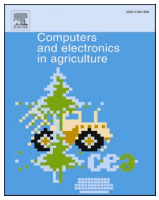
Computers and Electronics in Agriculture 185 (2021) 106123

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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



Original papers

DeepPhenology: Estimation of apple flower phenology distributions based on deep learning

Xu (Annie) Wang, Julie Tang, Mark Whitty *

School of Mechanical and Manufacturing Engineering, UNSW Sydney, Sydney, NSW 2052, Australia

ARTICLE INFO

Keywords:
Phenology
Deep learning
Apple
Maturity
Image processing

ABSTRACT

Estimation of phenology distribution in horticultural crops is very important as it governs the timing of chemical thinning in order to produce good quality fruit. This paper presents a novel phenology distribution estimation method named *DeepPhenology* for apple flowers based on CNNs using RGB images, which is able to efficiently map the flower distribution on an image-level, row-level, and block-level. The image classification model VGG-16 was directly trained with relative phenology distributions calculated from manual counts of flowers in the field and acquired imagery. The proposed method removes the need to label images, which overcomes difficulties in distinguishing overlapping flower clusters or identifying hidden flower clusters when using 2D imagery. *DeepPhenology* was tested on both daytime and night-time images captured using an RGB camera mounted on a ground vehicle in both Gala and Pink Lady varieties in an Australian orchard. An average Kullback-Leibler (KL) divergence value of 0.23 over all validation sets and an average KL value of 0.27 over all test sets was achieved. Further evaluation has been done by comparing the proposed model with YOLOv5 and shown to outperform this state-of-the-art object detection model for this task. By combining relative phenology distributions from a single image to a row-level or block-level distribution, we are able to give farmers a precise and high-level overview of block performance to form the basis for decisions on chemical thinning applications.

1. Introduction

Apple flower phenology covers a period from the emergence of the small green fruit buds to the stage when flower petals start falling. The flowers grow over a period of 2–3 weeks and within this period, farmers need to carefully inspect the phenology stages in the orchard to make thinning decisions. Excess numbers of flowers are common and mean some need to be removedto produce high-quality fruit. Flower thinning has the greatest potential to increase fruit size and can promote return bloom in the following season. The efficacy of blossom thinners is known to vary with the percentage of open blossoms and therefore the timing of application (Kon et al., 2018). Through improved timing of application, apple growers are able to increase the quality of production and be more competitive in the market (DeLong et al., 2018; Peck et al., 2016). In general, the percentage of open blossoms is visually estimated by the grower on a small subset of trees within an orchard block which may not be representative and is time-consuming. Recent developments in computer vision and deep learning have the potential to increase productivity by efficiently gathering detailed information to help farmers' decision-making and management practices. The information and management can be down to specific trees within the orchard which makes farm management more precise and efficient, enabling management of on-farm variability.

To first illustrate the inputs for the computer vision algorithms, images of apple flowers, as well as their corresponding phenology, are shown in Fig. 1. In Australia, apple tree blossom usually occurs between September and October. We have adopted a definition of eight stages during flower growth and the progression from the first to the final stage generally lasts 2–3 weeks. Based on the images in Fig. 1, the first stage of blossom is 'Green Tip' with fruit buds showing light green colour at the tip (Fig. 1a). Once the bud bursts, with little spiky leaves at around half-inch size, it enters the second stage called 'Half-inch Green' (Fig. 1b). Generally speaking, an apple bud will produce 5–6 flowers. The next stage is when the bud opens into a very tight and green cluster of flowers surrounded by early spur leaves, called 'Tight Clusters' (Fig. 1c). After these three green stages, the flowers will start to exhibit a pink colour, giving rise to the 'Pink' stage (Fig. 1d). When the pink flowers begin to separate and swell and look like small pink balloons, they enter the next

* Corresponding author.
E-mail addresses: xu.wang@unsw.edu.au (X.(A. Wang), julie.tang@unsw.edu.au (J. Tang), m.whitty@unsw.edu.au (M. Whitty).

https://doi.org/10.1016/j.compag.2021.106123
Received 17 November 2020; Received in revised form 18 March 2021; Accepted 21 March 2021
Available online 5 April 2021

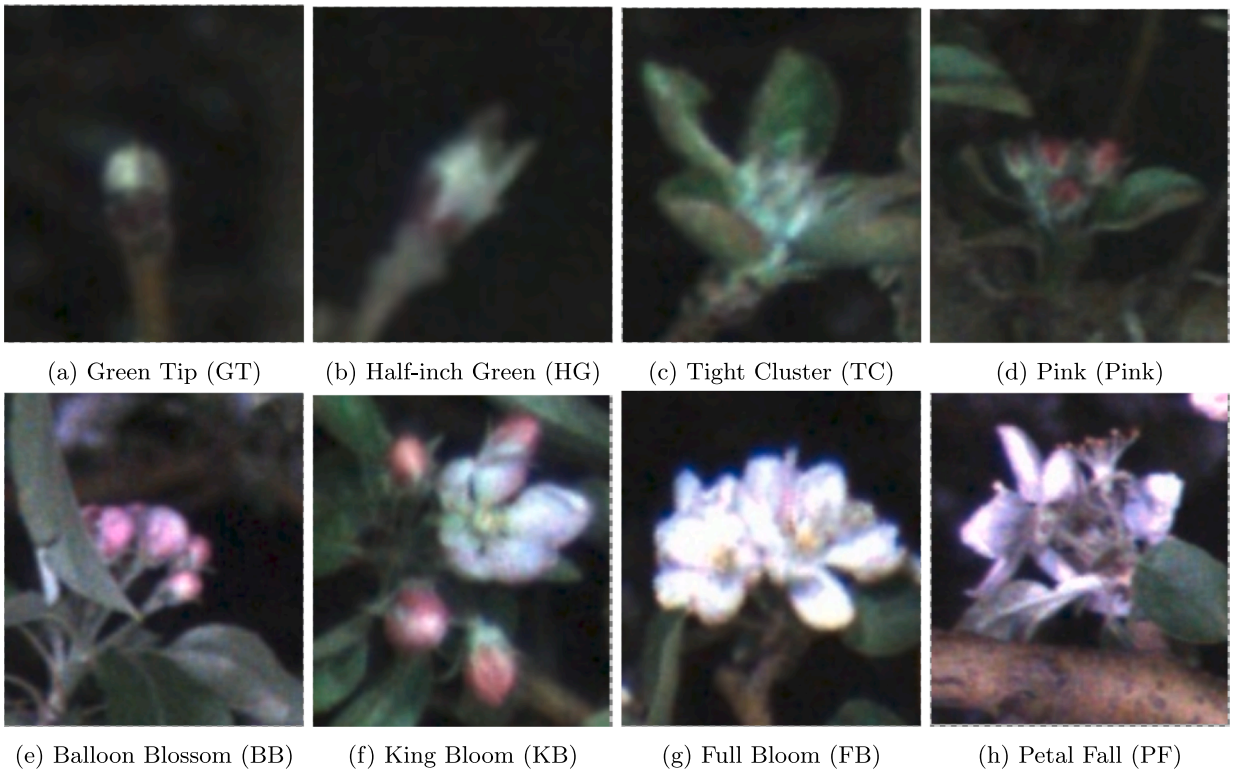


Fig. 1. Apple Flower Phenology Stages.

stage called 'Balloon Blossom' (Fig. 1e). By this stage, we will start to see the 'King Bloom' (the one in the center which appears most advanced, and has the highest fruiting potential) start to change colour and open, forming a white or light pink flower (Fig. 1f). The next stage is when the remaining pink balloon blossoms progressively open around the king bloom, denoted the 'Full Bloom' stage (Fig. 1g). When petals start falling, the final stage is called 'Petal Fall' (Fig. 1h). This paper aims to propose an automatic method of estimating the percentage of each phenological stage using images to get a row-level or block-level overview, which will help farmers decide the best timing of chemical thinning.

