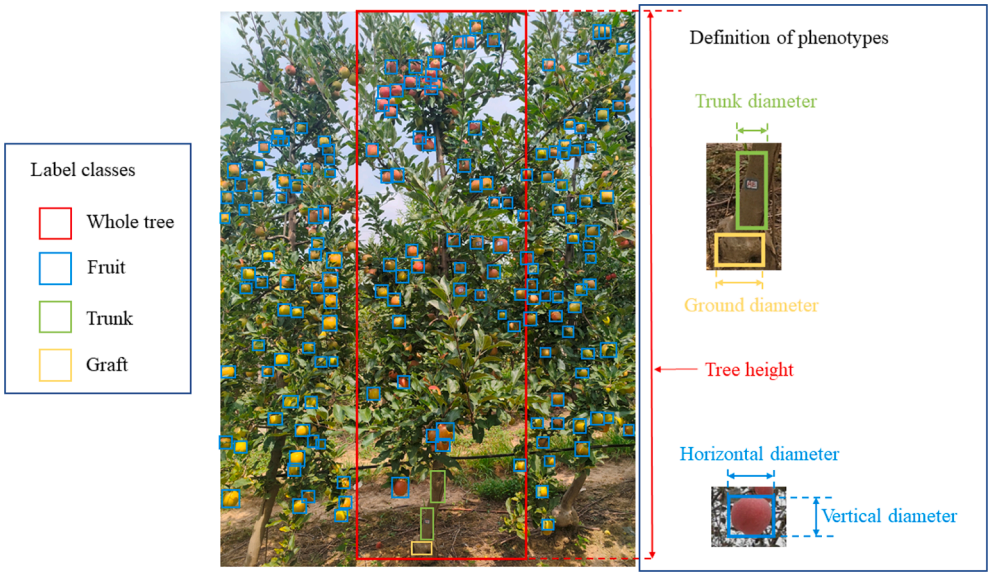

Fig. 3. The labels and definitions of apple tree phenotypes.

Afterwards, data augmentation was adopted to improve variety of the labeled dataset. For image capturing, lenses of the heterogeneous binocular camera are too difficult to capture the same scene with identical aperture, exposure time and shutter speed, etc. Even if they can fulfill that, the scene will still be captured with different illuminations (Fig. 5) because of their hardware differences (focal, imaging resolution, etc.). Thus, illumination change (with a brightness range from 0.65 to 1.2), rotation (with a 30° rotation of clockwise or anti-clockwise) and contrast change (with histogram normalization) were utilized to augment the training dataset randomly, each of the augmentation step generates 300 images (every augmentation step will only apply to an image once). This improved the robustness of trained model to better detect objects with different illuminations, inclinations and colors. The entire dataset comprised 900 augmented images and 300 original images. Out of these, all the augmented images and 225 original images were used as the training dataset, while the remaining 75 original images were used as the validation dataset.

2.3. YOLOv5s model

The well-known object detection model YOLOv5s was adopted to detect phenotypes of apple tree from the heterogeneous binocular images. YOLOv5s is the model which has the second least layer and second fastest inference speed among YOLOv5 series. And the model was training on the augmented dataset under PyTorch framework with training parameters specified in Table 1.

The model was training on a laptop embedded with Nvidia RTX 2060 (6 GigaByte), Intel Core i5 9300HQ and 16 GigaByte memory. Besides, CUDA (Compute Unified Device Architecture) 11.7 and PyTorch 1.12.1 were also installed to fit the needs of model training. After training, the trained model was converted to format of NCNN framework (an Android-based deep learning inference framework by Tencent) for deployment on Android smartphone, which also contains a model simplification by ONNX framework (Open Neural Network Exchange, a deep learning framework by Facebook and Microsoft).

2.4. Generation of standard binocular images from heterogeneous binocular images by virtual focal

The virtual focal method was developed to generate standard binocular images from heterogeneous binocular images of smartphone. At first, differences between heterogeneous binocular vision and standard binocular vision should be considered. The depth equation of standard binocular vision is described in Eq. (1).

Table 1
Training parameters of YOLOv5s.

Parameters	Learning rate	Momentum	Weight decay	Batch size
Value	0.01	0.937	0.0005	4

$$D = \frac{PixF \times B}{Disp} \tag{1}$$

Where D, $PixF$, B and $Disp$ are depth, pixel focal (focal in the unit of pixel) of standard binocular camera, baseline of binocular camera and disparity of feature points, respectively. Their units are millimeter, pixel, millimeter and pixel, respectively. However, unlike standard binocular camera, focal of two lenses of heterogeneous binocular camera are unequal. Therefore, identical pixel focal of the two lenses of heterogeneous binocular camera is essential to obtain depth by standard binocular vision. Besides, the disparity of standard binocular vision is calculated under identical imaging resolution, which indicates that the heterogeneous binocular vision lack of consistency of imaging resolution. Hence, the identical imaging resolution is also needed to obtain depth in addition to the identical pixel focal.