Although estimation of apple flower phenology distribution is very important, it is very challenging because of complex and overlapping phenological stages as introduced above. To the authors' knowledge, no published research to estimate on-tree flower phenology distribution comprising up to 8 stages using an automatic process is available at the time of writing. Some work exists for single-class on-tree flower or flower cluster detection either at the pixel-level (Wang et al., 2020; Dias et al., 2018; Dias et al., 2018) or object-level (Wu et al., 2020; Lim et al., 2020). However, single-class flower detection is not the focus of this research, so the literature discussed below is multi-class related as the emphasis is on estimating phenological stages. Most previous work has focused on fruit maturity and can be divided into two approaches: traditional handcrafted features and deep learning methods. Within traditional approaches, handcrafted features such as colour and texture are first extracted and then fed into a classifier such as Support Vector Machine (SVM), K-Nearest Neighbour (KNN), or Random Forest (RF). Indriani et al. (2017) proposed a traditional algorithm to determine the maturity stage of tomatoes in 2017. HSV colour features, as well as texture features based on Gray Level Co-occurrence Matrix (GLCM), were first analysed and then classified by KNN into 5 classes from raw, quite mature, mature, very ripe to rotten. Classification accuracy of up to 92% was achieved in this research. Similarly, in 2018, Pereira et al. (2018) used RF based on 21 hand-crafted colour features to predict the ripening of papaya fruit with a classification accuracy of 94%, while Kipli et al. (2018) used decision trees (DT) based on L*a*b colour features to classify the ripeness of the banana into three different maturity stages with an accuracy of 96%. Instead of extracting one colour feature value from the whole region of the object, Wan et al. (2018) then proposed a new method which extracts 5 colour features from 5 sub-regions divided from the whole fruit region in the image. These features were fed into a backpropagation neural network to do maturity classification of tomatoes, which achieved 99.31% accuracy. Other than colour features, Harel et al. (2020) also utilised morphological features such as eccentricity, extent, and solidity to classify maturity stages for sweet peppers by using RF, which achieved accuracy of up to 96%. Similar work for palm oil fruit and coconut fruit can be found in Septiarini et al. (2019), Hendrawan et al. (2019). Based on the above work, it is worth noting that traditional approaches rely heavily on reasonably good hand-crafted features. However, hand-crafted feature extraction is fruit-dependent, labour intensive, and also becomes challenging when it comes to objects with more complex colour and shape profiles like the phenology stages presented here in Fig. 1. The trend shifted towards using deep learning models that can extract and learn the features themselves. Nasiri et al. (2019) used VGG-16 to predict the ripening stage of healthy dates at the image-level, which performs feature extraction and selection automatically. This method achieved a classification accuracy of 96%. Yu et al. (2019) then used Mask-RCNN to do strawberry detection as well as maturity estimation. The model was trained on 1900 images and the fruit detection results of 100 test images showed that the average detection precision, recall, and mIoU rates were 95.78%, 95.41%, and 89.85%, respectively. As a result, deep learning has a high potential to achieve both high accuracy and robustness in fruit detection and maturity classification because feature extraction can be done automatically in the network. Based on the discussion above, the two types of deep learning methods that have been used to estimate fruit phenology are multi-class image classification and multi-class object detection. Multi-class image classification can only identify a single flower stage at an individual image-level, which means an individual image should contain at most a single flower cluster. Undesirable pre-

(a) Block 2 with row numbers (b) 3D central leader tree (c) Imaging system

Fig. 2. Block information and the imaging system.

processing steps are then required to generate sub-images of single flower cluster size. Multi-class object detection with 8 stages will require a correspondingly large number of image annotations. To address the limitations above, we have adjusted an image classification model (VGG-16) to directly estimate the flower phenology distribution without the need of generating single flower cluster images. Secondly, we have replaced the image labelling used in the object detection model with direct ground truth measurements obtained in the field.

Another aspect to note in computer vision in agriculture is the environment. Whether the evaluation of the work has been done in an indoor or outdoor environment is important. Most traditional methods require image pre-processing to do fruit segmentation, such as that seen in Wan et al. (2018) and Harel et al. (2020). In indoor environments, lighting conditions are easily controlled and segmentation is achievable using simple image processing such as colour thresholding. The other advantage of indoor environments is that all images can be captured without much variation so the trained model works very well with relatively high accuracy up to 96% (Nasiri et al., 2019). In outdoor or unstructured environments, many variables become uncertain. For example, the sun angle can result in severe overexposure in images which will affect the colour of the fruit. Complex backgrounds or fruit in the background row can significantly confuse the model, leading to lower accuracy. For example, Tu et al. (2018) used natural outdoor RGB-D images to identify the maturity stage of passionfruit. All passionfruit regions were first detected by Faster R-CNN and then classified into five maturity stages by using the Dense Scale Invariant Features Transform (DSIFT) algorithm combined with Locality-constrained Linear Coding (LLC). The overall classification accuracy was 91.52% but the classification accuracy at Near-Mature(NM) stage was only 58%. One main reason could be that fruit at the Near-Mature stage look similar to fruit at the Mature stage, especially under uncertain lighting conditions. Tan et al. (2018) also used images taken in outdoor scenes to recognise different maturity stages of blueberry fruit. Images were first cropped into patches, which were used to calculate Histogram of Oriented Gradients (HOG) feature vectors and a* & b* features. These features from fruit-like patches were then classified into three maturity stages using a KNN classifier. The classification accuracy for mature and intermediate fruit was up to 96%, but the accuracy for young fruit was only about 86%. Based on example images in the paper, it was found that leaf-like patches were easily misidentified as young fruit due to similarity in colour. Besides that, this method utilised a sliding window technique to generate hundreds of potential positions for blueberry in order to get a single blueberry in each patch, which was not considered efficient as a commercial solution. These two works still fall under the category of the traditional methods mentioned before, with the only difference in the method of segmenting single fruit image patches. Within deep learning methods applied in outdoor scenes, Chen et al. (2019) *et al.* used Faster-RCNN on aerial images to detect strawberries at different maturity stages. It is worthwhile to mention that this is the only work that evaluated detection results from images with the manual counts collected in the field. The effect of occlusion is unavoidable when counting fruits and flowers from imagery so it needs to be considered. In this paper, the average deep learning counting accuracy was 84.1% with average occlusion of 13.5%. Based on the above, we found the accuracy from images taken outdoors was lower than for images taken indoors. However, on-branch phenology estimation in uncontrolled environments is more practical for apple flowers.

The overall objective of this paper is to develop an automatic method that is able to directly estimate the phenology distribution for apple flowers from images without using normal bounding box annotation. The estimated phenology distribution can be also combined to get a row-level or block-level phenology distribution. Firstly, a novel method named DeepPhenology was proposed to directly estimate the phenology distribution for apple flowers from images based on a deep learning classification model by using a new type of ground truth collected in the field. Testing was then thoroughly conducted on both daytime and night-time images collected from a commercial apple orchard including both Gala and Pink Lady varieties. The evaluation and comparison were also conducted in comparison to the results calculated from YOLOv5. Finally, a simple way of combining relative distribution results from images to row-level or block-level mapping over the orchards was proposed and evaluated. Our main contributions are listed in the following:

- A novel method to estimate apple flower phenology distribution from images by building direct relationships with real counts in the field.
- Validation of the algorithm on both night-time and daytime datasets including two varieties captured on a moving vehicle in an uncontrolled environment.
- Validation of combining relative distribution results from the proposed model to row-level phenology distribution.

Table 1
Description of the imaging system.

Data Collection System	
Timing	Day and Night
Illumination	4 × 40 Watt LED lights
Type	RGB
Lens	8 mm or 12 mm [a]
Resolution	4196 × 2160
Vehicle Speed	5 km/h

[a] Images collected between 2019-09-20 and 2019-10-16 were using a 12 mm lens and others were using an 8 mm lens.

Table 2
Summary of datasets in this study that includes the date of capture, the mode of the phenology in the orchard when images were collected, daytime or night-time images, rows where images have been taken and number of images collected on that date. (See Fig. 1 for an explanation of phenology abbreviations.)

Date (Y-M-D)	Mode Phenology	Timing	Rows	No. of Images
2019–09-20	GT/HG/TC	Day & Night	14, 15, 16, 19	80 & 78
2019–09-23	HG/TC/Pink	Day	14, 15, 16, 19	80
2019–09-25	HG/TC/Pink	Day	14, 15, 16, 19	90
2019–09-26	TC/Pink	Day	14, 15, 16, 19	80
2019–09-30	P/BB/KB	Day & Night	14, 15, 16, 19	80 & 99
2019–10-04	KB/FB	Night	14, 15, 16, 19	79
2019–10-05	KB/FB	Day & Night	14, 15, 16, 19	80 & 80
2019–10-06	FB/PF	Night	14, 15, 16, 19	70
2019–10-07	FB/PF	Night	14, 15, 16, 19	80
2019–10-09	FB/PF	Day & Night	14, 15, 16, 19	80 & 72

The remainder of the paper is organized as follows. Section 2 describes our imaging system and data collection as well as ground truth data collection. Section 3 describes our novel algorithm DeepPhenology and experimental setup. Results from the proposed method are shown in Section 4. Discussion and some potential avenues for future work are shown in Section 5. Our concluding remarks are presented in Section 6.