A general imaging model of camera, namely, pinhole camera model was considered as the basis of the virtual focal. The simplified pinhole camera model is demonstrated in Fig. 6. It is obvious that focal of camera is inversely linear proportional to FoV (Field of View), while CMOS (Complementary Metal Oxide Semiconductor, the most common-use imaging sensor of smartphone) size of camera is linear proportional to it. As mentioned, the identical pixel focal and the identical imaging resolution (imaging resolution is linear proportional to CMOS size) are

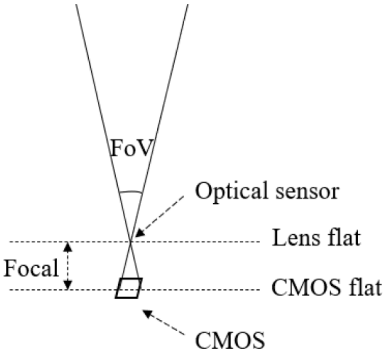

Fig. 6. Pinhole camera model.

Fig. 5. A pair of heterogeneous binocular image with different illumination for the same scene by Realme GT neo2. (a) Image captured by main camera. (b) Image captured by ultra-wide camera, which is darker.

essential for standard binocular vision. However, the heterogeneous binocular camera in this study was composed of two cameras of smartphone with different pixel focal and CMOS size, result to different FoV and imaging resolution, while the standard binocular camera captured binocular images with identical FoV and imaging resolution. The FoV relation between ultra-wide camera and main camera is demonstrated in Fig. 7, in which the FoV of main camera is smaller than ultra-wide camera. The comparison in Fig. 7 also applies to the most common smartphones that with at least one main camera and one ultra-wide camera.

Generation of a pair of binocular images captured by binocular camera with the same pixel focal and imaging resolution from the heterogeneous binocular image was fulfilled by the virtual focal. According to the linear proportions between pixel focal, CMOS size and FoV, the identical pixel focal between two lenses of the heterogeneous binocular camera can be made by only scale image of one of the lenses (ultra-wide in this study). The image scale operation indicates the identical scaling change of pixel focal and CMOS size of the lens. When pixel focal and CMOS size of a lens scaled at the same proportion, the FoV of image will not change because the influences they made have been offset by each other. At the same time, the image will be scaled the same proportion as the change of CMOS size. The scaling proportion to make an identical pixel focal between the heterogeneous binocular camera is *VFSF* (virtual focal scaling factor), whose definition is described in Eq. (2).

$$VFSF = \frac{PixF_m}{PixF_{uw}} = \frac{F_m}{PSoP_m} \div \frac{F_{uw}}{PSoP_{uw}} \tag{2}$$

Where $PixF_{\mathrm{m}}$ and $PixF_{\mathrm{uw}}$ are pixel focal of main camera and ultra-wide camera, respectively; $PSoP_{\mathrm{m}}$ and F_{m} indicate physical size of individual pixel in main camera and physical focal of main camera (focal in the unit of millimeter), respectively, so do $PSoP_{\mathrm{uw}}$ and F_{uw}. These parameters can be read from the system of smartphone. In this study, the *VFSF* was applied to scale ultra-wide image to align pixel focal of ultra-wide camera with pixel focal of main camera, which basically achieved identical pixel focal between lenses of the heterogeneous binocular camera. The hardware parameters and *VFSF* of the three smartphones are listed in the Table 2.

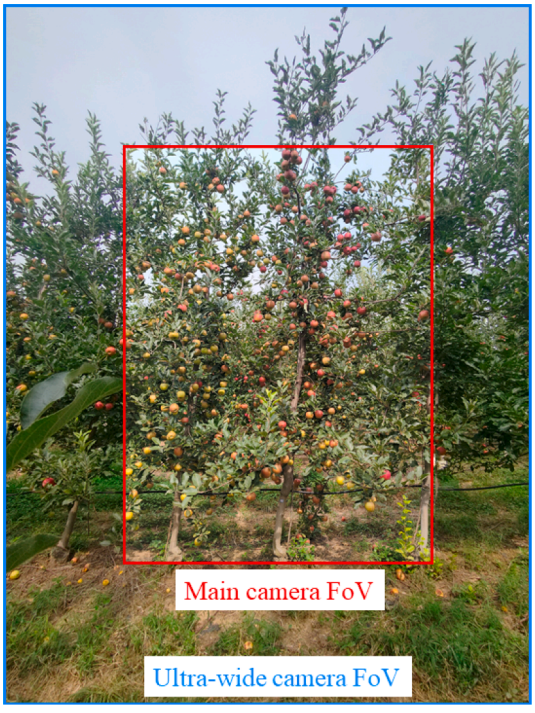

Fig. 7. Comparison of FoV between main camera and ultra-wide camera of Realme GT neo2.

After scaling the ultra-wide image, an image (scaled ultra-wide image) captured with pixel focal of main camera and CMOS size of scaled ultra-wide camera (namely, ultra-wide camera with virtual focal) was obtained. However, imaging resolution of the scaled ultra-wide image and the image of main camera are still different due to the different CMOS size. Therefore, the scaled ultra-wide image and the image of main camera are regarded as images captured by cameras with same pixel focal but different CMOS size, which still cannot be deemed as a pair of standard binocular camera.

Finally, an image which has the same resolution as image captured by main camera can be generated from cropping the scaled ultra-wide image. As is known to all, resolution of image is only depending on CMOS. If physical size of individual pixel is stable, the smaller CMOS is, the smaller resolution it has. Hence, it is feasible to minify CMOS size of ultra-wide camera with the virtual focal to make its CMOS size the same with CMOS of main camera. The image captured by CMOS size minified ultra-wide camera with virtual focal can be straightly obtained by cropping the scaled ultra-wide image. The cropped scaled ultra-wide image should have the same resolution as the image of main camera. Thus, an image that captured by a camera which has the same focal and CMOS size as main camera was obtained from original ultra-wide image, which has the same resolution and FoV as the image captured by main camera. A comparison of the cropped scaled ultra-wide image and original image of the heterogeneous binocular camera was demonstrated in Fig. 8.

2.5. Depth estimation by standard binocular vision

After generation of standard binocular images, the standard binocular vision is applied to estimate depth of detected objects. It uses the cropped scaled ultra-wide image (Fig. 8c) and the image captured by main camera (Fig. 8a) to estimate depth. Then a series of operations based on template matching were adopted to match detected objects and calculate disparity to quickly estimate depth of detected objects.

Template matching was to find the most similar region of matching template in target image. In this study, the detected objects of image captured by main camera were regarded as the matching template, while the cropped scaled ultra-wide image was regarded as the target image. Before that, these images were scaled to half of their original size to fulfill a faster template matching. A limitation based on epipolar constraint was adopted to limit the matching area (Li et al., 2021). The detected objects (or matching templates) with higher matching costs should be ignored to get a better performance. Finally, disparity of a detected object can be calculated by its coordinate of center point of original detected bounding box (in image captured by main camera) and coordinate of center point of matching bounding box in the cropped scaled ultra-wide image. After that, depth of detected objects can be determined by principle of standard binocular vision in Eq. (1).

Table 2
Hardware parameters and *VFSF* of the three smartphones.

Smartphones		Realme GT neo2	Huawei P30	Xiaomi 10
Focal (mm)	Main	5.58	5.58	6.72
	Ultra-wide	1.66	2.26	2.13
Imaging resolution (pixel × pixel)	Main	6944 × 9280	5472 × 7296	9024 × 12032
	Ultra-wide	2448 × 3264	3456 × 4608	3120 × 4208
Pixel Size (μm)	Main	0.8	0.8	0.8
	Ultra-wide	1.12	1	1.12
VFSF		4.706	3.086	4.417

(a) (b) (c)

Fig. 8. Images captured by heterogeneous binocular camera of smartphone (take the Realme GT neo2 as example) and the cropped scaled ultra-wide image. (a) Image captured by main camera of smartphone. (b) Image captured by ultra-wide camera of smartphone. (c) The cropped scaled ultra-wide image, which has the same FoV and resolution as image captured by main camera (Fig. 8a).

2.6. Phenotyping by pinhole camera model

Pinhole camera model was adopted to ulfil phenotyping of apple tree. To the best of our knowledge, most cameras are build based on the pinhole camera model (Fig. 9), which is the general imaging model of camera. In Fig. 9, there is a pair of similar triangles between apple and its image, where an equation (Eq. (3) was developed to describe relations between them.

$$\frac{PS}{IPS} = \frac{D}{F} \tag{3}$$

Where *PS* is physical size of apple in real world, which is unknown; *IPS* is physical size of apple image in CMOS, which can be calculated by pixel size of apple in image and *PsoP*; *D* is depth, which can be estimated by the heterogeneous binocular vision; *F* is focal of camera, which can be straightly read from system of smartphone. The Eq. (3) usually works no matter what kind of object is in Fig. 9. After that, three parameters of the equation are obtained to calculate the *PS* by Eq. (4).

$$PS = \frac{D \times IPS}{F} \tag{4}$$

Besides, angle of inclination was applied to correct measured trunk diameter, which is described in Fig. 10. The angle (α in Fig. 10) between yellow line (which connects center points of the lowest two detect bounding boxes of trunk) and red vertical dash line is angle of inclination. Perpendicular lines of the two lines are correct trunk diameter (orange line), measured trunk diameter (red horizontal line), respectively. The correct value of trunk diameter should be about cosine α times of the measured trunk diameter. Moreover, for those trunks with only one detect bounding box, the correction should be ignored.

2.7. Alternative schemes for errors in detecting or matching

Dozens of alternative schemes were developed to avoid errors in object detection or template matching that might lead to mistakes or interruption. The potential errors mainly consist of two situations as following:

(1) Some phenotypes are not detected, whose bounding box is nothing.

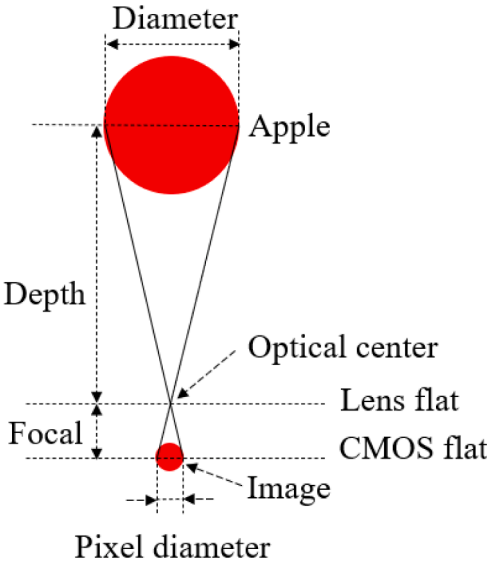

Fig. 9. A simplified pinhole camera model.