2. Materials and methods

2.1. Data collection

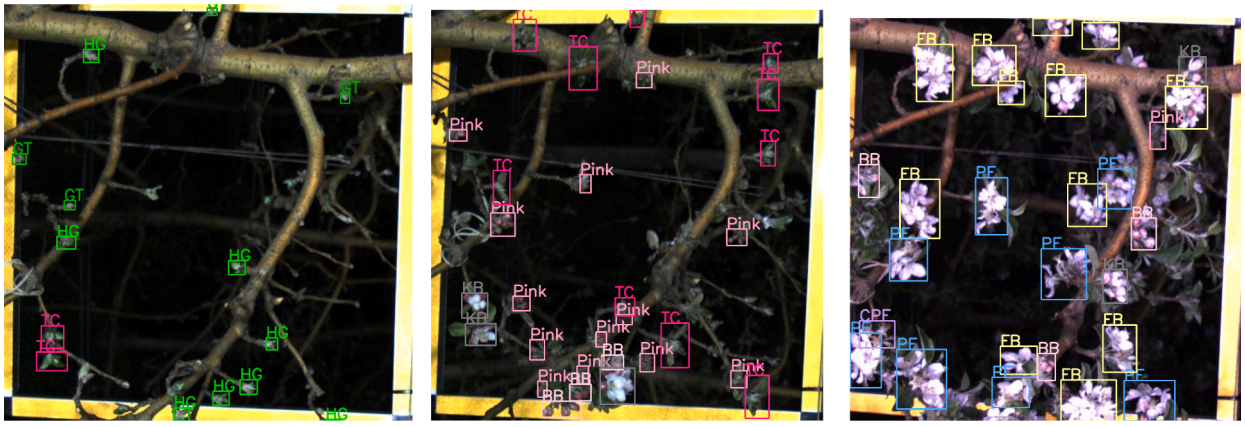
All data were acquired from a commercial apple orchard block located in Shepparton, Victoria, Australia, shown in Fig. 2a. The trees in Block 2 were planted in 2012 and are all 3D central leader trained as shown in Fig. 2b, with tree spacing of approximately 1.5 m and row spacing of approximately 4 m. The whole block contains 31 rows with approximately 2800 trees. For cross-pollination, every fourth pink lady row is followed by a gala row. In order to evaluate our proposed algorithm thoroughly, our four study rows cover both varieties. Rows 14 and 19 are of the Gala variety whereas rows 15 and 16 are of the Pink Lady variety. It is also noted that Gala usually progresses faster than Pink Lady in this region.

For this application, the mobile platform has been improved based on a subset of sensors mounted on a mobile platform built by Swarm-Farm Robotics in 2018, shown in Fig. 2c. Instead of using two separate cameras for daytime and night-time as in our previous research (Wang et al., 2020), the image acquisition system consists of only one RGB camera and four 40-Watt LED lights as shown in Table 1. The camera is mounted approximately perpendicular to the row at 2.2 metres from the ground in order to capture the full canopy. The vehicle travelled forward at approximately 5 km/h with the camera facing to the left to capture images of apple trees. RTK-GPS is used for positioning and the system includes a computer for data logging which was able to capture images every 0.5 m, approximately 1–2 images per second at the above speed. All datasets are summarized in Table 2. Datasets were collected from the 'Green Tip' stage to 'Petal Fall' stage which lasted about two weeks. All night and day images were taken for both sides of the row on the same day to ensure they were consistent with manual counts.

2.2. Ground truth

Doing object detection for apple flower clusters for individual phenology classes and then calculating the distribution based on detection results is one of the most common methods based on literature in Section 1. In this case, each visible flower cluster needs to be selected using a rectangular bounding box and assigned a phenology class as shown in Fig. 3. However, apple flowers are usually surrounded by messy leaves and occluded by other flower clusters of complex shapes as shown in Fig. 4. It is especially difficult to identify the correct class in the case that only one white flower is visible and is not surrounded by anything in the image because it could be the white flower from King Bloom, Full Bloom, or Petal Fall. This makes labelling very difficult when using 2D images.

Based on the occlusion issues mentioned above, a new concept of generating ground-truth is proposed in this paper. Instead of counting phenology class instances from images, manual counting is conducted in the field at the same time as images are collected. It is unrealistic to do manual counting over the whole tree so coloured square section markers (herein termed quads) were placed to limit the counting area as shown by Hu and Whitty (2019). In this paper, 40 orange quads with unique QR



(a) Example 1 at mostly green stages (b) Example 2 at pink stages (c) Example 3 at bloom stages

Fig. 3. Examples of bounding box labelling for a single tree section at various times in the flowering period.

(a) Example 1 (b) Example 2

Fig. 4. Difficult examples observed while labelling flower clusters from 2D images.

(a) Side A (b) Side B

Fig. 5. Images of Quad 27, pictured from both sides of the row.

Table 3
Manual counts for Quad 27 in Fig. 5.

Side	Date	Row	GT	HG	TC	Pink	BB	KB	FB	PF
A	5 October 2019	16	0	0	3	6	1	1	8	0
B	5 October 2019	16	0	0	1	2	1	1	2	0
	Total		0	0	4	8	2	2	10	0
	Distribution		0%	0%	15.4%	30.8%	7.7%	7.7%	38.4%	0%

codes have been placed in our four study rows with 10 quads evenly spaced in each row. Manual counting of flower clusters inside the quad was conducted during the day in which images were collected, which was much easier and more accurate than counting from images. The manual counting was divided into two sides (A and B) based on the wire trellis (See Fig. 5). The dates on which manual counts were undertaken are listed in Table 2 and they covered 8 phenological stages from 'Green Tip' to 'Petal Fall'. Fig. 5 shows images of Quad 27 with side A and side B respectively and Table 3 shows manual counts of each side on 2019-10-05. Based on counts of each phenological stage from both sides, we calculate the relative distribution over the phenological stages for the quad. An example for a quad is shown in Table 3, which shows FB as the majority class with Pink as a close second. Note that the counts were done during the daytime so flowers in night-time images may have progressed slightly.

3. Apple flower phenology estimation

3.1. DeepPhenology

Estimating the phenology distribution of apple flowers is formulated as a classification task with 8 classes. In normal image classification, the

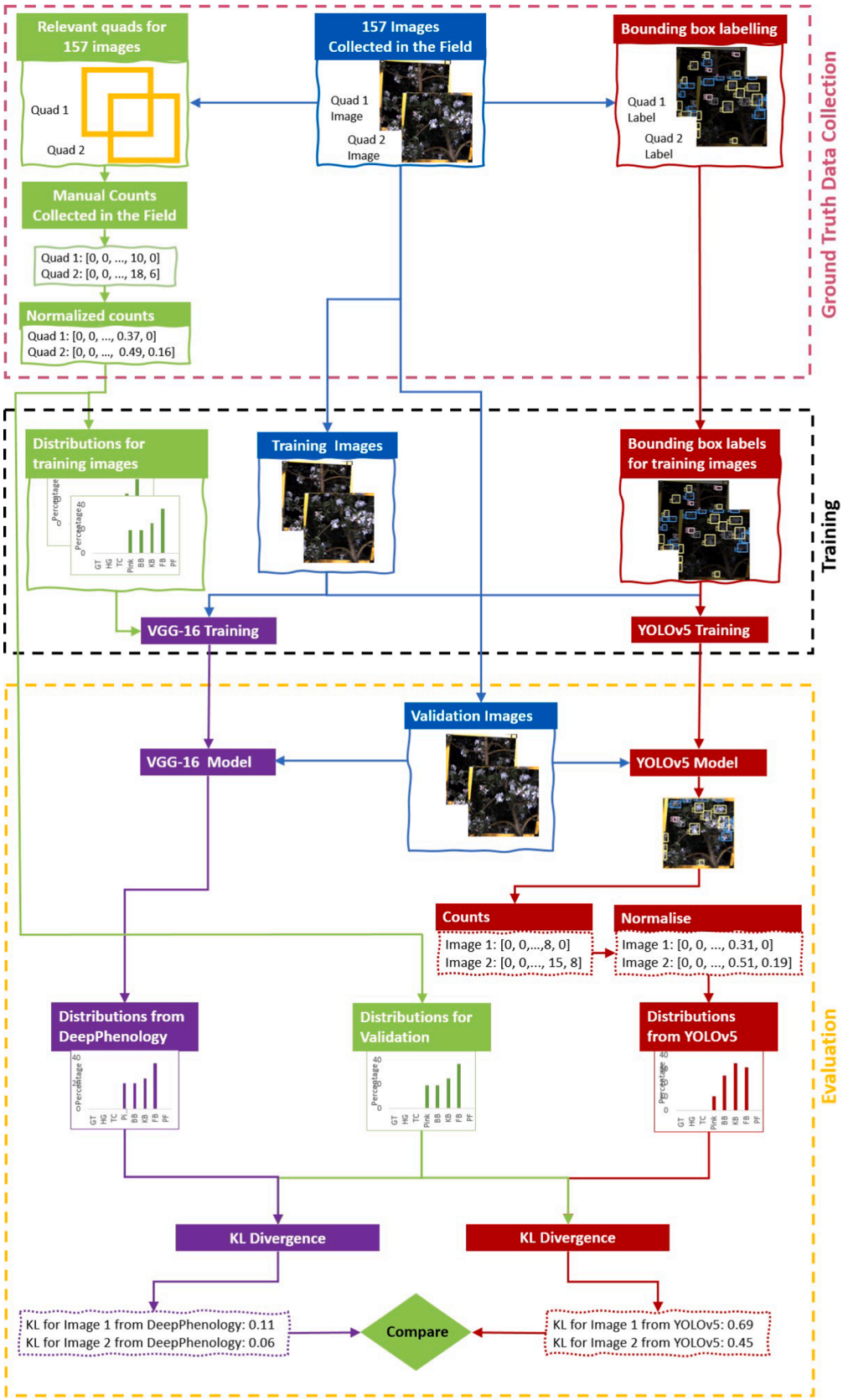


Fig. 6. The pipeline of DeepPhenology and YOLOv5 for estimating apple flower phenology distribution. The green boxes show how the new type of ground truth is made and used by the VGG-16 model, which the forms the basis of our DeepPhenology method. The red colour boxes show how YOLOv5 is trained and the process of generating phenology distributions based on detection results from YOLOv5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