Fig. 10. Definition of angle of inclination.

(2) Some phenotypes are not well matched, which means that its phenotype is hard to be estimated because of its wrong depth.

There are several methods to fulfill phenotyping for phenotypes that have not been detected. If a trunk or a graft is not detected, then its diameter can be estimated by the known diameter of graft or trunk, respectively. That is, the mean ratio between trunk diameter and graft diameter of the ground truth dataset is about 2.18 (might be different for different period of different apple tree). For the other phenotypes that are not detected, it is hard to make a logical strategy to deal with their situation. Furthermore, for phenotypes that not well matched (usually comes with large matching cost), their depth can be straightly use depth of its closest correct matching object instead.

2.8. Evaluation criteria

The *MAE*, *MAPE* and *RMSE* (Root Mean Square Error) were adopted to evaluate performance of phenotyping by smartphone-based heterogeneous binocular vision, which were defined by Eqs. (5)–(7).

$$MAE_t = \frac{1}{n_t}\sum_{k=1}^{n_t}|P_k - P'_k| \times 100\,\% \tag{5}$$

$$MAPE_t = \frac{1}{n_t}\sum_{k=1}^{n_t}\frac{|P_k - P'_k|}{P} \times 100\,\% \tag{6}$$

$$RMSE_t = \sqrt{\frac{\sum_{k=0}^{n_t}(P_k - P'_k)^2}{n_t}} \tag{7}$$

Where P_k, P'_k are phenotypes of a single class measured manually and by smartphone-based heterogeneous binocular vision, respectively; k indicates which P and P' is; n_t is the quantity of t (t indicates a certain phenotype).

The *mAP* (mean average precision) was adopted to evaluate the detection performance of YOLOv5s model, which defined by Eqs. (8) and (9).

$$mAP = \frac{1}{c}\sum_{i=0}^{c} AP_i \tag{8}$$

$$AP_i = \int_0^1 P(R_i)\mathrm{d}R_i \tag{9}$$

Where c is the number of labeled classes, which is four in this study; AP_i is the average precision of a certain class, which is the area under *P-R* curve of the i^{th} class; P and R are precision and recall, which are defined in Eqs. (10) and (11), respectively. The *TP*, *TN*, *FP* and *FN* are indicated four types of detected datasets: true positive, true negative, false positive and false negative according to integration of ground truth and predicted class, respectively.

$$P = \frac{TP}{TP + FP} \tag{10}$$

$$R = \frac{TP}{TP + FN} \tag{11}$$

3. Results and discussions

3.1. Training assessment and performance of the model

As mentioned, a four-class training dataset of apple tree heterogeneous binocular images was applied to train YOLOv5s model under PyTorch framework, whose loss curve and *mAP* curve are displayed in Fig. 11. The model was convergent around the epoch of 350 according to the *mAP* curve of it, which basically fits the need of phenotyping of apple tree. The *mAP* reached 83.86 % after 400 Epoch of training, where AP_i for each class of the four-class dataset (fruit, graft, trunk, whole tree) under PyTorch are 95.31 %, 79.32 %, 84.29 % and 76.53 %, respectively.

After training of the YOLOv5s model, the model was converted to format of NCNN framework to fit the deployment in Android smartphone. Besides, the model is also simplified by onnx-simplifier and onnx-optimizer (two modules from ONNX). The final YOLOv5s NCNN model performed a similar performance with YOLOv5s PyTorch model, which achieved *mAP* of 84.62 %. The *P-R* curves of each class for the YOLOv5s NCNN model were demonstrated in Fig. 12. The detection speed of the YOLOv5s NCNN model is about 190 ms for each image captured by main camera or ultra-wide camera.

3.2. Evaluation of phenotyping

Phenotyping with smartphone-based heterogeneous binocular vision and YOLOv5s for the five phenotypes of apple tree reached an acceptable performance. The *MAE*, *MAPE* and *RMSE* of five phenotypes of apple tree are demonstrated in Table 3. It can be seen that phenotyping for five phenotypes reached *MAPE* ranging from 6.00 % to 13.73 %. All the evaluation criteria were calculated with the datasets in Sections 2.1 and 2.2.

The phenotyping deviation was primarily due to non-parallel alignment between CMOS flat and phenotype flat. As what demonstrated in Fig. 13, the non-parallel alignment between flats resulted in varying results depending on depth it obtained. More specifically, the smaller

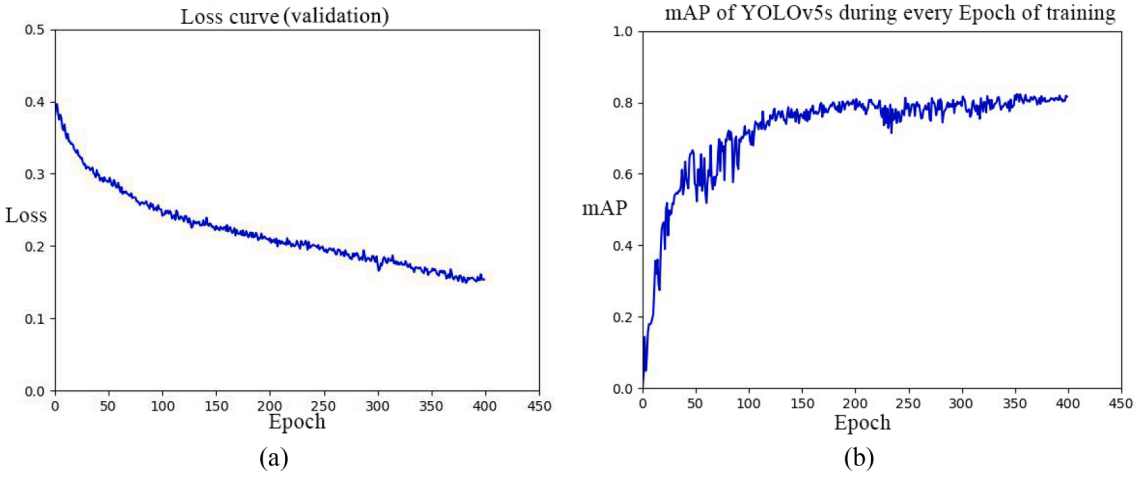

Fig. 11. Loss curve and *mAP* curve of YOLOv5s. (a) Loss curve of YOLOv5s. (b) *mAP* curve of YOLOv5s.

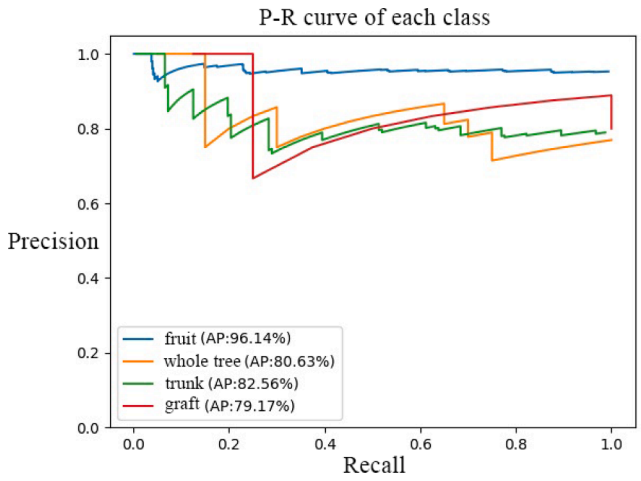

Fig. 12. *P-R* curves of each class under the YOLOv5s NCNN model.

Table 3
MAPE, *MAE* and *RMSE* of the five phenotypes.

Phenotypes	Trunk diameter	Tree height	Ground diameter	Fruit horizontal diameter	Fruit vertical diameter
MAPE (%)	11.20	6.00	11.45	12.96	13.73
MAE (mm)	6.78	184.15	15.50	8.80	8.50
RMSE (mm)	8.33	238.58	20.59	10.84	10.31

the depth of the phenotype, the smaller the measured phenotype. It is possible to minify the deviation by using posture of smartphone obtained by its gyroscope during image capture to fulfill a more accurate phenotyping in the future.

3.3. Results from other studies on phenotyping

Herein, the feasibility of applying the smartphone-based heterogeneous binocular vision to fulfill phenotyping of apple tree is proved. To the best of our knowledge, this is the first study uses the smartphone-based heterogeneous binocular vision to fulfill phenotyping. The results indicated the great potential of phenotyping with the smartphone-based heterogeneous binocular vision. Furthermore, our method has the lowest cost among the existing automated phenotyping researches. It even can be no cost for those who using smartphone with multi-camera.

Compared with the existing studies, our method achieved a similar accuracy for phenotyping of some phenotypes at an extremely much lower cost. A comparison with studies of the same phenotypes for different instruments were made, whose details are demonstrated in Table 4. For fruit size, Mirbod et al. (2020) applied measurement of apple fruit size and yield by stereo vision cameras, which obtained a system capable of capturing variability by 10.00 mm. Gené-Mola et al. (2021) used a Canon DSLR (Digital Single Lens Reflex Camera) with structure from motion and multi-view geometry to estimate apple fruit size by four different methods, which achieved *MAE*s of 3.70 mm to 15.00 mm for apples with different occluded situations. Compared with that, our method achieved *MAE* of 8.50 mm to 8.80 mm for apple fruit size. For tree height, Nasiri et al. (2021) applied UAV to determining tree height with *RMSE* of 325.22 mm, while our method reached a 238.58 mm *RMSE* of tree height measurement. For DBH and trunk diameter, Shao et al. (2022) estimated DBH with monocamera of smartphone and laser ranger mounted on a tripod, which obtained a *MAE* of 11.20 mm for DBH estimation. Compared with studies of Shao et al. (2022) and Sun et al. (2022), our method achieved *MAE* of 6.78 mm for trunk diameter. In the following research, a comparison between phenotyping by different instruments in the same grow stage of apple tree and experimental orchard will be conducted to evaluate the performance.