CNN model accepts the given input images and produces an output probability over all classes. The class with the highest probability is then considered as the one that the image belongs to. The ground-truth for multi-class image classification is a one-hot type which means images can only be one class as discussed in Section 1, which is not the aim of this paper. Inspired by soft labelling (Müller et al., 2019), the ground-truth that we feed into the CNN model is a normalized 8-class distribution based on manual counts collected in Section 2.2, instead of a one-hot vector. The raw output from the CNN model followed by Softmax calculation then becomes a discrete normalized probability distribution. Given the good accuracy that we previously achieved using VGG16 for apple flower density estimation (Wang et al., 2020), VGG16 was also used in this paper.

3.2. Loss function

Because the output and the target are both distributions, the loss function of the deep learning model needs to be able to measure the difference between these two distributions. The Kullback-Leibler(KL) Divergence value (Kullback and Leibler, 1951) gives a measure of this and was chosen as it is differentiable and thus appropriate for a gradient based CNN. As a result, the KL value was utilised as the loss function as well as the evaluation index; with a smaller KL value indicating a distribution closer to the reference. The KL Divergence between two probability distributions P and Q is calculated as follows:

$$D_{KL}(P\|Q) = \sum_{x\in\chi} \log\left(\frac{P(x)}{Q(x)}\right). \quad (1)$$

3.3. Training

To better investigate the robustness, all daytime and night-time images from one specific quad that we randomly chose (e.g. Quad 27) were first left out to form the test set as this will enable us to test the performance across all phenological stages. The rest of the daytime and night-time images were randomly split into training and validation sets in a 70% and 30% split respectively. To avoid disproportionate splitting between daytime and night-time, the splitting was done on daytime and night-time images separately and those for the respective training set or validation set combined together. During the experiment, we chose 12 out of 40 quads covering both Pink Lady and Gala to do cross validation with the proposed model. For each cross validation, we have a training set, a validation set and a test set. The deep learning framework used for the implementation of VGG-16 with Batch Normalisation was Pytorch 1.1.0. The network was initialized using the weights previously computed for ImageNet. No layer was frozen during training, so all weights could be updated by the training on the apple flower dataset. The training for the network was run on a local machine with a single GeForce RTX 2080 Ti (12 GB), an Intel(R) Core(TM) i9-9900KF CPU @ 3.60 GHz with Ubuntu 18.04 LTS. The overall training process took approximately 2 h for 200 epochs using Stochastic Gradient Descent (SGD) with Nesterov momentum at 0.9, a batch size of 8, and a learning rate of 0.00005. Random horizontal or vertical flip and HSV augmentation was also used to improve robustness during training.

3.4. Evaluation

To better investigate our proposed method, three evaluations have been conducted in this study.Two evaluations were performed in terms of image-level phenology distribution while the third evaluation was performed in terms of row-level phenology distribution.

Table 4
KL value of validation and test set.

No.	No. of Training/ Validation Images	KL validation	Quad ID	Variety	No. of Quad Images	KL test
1	769/330	0.222	25	Pink Lady	29	0.100
2	770/330	0.225	26	Pink Lady	28	0.172
3	770/330	0.231	27	Pink Lady	28	0.182
4	771/331	0.225	28	Pink Lady	26	0.186
5	770/330	0.231	29	Pink Lady	28	0.174
6	770/331	0.244	30	Pink Lady	27	0.103
7	771/331	0.231	31	Gala	26	0.626
8	771/331	0.240	32	Gala	26	0.194
9	771/331	0.240	33	Gala	26	0.199
10	771/331	0.243	34	Gala	26	0.142
11	771/331	0.239	35	Gala	26	0.155
12	771/331	0.235	36	Gala	26	0.460
Avg.	–	0.236	–	–	–	0.278

The first evaluation of single image phenology distribution was to calculate the average KL values mentioned in Section 3.2 for images from each validation set as well as each test set among 12 cross validations. This aims to prove that DeepPhenology provides relatively good accuracy and is also robust enough to different datasets. The second evaluation was to compare the performance from the proposed model with the object detection method which is a highly possible way of generating a phenology distribution based on previous relevant literature in Section 1. To explain it further, the object detection model was first trained using the same input images with bounding box annotations for 8 phenological stages defined in Fig. 1. The detection results for the image were then accumulated to obtain counts for each phenological stage. The same format phenology distribution as DeepPhenology was built by normalizing these counts, from which a KL value could be computed relative to the manual ground truth distribution used in DeepPhenology. For a fair comparison, the object detection model selected needs to be a state-of-the art model in terms of both accuracy and speed. Since YOLO models have shown high potential in speed while achieving high accuracy in the literature (Santos et al., 2020; Koirala et al., 2019), YOLOv5 (Jocher et al., 2020), the latest version of the YOLO family, has been chosen as the object detection model. Among the YOLOv5 models, YOLOv5l was selected for implementation as this model has the highest potential of achieving high detection accuracy. Another thing to note is that this evaluation with YOLOv5l was only done against night-time imagery. To train the object detection model, bounding box labelling was done for 157 night-time quad images. Based on our experience, the labelling time was roughly 8–10 images per hour, depending on how many objects were in each image. These images were assigned randomly into Training-YOLO and Testing-YOLO datasets in a 70% and 30% split respectively. No layer was frozen during training and the training for the network was run on the same local machine mentioned in 3.3. The overall training process took approximately 40 min with a batch size of 16 for 300 epochs, and a learning rate of 0.01. To be a fair comparison, DeepPhenology was also retrained using the same Training-YOLO and Testing-YOLO sets. Note that the input size for YOLOv5l is 640*640 whereas the input size of our model is still 448*448. However, better accuracy is generally expected with higher input resolutions. It is noted that the ground truth (manual counts

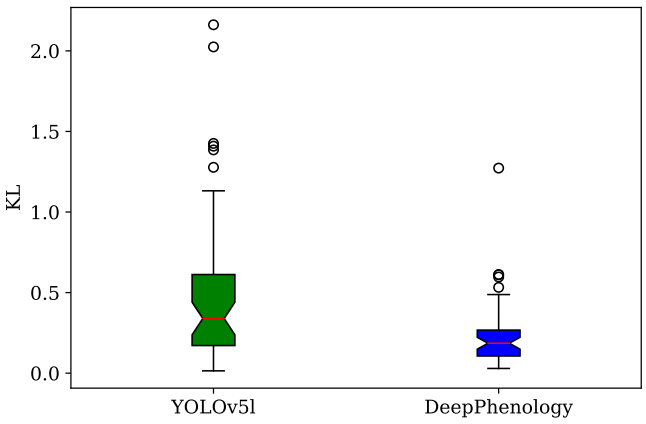


Fig. 7. Boxplot of KL values of the validation set from YOLOv5l and our proposed method.

collected in the field) was the same for YOLOv5l and DeepPhenology but YOLOv5l has additional bounding-box labelling for training which our DeepPhenology does not require. The whole pipeline of estimating apple flower phenology distribution based on DeepPhenology and YOLOv5 has been summarized in Fig. 6.

After evaluation on a single image patch, it is important to evaluate the performance of combining different image patches in order to get a row-level or block-level phenology distribution. Here we propose a simple method of evaluating the combination performance. All the validation images were first separated based on date and variety and this left 10 dates for two varieties across the whole validation set; 20 small sets in total. For each small set, phenology estimation for each stage was simply averaged across all images so a combined phenology distribution was achieved. Equivalently, the ground-truth distribution for each small set was also simply averaged over all samples.

4. Results

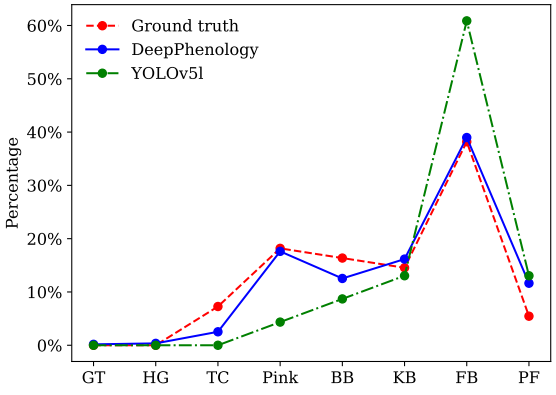
4.1. Single quad image evaluation

To verify our method on single quad images, the KL value was calculated for each image from the validation/test set and all KL values were averaged to get the final KL value for the validation/test set. As shown in Table 4, there are 12 cross validation rounds in total. Quads 25 to 36 were selected as our quad test sets and the rest of the images were split as mentioned in 3.3. Among all quad test sets, Quads 25 to 30 are of the Pink Lady variety whereas Quads 31 to 36 are of the Gala variety. We can see that the validation set contains about 330 images and the quad test set contains about 27 images. Across all 12 validation rounds, the average KL values of each validation set were consistently around 0.23. The KL values from the quad test set had a slightly slower median, but two outliers (0.46 and 0.62 for Quads 31 and 36 respectively) out of 12 values skewed the mean to be slightly higher than the validation set. See Section 5 for a discussion of these results.

4.2. Comparison with YOLOv5

The comparison between our proposed model DeepPhenology and YOLOv5l has been conducted as mentioned in Section 3.4. The proposed model and YOLOv5l were trained on the same training set and tested on the same test set. The results show our proposed algorithm achieved much better accuracy with a KL value of 0.235 whereas the KL value for YOLOv5l was almost double this, at 0.511. Fig. 7 shows boxplots of KL values for each test image from both YOLOv5l (left) and DeepPhenology (right). It is obvious that KL values from YOLOv5l have some unusual numbers up to 2.4, whereas most KL values from our proposed algorithm are less than 0.5 which indicates our proposed model gave more stable results. Fig. 8a shows detection overlay results from YOLOv5l and Fig. 8b shows a comparison between phenology distributions from YOLOv5l and the proposed algorithm as well as ground truth. The image

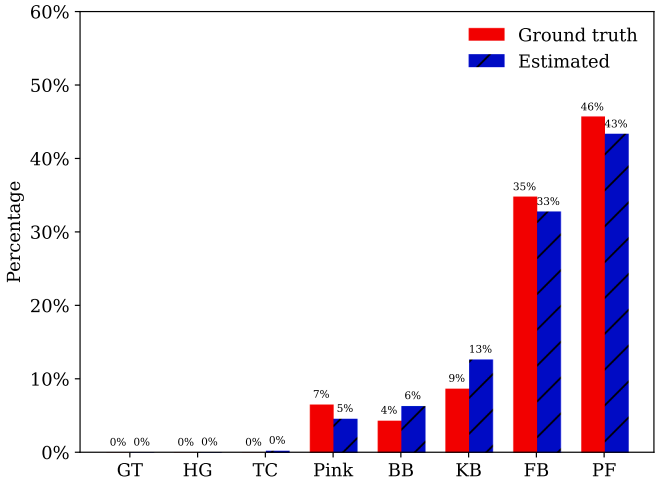
(a) The YOLOv5l bounding box detection results



(b) The estimated phenology distribution between DeepPhenology and YOLOv5l in comparison to the ground truth

Fig. 8. Comparison between YOLOv5l and DeepPhenology.

(a) Quad 35 image on 2019-10-07



(b) Ground truth vs Estimate from DeepPhenology

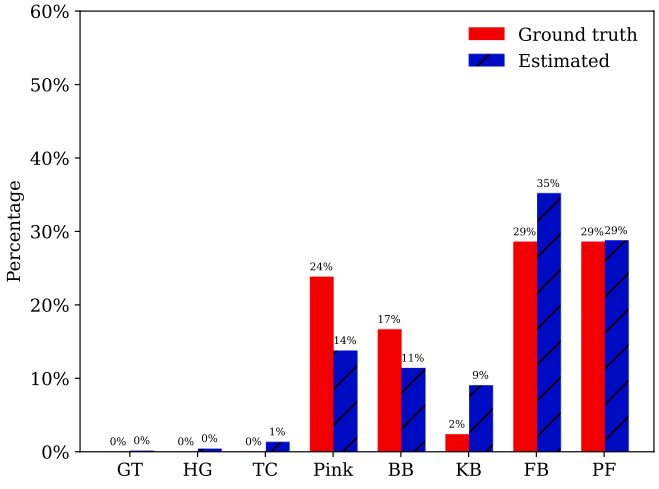
Fig. 9. One example result with a KL value of 0.01 from DeepPhenology.

in Fig. 8a is from Quad 30 Side B on 2019-10-06 and the KL value from YOLOv5l is 0.3 compared with a much smaller value of 0.06 from DeepPhenology. There is a minor difference in the inference time between YOLOv5l (0.016 s) and our proposed algorithm (0.012 s).

4.3. Combination evaluation

The first validation set consisting of 330 images in Table 4 was selected in this section as the basis for evaluating the combination of images, simulating the effect of imaging at a row or block level. Fig. 13 shows the combined phenology distribution for all small sets from 2019-09-20 to 2019-10-09 for both varieties. Blue bars in images mean the estimated combination distribution from DeepPhenology whereas red bars mean the ground-truth. The title of each sub-figure indicates the date and the variety as well as timing of images in parentheses, and the left top corner of each sub-figure also indicates its related KL value. Most KL values were very low, less than 0.08, which is much smaller than the average KL values on validation/test set calculated in Section 4.1. There are only two dates with relatively large KL values which are 2019-09-25 Pink Lady (Day) with a KL value of 0.2 and 2019-09-26 Gala (Day) with a KL value of 0.11. For 2019-09-25 Pink Lady, the proposed model indicates different distribution for GT and TC that around 20% of TC were expected as GT. On 2019-09-26 Gala, the proposed model expected 12% more Pink and less TC as well as BB. However, the KL values of the combined distributions for these two dates (2019-09-25 Pink Lady (Day) and 2019-09-26 Gala (Day)) are still smaller than the average KL values on validation/test set calculated per quad in Section 4.1, which proves the combined row-level phenology distribution has relatively better performance. It is also noted that both large KL values are from daytime-only images. The small KL values on all dates for two varieties also show the proposed method is quite robust to different varieties. As shown in all figures, Gala was often more progressed than Pink Lady, which

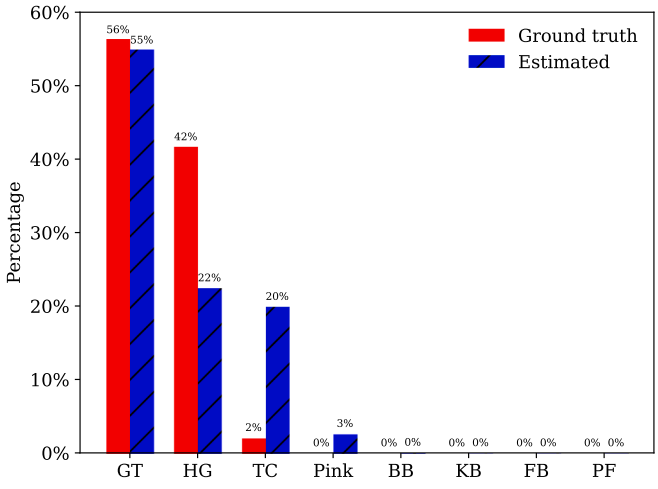
(a) Quad 29 image on 2019-10-09



(b) Ground truth vs Estimate from DeepPhenology

Fig. 10. One example result with a KL value of 0.10 from DeepPhenology.

(a) Quad 27 image on 2019-10-06



(b) Ground truth vs Estimate from DeepPhenology

Fig. 11. One example result with a KL value of 0.26 from DeepPhenology.

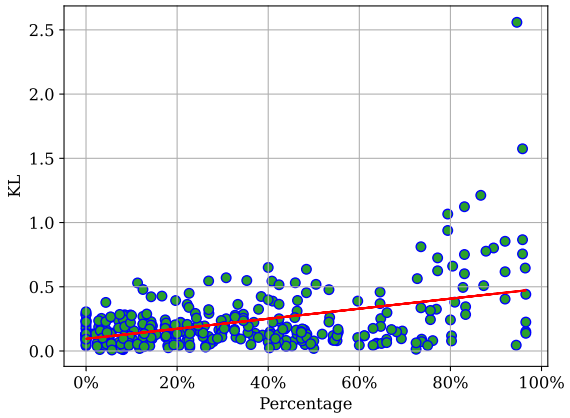


Fig. 12. KL values vs the proportion of leaf clusters to total clusters.

matched our ground-truth. For example, Gala on 2019-09-20 already contained around 40% TC (Fig. 13a) whereas Pink Lady on 2019-09-20 was still at GT stage (Fig. 13b).