Even though the phenotyping using smartphone-based heterogeneous binocular vision method is workable without stereo calibration, smartphone-based heterogeneous binocular vision method with stereo calibration resulted to a better performance. This is mainly because deviations of the method are also affected by potential manufacturing errors in the multi-camera module, as illustrated in Fig. 14, whose red line and red rectangle indicate where the CMOS should be if it was correctly manufactured (Wu et al., 2010). The black solid lines and black rectangles of Fig. 14 represent the CMOSs of multi-camera module from a top and front view, respectively. The dotted lines denote the optical axes of multi-camera. O_L and O_R refer to the optical centers of left camera and right camera of heterogeneous binocular camera, respectively.

As what mentioned, the deviation caused by manufacturing error can be solved through stereo calibration. However, to maintain the user-friendly characteristic of the method, the commonest chessboard-based stereo calibration is not considered. Instead, a feature-based method would be a preferable way to fulfil stereo calibration and epipolar constraint of smartphone-based heterogeneous binocular vision method in the future.

4. Conclusions

In a word, the study developed one of the most low-cost automated phenotyping methods for individual apple tree, which only need a smartphone with multi-camera. The smartphone-based heterogeneous binocular vision method reached 11.20 %, 6.00 %, 11.45 %, 12.96 %, and 13.73 % *MAPE*s for trunk diameter, tree height, ground diameter, fruit horizontal diameter, and fruit vertical diameter, respectively. And *MAE*s of them reached 6.78 mm, 184.15 mm, 15.50 mm, 8.80 mm and

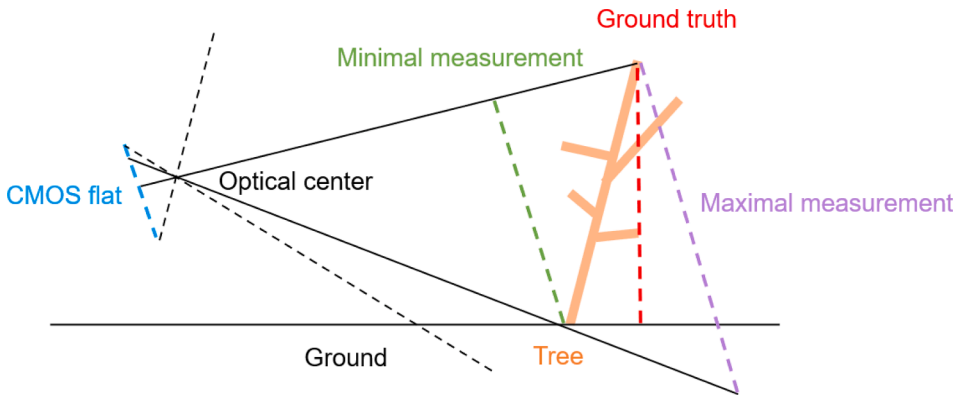

Fig. 13. Geometric representation of phenotyping deviation.

Table 4
Results from other studies on apple tree phenotyping.

Phenotype	Reference	Instruments	Comparison	
			Others	Ours
Fruit size	Mirbod et al. (2020)	Stereo camera	Capturing variability:10.00 mm	MAE: maximum of 8.80 mm
	Gené-Mola et al. (2021)	Canon EOS 60D DSLR camera	MAE: maximum of 15.00 mm for different occluded apples	
Tree height	Nasiri et al. (2021)	UAV	RMSE: 325.22 mm	RMSE: 238.58 mm
DBH (Trunk diameter)	Shao et al. (2022)	Smartphone, laser ranger	MAE: 11.20 mm RMSE: 15.50 mm	MAE: 6.78 mm RMSE: 8.33 mm
	Sun et al. (2022)	Kinect V2	MAE: 3.01 mm RMSE: 3.79 mm	

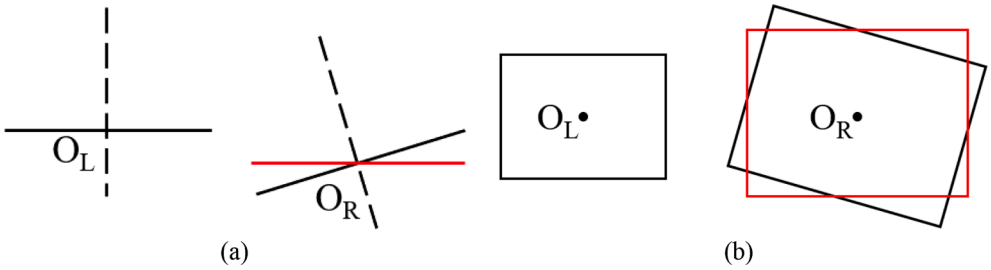

Fig. 14. Two cases of manufacturing error of the multi-camera module that might cause deviations of heterogeneous binocular method. (a) Optical axes unparallel caused by deflection of one of the cameras. (b) Rotational distortion caused by rotation of one of the cameras.

8.50 mm, respectively. The phenotyping of an individual apple tree can be fulfilled within 10 s (depends on the quantity of apples).

More specifically, the feasibility of phenotyping with only a smartphone instead of any instruments such as RGB-D camera or LiDAR is proved. Besides, the use of multi-camera of smartphone to obtain depth has been confirmed. Moreover, the phenotyping performance of smartphone is tested, which is similar to performance of existing researches. A future work will focus on the improvement of phenotyping accuracy by using gyroscope of smartphone and feature-based stereo calibration.