5. Discussion

Fruit tree flowering phenology is important to decide the timing of the thinning application. Currently, the common way of estimating the flower phenological development is by human inspection, which is unreliable and time-consuming. Due to the success of computer vision and deep learning, researchers start utilising these techniques to increase productivity and accuracy in phenology estimation. Yu et al. (2019) has used Mask-RCNN to do strawberry fruit detection at different maturity stages, which could then be used to calculate maturity distribution in images. However, normal bounding box labelling for complicated objects such as apple flower clusters is considered much more difficult than fruit because flower clusters have more complex shapes than simple round shape fruits and there is less colour change between certain stages, unlike fruits during progression as shown in Fig. 4. Similarly, Chen et al. (2019) *et al.* used Faster-RCNN on aerial images to detect strawberries at different maturity stages as well as flowers. Besides detection, Chen has also evaluated the detection results from images with the real count in the field. Because of unavoidable occlusion, this is considered an important evaluation but easily ignored in other previous research. This evaluation can prove whether the counting from images is meaningful as an input for the thinning application. Anecdotally, the farmer found a phenological distribution at different stages in the block is a useful input when making decisions on when and where to spray. The effective use of this data in a decision support tool is the subject of an ongoing project.

Considering the results above, our proposed method DeepPhenology has directly learnt the relationship between images of apple trees and phenology distributions from in-field manual counts using deep learning, which is a novel application. Because the output and the target are both normalized 8-class distribution distributions, the KL value was selected as the main index to measure the performance of DeepPhenology. To better understand KL values, three examples with different KL values from 0.01 to 0.22 have been shown in Figs. 9–11. The red bars in the right-hand images indicate the ground truth distribution whereas the blue bars indicate the predicted phenology distribution. A KL value of 0.01 is considered a near-perfect prediction as shown in Fig. 9. The distributions across all phenology stages are very similar to ground truth with only 1 or 2% difference. Fig. 10 shows an example prediction with a KL value of 0.10 from daytime images. As the first three stages have a very small distribution, we can ignore them and observe the performance from 'Pink'. We can see the order of the

predicted distribution from large to small was 'FB', 'PF', 'Pink', 'BB', and 'KB'. This is the same as the ground truth distribution except that 'FB' and 'PF' have the same amount whereas the proposed algorithm expected a slightly higher 'FB' than 'PF'. Besides this, 'Pink' and 'BB' were predicted to be lower but 'KB' was predicted to be higher. One possible reason for this could be that flowers have progressed since the time of manual counting as not a lot of 'Pink' clusters were seen from the image. The difference between the ground truth and predicted distribution for each stage was also less than 10%. Fig. 11 shows an example prediction with a KL value of 0.26 at the early stages of bloom. The major phenology predicted by the distribution fitted the ground truth with a similar percentage. However, the percentages for the next two stages had a very large difference, more than 15%, although the order of distribution from large to small fitted the ground truth. DeepPhenology expected more 'TC' and less 'HG' than the ground truth. This could be caused by the case that many 'HG' clusters were out of focus and on the other side of the quad.

DeepPhenology has shown reasonably good performance in terms of KL values as detailed in Section 4. Particularly, in the evaluation conducted with YOLOv5l as seen in Section 4.2, our proposed method has achieved better performance in phenology distribution estimation with the same number of training images. The main reason is that the object detection methods such as YOLOv51 would require a large dataset for training the detection of apple flowers because flower clusters at the same phenological stage can be represented differently when images are taken at different angles. It is also noted that our proposed method is stable when only 15% of the total images are used because the average KL value achieved of 0.235 in Section 4.2 is similar to that achieved of 0.236 in the validation and test sets of the full dataset in Table 4. As a result, with the same number of images, our proposed method is recommended for phenology estimation. Besides the evaluation of single image patches, the combination evaluation in Section 4.3 also shows that relative phenology distribution results can be simply averaged to form a row-level or block-level distribution.

To investigate the limitation of DeepPhenology, an additional investigation was done in the validation set. The relationship between KL values and the proportion of leaf clusters to total clusters has been summarized as shown in Fig. 12. It is noticeable that the performance becomes worse when there are more than 80% leaf clusters in quad images. More leaf clusters usually occlude flower clusters heavily, so this largely affects the estimation accuracy. The other reason could be because there were only a few images with more than 80% leaf clusters in the dataset during training so the model could not learn this case very well.

6. Conclusion

This work has proposed a robust and automatic method of estimating apple flower phenological stages based on deep learning, which also utilises a new type of ground truth. The main conclusions are as follows:

- A novel method named DeepPhenology was proposed to estimate apple flower phenology distribution directly from images based on the VGG-16 image classification model. An average KL value of 0.23 was consistently achieved among 12 cross validation rounds, which shows reasonably good and robust performance. The proposed method was also specially compared with one common approach used in previous literature, which is to utilize the normalized counts from the object detection results. However, the average KL value from YOLOv5l was only 0.511, which was almost double the average KL value of 0.235 from DeepPhenology.
- Relative phenology distribution results from single images based on DeepPhenology can be easily combined to build a row-level or block-level flower phenology distribution. The KL values for the combined phenology distribution based on 10 dates with 2 varieties are very low, less than 0.08, which is much smaller than the average KL values of 0.23 calculated per image on the validation set. The block-level distribution was particularly useful for the farmer as a first estimate, while the method provides the ability to examine the phenological stages at a sub-tree level.
- The DeepPhenology method was evaluated using manual ground truth data rather than labelling from 2D images, thereby avoiding the issues of compensating for occluded objects or classifying complex, densely packed shapes. This thorough evaluation gives greater confidence in the robustness of the solution when compared with methods which have been evaluated only on labelled images.
- The robustness of DeepPhenology has also been thoroughly evaluated on daytime and night-time images taken in an uncontrolled environment. The inference speed of DeepPhenology is slightly faster than YOLOv5l which is a current state-of-the-art algorithm, showing that the proposed method is feasible for commercial application.
- Future improvements will be made by collecting consecutive data in the coming season, which will allow the algorithm to be evaluated on farms with different trellis styles such as V-trellis and a wider range of varieties.

CRediT authorship contribution statement

Xu Wang (Annie): Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing - original draft, Writing - review & editing. **Julie Tang:** Data curation, Supervision, Validation, Writing - original draft, Writing - review & editing. **Mark Whitty:** Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Visualization, Writing - review & editing.

Declaration of Competing Interest

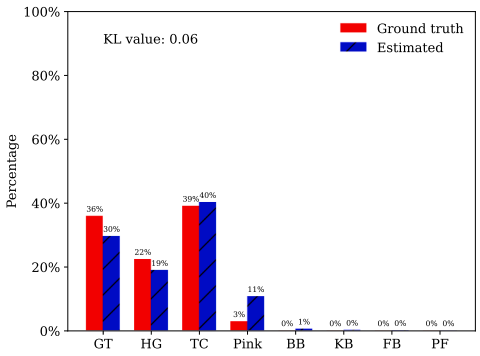
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

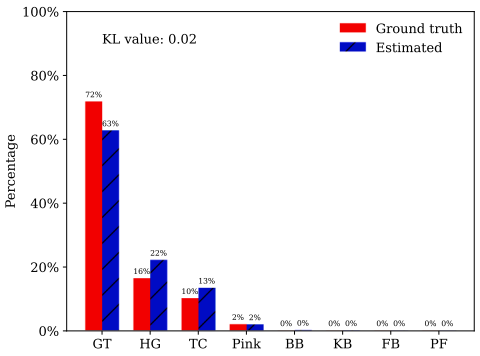
The authors would like to thank Angus Ross, Angus Hogan, Tom Brodie, and William McCarthy from SwarmFarm for building the hardware and collecting ground truth data along with Michelle Egan and the team from IK Caldwell. Thanks are due to Hiranya Jayakody and Valerie Mengying Hu for helping with data collection and to the team who tediously conducted manual labelling for the YOLOv5l comparison. This work was supported through AP160005, a project which has been funded by Hort Innovation, using the apple and pear research and development levies and contributions from the Australian Government. Hort Innovation is the grower-owned, not-for-profit research and development corporation for Australian horticulture.