CRediT authorship contribution statement

Guanao Zhao: Data curation, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. **Ruizhe Yang:** Conceptualization, Investigation, Methodology, Writing – review & editing. **Xudong Jing:** Conceptualization, Investigation, Methodology, Writing – review & editing. **Haosen Zhang:** Conceptualization, Validation, Methodology. **Zhenchao Wu:** Methodology, Supervision, Writing – review & editing. **Xiaoming Sun:** Data curation, Methodology. **Hanhui Jiang:** Investigation, Writing – review & editing. **Rui Li:** Conceptualization, Data curation, Supervision. **Xiaofeng Wei:** Conceptualization, Supervision, Writing – review & editing. **Spyros Fountas:** Conceptualization, Supervision, Writing – review & editing. **Huijun Zhang:** Data curation, Writing – review & editing. **Longsheng Fu:** Conceptualization, Data curation, Methodology, Supervision, Writing – review & editing.

Declaration of Competing Interest

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

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (32171897); National Foreign Expert Project, Ministry of Science and Technology, China (DL2022172003L, QN2022172006L); Youth Science and Technology Nova Program in Shaanxi Province of China (2021KJXX-94); Science and Technology Promotion Program of Northwest A&F University (TGZX2021-29); Natural Science Foundation for Chongqing (No. cstc2020jscx-lygg0001).

References

Baek, S.H., Kim, M.H., 2016. Stereo fusion: combining refractive and binocular disparity. Comput. Vis. Image Underst. 146, 52–66. https://doi.org/10.1016/j.cviu.2016.02.006.

Bauer, A., Bostrom, A.G., Ball, J., Applegate, C., Cheng, T., Laycock, S., Rojas, S.M., Kirwan, J., Zhou, J., 2019. Combining computer vision and deep learning to enable ultra-scale aerial phenotyping and precision agriculture: a case study of lettuce production. Hortic. Res. 6, 70. https://doi.org/10.1038/s41438-019-0151-5.

Berk, P., Stajnko, D., Belsak, A., Hocevar, M., 2020. Digital evaluation of leaf area of an individual tree canopy in the apple orchard using the LIDAR measurement system. Comput. Electron. Agric. 169, 105158 https://doi.org/10.1016/j.compag.2019.105158.

Chang, A., Jung, J., Yeom, J., Maeda, M.M., Landivar, J.A., Enciso, J.M., Avila, C.A., Anciso, J.R., 2021. Unmanned aircraft system- (UAS-) based high-throughput phenotyping (HTP) for tomato yield estimation. J. Sensors 2021, 1–14. https://doi.org/10.1155/2021/8875606.

Chen, X., Zhang, W., Zheng, L., Gao, W., Wang, M., Affairs, R., Technology, B., 2020. PhenoStereo: a high-throughput stereo vision system for field- based plant phenotyping - with an application in sorghum stem diameter estimation Lirong. ASABE Meet. Present. 19, 2–12.

Corte, A.P.D., Rex, F.E., de Almeida, D.R.A., Sanquetta, C.R., Silva, C.A., Moura, M.M., Wilkinson, B., Zambrano, A.M.A., da Cunha Neto, E.M., Veras, H.F.P., de Moraes, A., Klauberg, C., Mohan, M., Cardil, A., Broadbent, E.N., 2020. Measuring individual tree diameter and height using gatoreye high-density UAV-lidar in an integrated crop-livestock-forest system. Remote Sens. 12 (5), 863. https://doi.org/10.3390/rs12050863.

Fan, Z., Sun, N., Qiu, Q., Li, T., Feng, Q., Zhao, C., 2022. In situ measuring stem diameters of maize crops with a high-throughput phenotyping robot. Remote Sens. 14 (4), 1030. https://doi.org/10.3390/rs14041030.

Gené-Mola, J., Sanz-Cortiella, R., Rosell-Polo, J.R., Escolà, A., Gregorio, E., 2021. In-field apple size estimation using photogrammetry-derived 3D point clouds: comparison of 4 different methods considering fruit occlusions. Comput. Electron. Agric. 188, 106343 https://doi.org/10.1016/j.compag.2021.106343.

Herzig, P., Borrmann, P., Knauer, U., Klück, H.C., Kilias, D., Seiffert, U., Pillen, K., Maurer, A., 2021. Evaluation of RGB and multispectral unmanned aerial vehicle (UAV) imagery for high-throughput phenotyping and yield prediction in barley breeding. Remote Sens. 13 (14), 2670. https://doi.org/10.3390/rs13142670.

Hobart, M., Pflanz, M., Weltzien, C., Schirrmann, M., 2020. Growth height determination of tree walls for precise monitoring in apple fruit production using UAV photogrammetry. Remote Sens. 12(10), 1656. doi: 10.3390/rs12101656.

Kim, C.G., 2013. A characterisitc analysis study of android based stereoscopic 3D technology. J. Satell. Inf. Commun. 8, 68–73.

Kim, S.K., 2021. Generation of stereo images from the heterogeneous cameras. Instrum. Mes. Metrol. 20, 73–78. https://doi.org/10.18280/i2m.200202.
Krisanski, S., Taskhiri, M.S., Turner, P., 2020. Enhancing methods for under-canopy unmanned aircraft system based photogrammetry in complex forests for tree diameter measurement. Remote Sens. 12 (10), 1652. https://doi.org/10.3390/rs12101652.
Li, T., Fang, W., Zhao, G., Gao, F., Wu, Z., Li, R., Fu, L., Dhupia, J., 2021. An improved binocular localization method for apple based on fruit detection using deep learning. Inf. Process. Agric. doi: 10.1016/j.inpa.2021.12.003.
Li, D., Shi, G., Kong, W., Wang, S., Chen, Y., 2020. A leaf segmentation and phenotypic feature extraction framework for multiview stereo plant point clouds. IEEE J. Sel. Top. Appl. EARTH Obs. Remote Sens. 13, 2321–2336.
Liao, L., Cao, L., Xie, Y., Luo, J., Wang, G., 2022. Phenotypic traits extraction and genetic characteristics assessment of eucalyptus trials based on UAV-Borne LiDAR and RGB images. Remote Sens. 14 (03), 0765. https://doi.org/10.3390/rs14030765.
Lin, D., Wang, Z., Shi, H., Chen, H., 2021. Modeling and analysis of pixel quantization error of binocular vision system with unequal focal length. J. Phys. Conf. Ser. 1738, 012033 https://doi.org/10.1088/1742-6596/1738/1/012033.
Liu, Y., Xing, M., Zhou, X., Song, Y., Wang, D., 2020. Tree height extraction in sparse scenes based on UAV. IGARSS 2020, 6499–6502.
Machimura, T., Fujimoto, A., Hayashi, K., Takagi, H., Sugita, S., 2021. A novel tree biomass estimation model applying the pipe model theory and adaptable to UAV-derived canopy height models. Forests 12 (2), 1–16. https://doi.org/10.3390/f12020258.
Maesano, M., Khoury, S., Nakhle, F., Firrincieli, A., Gay, A., Tauro, F., Harfouche, A., 2020. UAV-based LiDAR for high-throughput determination of plant height and above-ground biomass of the bioenergy grass arundo donax. Remote Sens. 12 (20), 3464. https://doi.org/10.3390/rs12203464.
Mirbod, O., Choi, D., Heinemann, P., Marini, R., 2020. Towards image-based measurement of accurate apple size and yield using stereo vision cameras. ASABE 2020 Annu. Int. Meet., pp. 1–6. doi: 10.13031/aim.202001115.
Nasiri, V., Darvishsefat, A.A., Arefi, H., Pierrot-Deseilligny, M., Namiranian, M., Le Bris, A., 2021. Unmanned aerial vehicles (Uav)-based canopy height modeling under leaf-on and leaf-off conditions for determining tree height and crown diameter (case study: Hyrcanian mixed forest). Can. J. For. Res. 51 (7), 962–971. https://doi.org/10.1139/cjfr-2020-0125.
Sengupta, S., 1997. Effects of unequal focal lengths in stereo imaging. Pattern Recognit. Lett. 18 (4), 395–400. https://doi.org/10.1016/S0167-8655(97)00024-X.
Shao, T., Qu, Y., Du, J., 2022. A low-cost integrated sensor for measuring tree diameter at breast height (DBH). Comput. Electron. Agric. 199, 107140 https://doi.org/10.1016/j.compag.2022.107140.
Solvin, T.M., Puliti, S., Steffenrem, A., 2020. Use of UAV photogrammetric data in forest genetic trials: measuring tree height, growth, and phenology in Norway spruce (*Picea abies* L. Karst.). Scand. J. For. Res. 35 (7), 322–333. https://doi.org/10.1080/02827581.2020.1806350.
Sun, X., Fang, W., Gao, C., Fu, L., Majeed, Y., Liu, X., Gao, F., Yang, R., Li, R., 2022. Remote estimation of grafted apple tree trunk diameter in modern orchard with RGB and point cloud based on SOLOv2. Comput. Electron. Agric. 199, 107209 https://doi.org/10.1016/j.compag.2022.107209.
ten Harkel, J., Bartholomeus, H., Kooistra, L., 2020. Biomass and crop height estimation of different crops using UAV-based LiDAR. Remote Sens. 12 (01), 0017. https://doi.org/10.3390/RS12010017.
Westling, F., Underwood, J., Bryson, M., 2021. Graph-based methods for analyzing orchard tree structure using noisy point cloud data. Comput. Electron. Agric. 187, 106270 https://doi.org/10.1016/j.compag.2021.106270.
Wu, J., Snášel, V., Abraham, A., 2010. A vision-based navigation system of mobile tracking robot. In: Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern., pp. 3053–3059. doi: 10.1109/ICSMC.2010.5642253.
Wu, Z., Li, G., Yang, R., Fu, L., Li, R., Wang, S., 2022. Coefficient of restitution of kiwifruit without external interference. J. Food Eng. 327, 111060 https://doi.org/10.1016/j.jfoodeng.2022.111060.
Zhao, J., Zhang, X., Yan, J., Qiu, X., Yao, X., Tian, Y., Zhu, Y., Cao, W., 2021. A wheat spike detection method in UAV images based on improved YOLOv5. Remote Sens. 13 (16), 3096. https://doi.org/10.3390/rs13163095.