Appendix A

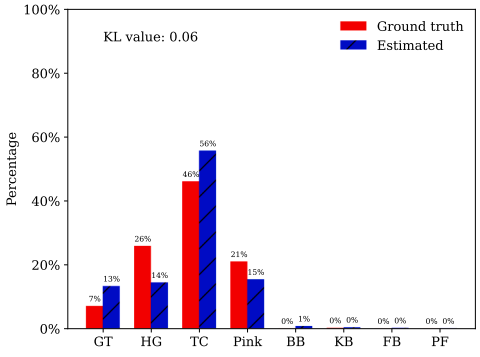
A.1. Combined phenology estimation results for all datasets



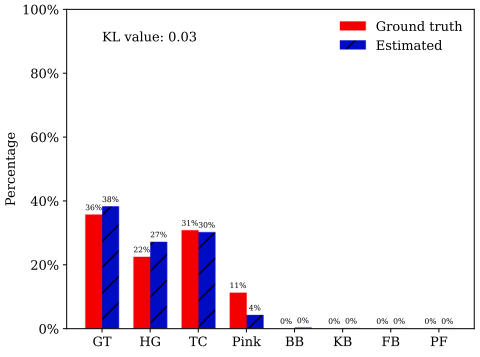
(a) 2019-09-20 Gala (Day & Night)



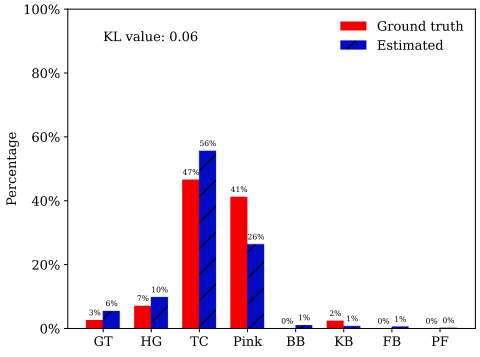
(b) 2019-09-20 Pink Lady (Day & Night)



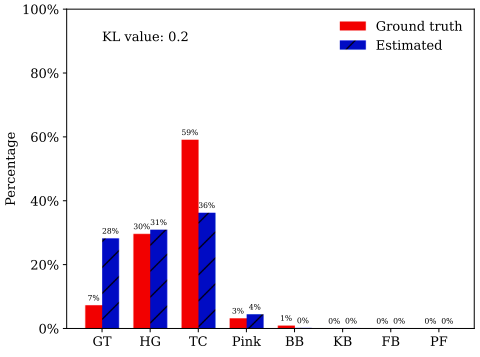
(c) 2019-09-23 Gala (Day)



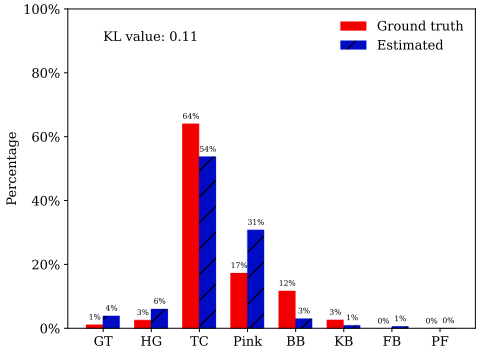
(d) 2019-09-23 Pink Lady (Day)



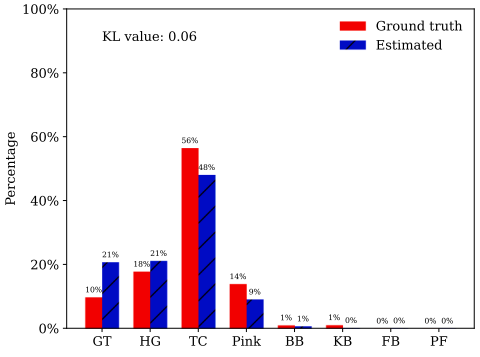
(e) 2019-09-25 Gala (Day)



(f) 2019-09-25 Pink Lady (Day)

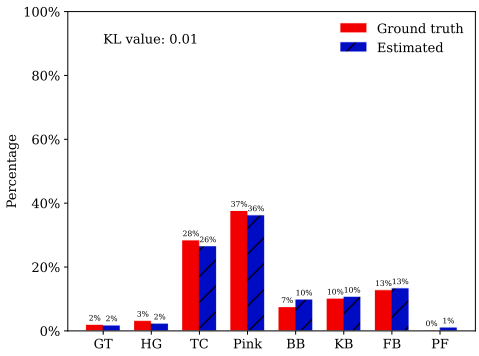


(g) 2019-09-26 Gala (Day)

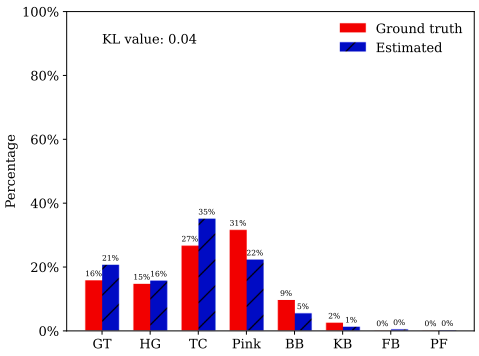


(h) 2019-09-26 Pink Lady (Day)

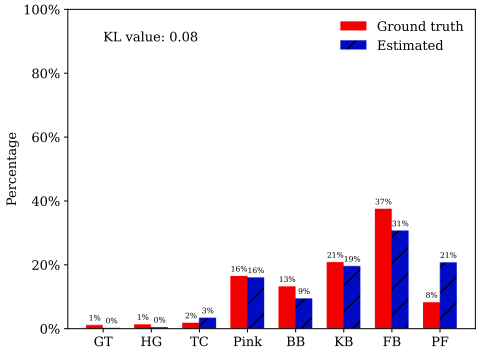
Fig. 13. Phenology distribution on all four study rows at different dates.



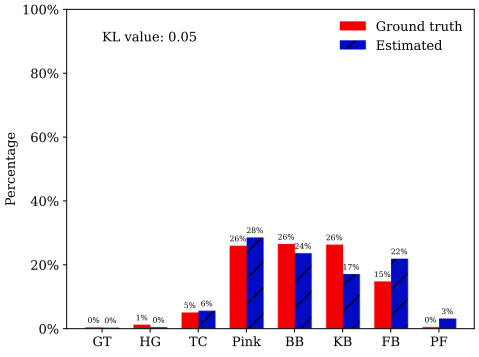
(i) 2019-09-30 Gala (Day & Night)



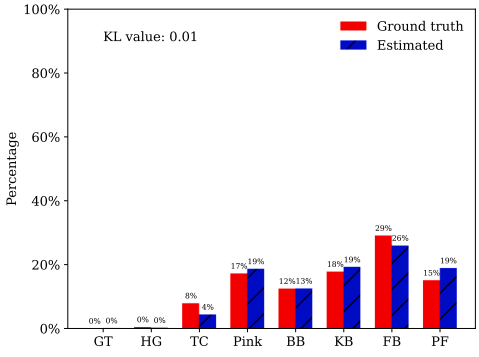
(j) 2019-09-30 Pink Lady (Day & Night)



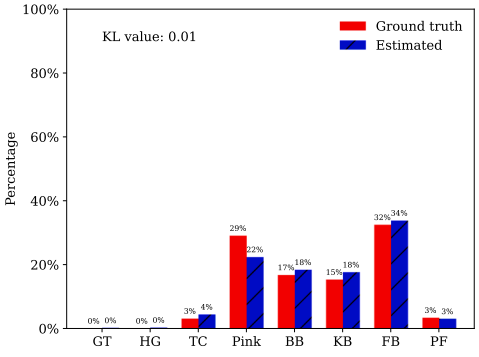
(k) 2019-10-04 Gala (Night)



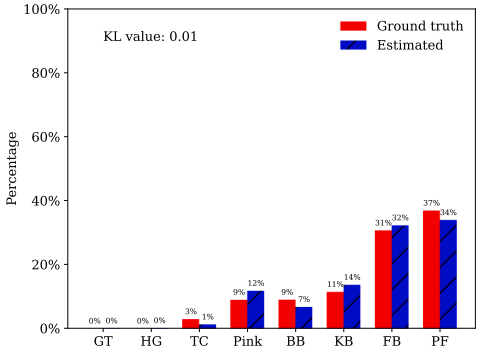
(l) 2019-10-04 Pink Lady (Night)



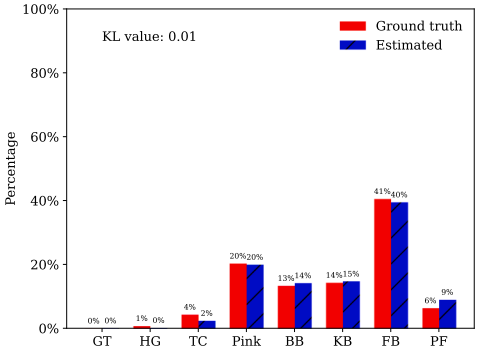
(m) 2019-10-05 Gala (Day & Night)



(n) 2019-10-05 Pink Lady (Day & Night)

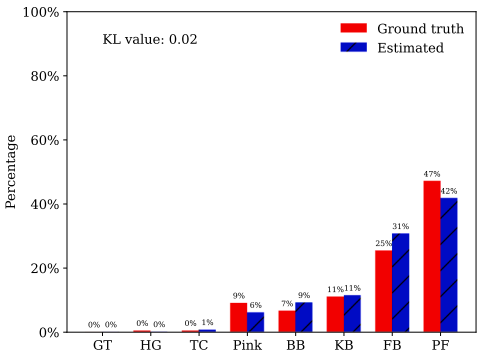


(o) 2019-10-06 Gala (Night)

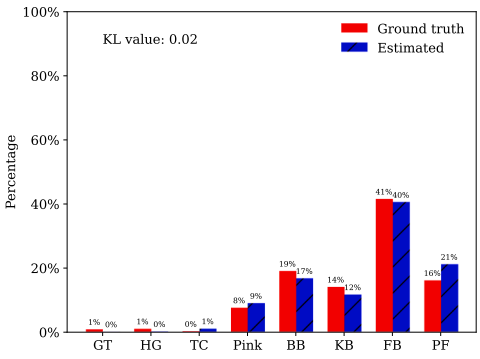


(p) 2019-10-06 Pink Lady (Night)

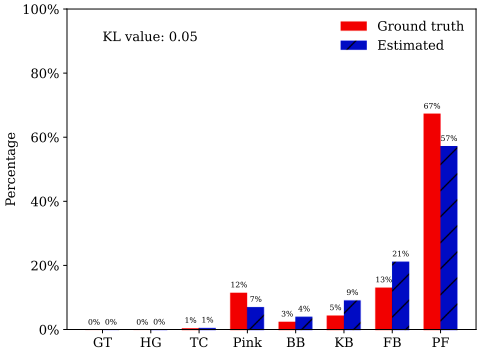
Fig. 13. (*continued*).



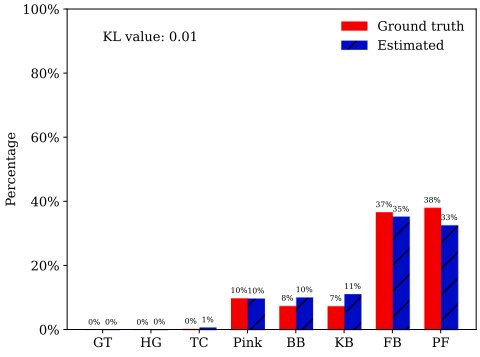
(q) 2019-10-07 Gala (Night)



(r) 2019-10-07 Pink Lady (Night)



(s) 2019-10-09 Gala (Day & Night)



(t) 2019-10-09 Pink Lady (Day & Night)

Fig. 13. (*continued*).

References

DeLong, C.N., Yoder, K.S., Cochran, A.E., Kilmer, S.W., Royston, W.S., Combs, L.D., Peck, G.M., 2018. Apple disease control and bloom-thinning effects by lime sulfur, regalia, and jms stylet-oil. Plant Health Progress 19 (2), 143–152. https://doi.org/10.1094/PHP-10-17-0065-RS.

Dias, P.A., Tabb, A., Medeiros, H., 2018. Multispecies fruit flower detection using a refined semantic segmentation network. IEEE Robot. Automat. Lett. 3 (4), 3003–3010. https://doi.org/10.1109/LRA.2018.2849498.

Dias, P.A., Tabb, A., Medeiros, H., 2018. Apple flower detection using deep convolutional networks. Comput. Ind. 99, 17–28. https://doi.org/10.1016/j.compind.2018.03.010.

Harel, B., Parmet, Y., Edan, Y., 2020. Maturity classification of sweet peppers using image datasets acquired in different times. Comput. Ind. 121, 103274. https://doi.org/10.1016/j.compind.2020.103274.

Chen, Y., Lee, W.S., Gan, H., Peres, N., Fraisse, C., Zhang, Y., He, Y., 2019. Strawberry yield prediction based on a deep neural network using high-resolution aerial orthoimages. Remote Sens 11 (13). https://doi.org/10.3390/rs11131584.

Hendrawan, Y., Amini, A., Maharani, D.M., Sutan, S.M., 2019. Intelligent non-invasive sensing method in identifying coconut (coco nucifera var. ebunea) ripeness using computer vision and artificial neural network. Pertanika J. Sci. Technol. 27 (3), 1317–1339.

Hu, M., Whitty, M., 2019. An evaluation of an apple canopy density mapping system for a variable-rate sprayer. IFAC-PapersOnLine 52 (30), 342–348. https://doi.org/10.1016/j.ifacol.2019.12.563.

Indriani, O.R., Kusuma, E.J., Sari, C.A., Rachmawanto, E.H., 2017. Tomatoes classification using k-nn based on glcm and hsv color space. IEEE, pp. 1–6. https://doi.org/10.1109/INNOCIT.2017.8319133.

Jocher, G., Stoken, A., Borovec, J., NanoCode012, ChristopherSTAN, Changyu, L., Laughing, Hogan, A., lorenzomammana, tkianai, yxNONG, AlexWang1900, Diaconu, L., Marc, wanghaoyang0106, ml5ah, Doug, Hatovix, J. Poznanski, L.Y., changyu98, Rai, P., Ferriday, R., Sullivan, T., Xinyu, W., YuriRibeiro, Claramunt, E.R., 2020. hopesala, pritul dave, yzchen, ultralytics/yolov5: v3.0 (Aug. 2020). doi:10.5281/zenodo.3983579.

Kipli, K., Zen, H., Sawawi, M., Mohamad Noor, M.S., Julai, N., Junaidi, N., Shafiq Mohd Razali, M.I., Chin, K.L., Wan Masra, S.M., 2018. Image processing mobile application for banana ripeness evaluation. In: 2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA), 2018, pp. 1–5. doi: 10.1109/ICASSDA.2018.8477600.

Koirala, A., Walsh, K.B., Wang, Z., McCarthy, C., 2019. Deep learning for real-time fruit detection and orchard fruit load estimation: Benchmarking of 'MangoYOLO. Precision Agriculture 20 (6), 1107–1135. https://doi.org/10.1007/s11119-019-09642-0.

Kon, T.M., Schupp, J.R., Yoder, K.S., Combs, L.D., Schupp, M.A., 2018. Comparison of chemical blossom thinners using 'golden delicious' and 'gala' pollen tube growth models as timing aids. HortScience horts 53 (8), 1143–1151. https://doi.org/10.21273/HORTSCI13087-18.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 22 (1), 79–86. https://doi.org/10.1214/aoms/1177729694.

Lim, J., Ahn, H.S., Nejati, M., Bell, J., Williams, H., MacDonald, B.A., 2020. Deep neural network based real-time kiwi fruit flower detection in an orchard environment. arXiv preprint arXiv: 2006.04343.

Müller, R., Kornblith, S., Hinton, G.E., 2019. When does label smoothing help?. In: Advances in Neural Information Processing Systems, pp. 4694–4703.

Nasiri, A., Taheri-Garavand, A., Zhang, Y.-D., 2019. Image-based deep learning automated sorting of date fruit. Postharvest Biol. Technol. 153, 133–141. https://doi.org/10.1016/j.postharvbio.2019.04.003.

Peck, G.M., Combs, L.D., DeLong, C., Yoder, K.S., 2016. Precision apple flower thinning using organically approved chemicals. Acta Hortic 1137, 47–52. https://doi.org/10.17660/ActaHortic.2016.1137.7.

Pereira, L.F.S., Barbon Jr., S., Valous, N.A., Barbin, D.F., 2018. Predicting the ripening of papaya fruit with digital imaging and random forests. Comput. Electron. Agric. 145, 76–82. https://doi.org/10.1016/j.compag.2017.12.029.

Santos, T.T., de Souza, L.L., dos Santos, A.A., Avila, S., 2020. Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association. Comput. Electron. Agric. 170, 105247. https://doi.org/10.1016/j.compag.2020.105247.

Septiarini, A., Hamdani, H., Hatta, H.R., Kasim, A.A., 2019. Image-based processing for ripeness classification of oil palm fruit. In: 2019 5th International Conference on Science in Information Technology (ICSITech), pp. 23–26.

Tan, K., Lee, W.S., Gan, H., Wang, S., 2018. Recognising blueberry fruit of different maturity using histogram oriented gradients and colour features in outdoor scenes. Biosyst. Eng. 176, 59–72. https://doi.org/10.1016/j.biosystemseng.2018.08.011.

Tu, S., Xue, Y., Zheng, C., Qi, Y., Wan, H., Mao, L., 2018. Detection of passion fruits and maturity classification using red-green-blue depth images. Biosyst. Eng. 175, 156–167. https://doi.org/10.1016/j.biosystemseng.2018.09.004.

Wan, P., Toudeshki, A., Tan, H., Ehsani, R., 2018. A methodology for fresh tomato maturity detection using computer vision. Comput. Electron. Agric. 146, 43–50. https://doi.org/10.1016/j.compag.2018.01.011.

Wang, X.A., Tang, J., Whitty, M., 2020. Side-view apple flower mapping using edge-based fully convolutional networks for variable rate chemical thinning. Comput. Electron. Agric. 178, 105673. https://doi.org/10.1016/j.compag.2020.105673.

Wu, D., Lv, S., Jiang, M., Song, H., 2020. Using channel pruning-based yolo v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments. Comput. Electron. Agric. 178, 105742. https://doi.org/10.1016/j.compag.2020.105742.

Yu, Y., Zhang, K., Yang, L., Zhang, D., 2019. Fruit detection for strawberry harvesting robot in non-structural environment based on mask-rcnn. Comput. Electron. Agric. 163, 104846. https://doi.org/10.1016/j.compag.2019.06.001.