Agricultural and Forest Meteorology 264 (2019) 225–234

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

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

Ear density estimation from high resolution RGB imagery using deep learning technique

Simon Madec[a,*], Xiuliang Jin[a], Hao Lu[b], Benoit De Solan[c], Shouyang Liu[a], Florent Duyme[c], Emmanuelle Heritier[c], Frédéric Baret[a]

[a] INRA, UMR EMMAH, UMT-CAPTE, Avignon, France
[b] National Key Laboratory of Science and Technology on Multi-Spectral Information Processing, School of Automation, Huazhong University of Science and Technology, Wuhan, 430074, China
[c] ARVALIS, Institut du végétal, Avignon, France

ARTICLE INFO

Keywords:
Wheat ear density
Object detection
Object counting
Convolutional neural networks
Phenotyping
Broad-sense heritability

ABSTRACT

Wheat ear density estimation is an appealing trait for plant breeders. Current manual counting is tedious and inefficient. In this study we investigated the potential of convolutional neural networks (CNNs) to provide accurate ear density using nadir high spatial resolution RGB images. Two different approaches were investigated, either using the Faster-RCNN state-of-the-art object detector or with the TasselNet local count regression network. Both approaches performed very well (rRMSE ≈ 6%) when applied over the same conditions as those prevailing for the calibration of the models. However, Faster-RCNN was more robust when applied to a dataset acquired at a later stage with ears and background showing a different aspect because of the higher maturity of the plants. Optimal spatial resolution for Faster-RCNN was around 0.3 mm allowing to acquire RGB images from a UAV platform for high-throughput phenotyping of large experiments. Comparison of the estimated ear density with in-situ manual counting shows reasonable agreement considering the relatively small sampling area used for both methods. Faster-RCNN and in-situ counting had high and similar heritability ($H^2 \approx 85\%$), demonstrating that ear density derived from high resolution RGB imagery could replace the traditional counting method.

1. Introduction

Wheat ear density in wheat crops is associated with components of crop yield related to plant population and tiller number per plant, but is a difficult and tedious trait for breeders to efficiently measure. Further, it is prone to sampling errors when the sampling area is small due to limited human resources. Computer vision approaches provide a potential solution to increase the throughput as well as the spatial representativeness, leading potentially to an improved accuracy. A number of studies based on high spatial resolution imaging systems applied to plant phenotyping under field conditions have received much attention in recent years (Li et al., 2014). Both ground-based and aerial platform (Araus and Cairns, 2014; Deery et al., 2014; Tardieu et al., 2015) have been exploited to image the microplots with a spatial resolution spanning within few centimeters to a fraction of millimeter.

Because of the typical size of wheat ears and the possible occlusions between them, a spatial resolution of few millimeters is required to identify non-ambiguously the ears. Therefore, most studies focused on high-resolution RGB images on which a high pass filter and morphological operators were applied (Journaux et al., 2010) (Fernandez-Gallego et al., 2018). Those methods provide promising results on small datasets. However, these types of algorithms may fail when applied to images acquired under different conditions and for different development stages: the change in illumination conditions, the occlusions, the variability of ear aspect due to genotype including the presence or absence of awns, the flowering status, the variability of the background and the image quality make the scalability of this phenotyping task challenging.

The advances in computation capacity along with the availability of very large collections of labelled images have fostered enhanced machine learning methods based on convolutional neural networks (CNNs) in the field of computer vision (Hinton and Salakhutdinov, 2006; LeCun et al., 2015). CNNs are currently achieving impressive performances for image classification (Singh et al., 2016; Krizhevsky et al., 2012). Because the number of label images required to train CNN models from scratch is important, pre-trained are often used as a starting point. Further, pretrained models generally improved the resulting accuracy (Mohanty et al., 2016) and limit overfitting issues (Yosinski et al.,

* Corresponding author.
E-mail address: simon.madec@inra.fr (S. Madec).

https://doi.org/10.1016/j.agrformet.2018.10.013
Received 18 May 2018; Received in revised form 15 October 2018; Accepted 22 October 2018
Available online 30 October 2018
0168-1923/ © 2018 Elsevier B.V. All rights reserved.

2014). Several network architectures have proven their effectiveness over benchmark computer vision database like Alexnet (Singh et al., 2016), VGG (Simonyan and Zisserman, 2014) and more recently residual network with inception layers like Inception-ResNet (Szegedy et al., 2016). These models are then fine-tuned (Dauphin et al., 2012) on a small training dataset specific to a particular classification task (Yosinski et al., 2014; Donahue et al., 2014; Sharif Razavian et al., 2014).

Some of these methods have already been applied to plant phenotyping. CNN models have demonstrated to be effective for discriminating features for wheat plants including highly accurate identification of ears (Pound et al., 2017) in glasshouse condition. Similar studies have shown that CNNs outperform classic hand-crafted feature descriptors and offer an alternative approach for classification problems (Pound et al., 2017; Allen et al., 2005; Madec et al., 2017). The detection algorithm needs to identify and localize each ear in the image. When the overlap between identified objects is a common pattern as for crowded scenes, counting by regression networks was recently demonstrated to be a relevant alternative (Huang et al., 2016; Hosang et al., 2017; The Open Images dataset, 2018): the TasselNet model was proposed for counting maize tassels (Lu et al., 2017). Tasselnet is based on a CNN with a regression output layer. The local counts regressed from individual sub-images are merged to provide a count map for the whole image. Likewise, a deep residual model with a regression output was used to count the number of wheat plants at emergence (Salton and McGill, 1983). The emergence counting is achieved through a two-stage process: segmenting wheat plants and regressing the counts from small image patches. This presents an alternative way to tackle occluded wheat plants, while the precision of the system is also affected by the segmentation algorithm used.

The main objective of this study is to evaluate deep learning approaches for high-throughput wheat ear counting under field conditions. For this purpose, two types of CNN architectures will be investigated: (i) local object detection and (ii) counting by regression. The influence of the spatial resolution of the RGB on the model performance will be analyzed to select the optimal resolution. Finally, the ear density estimated from the RGB images will be compared with the in-situ visual ear counting and the broad sense heritability is then quantified to evaluate the suitability of the proposed method for field phenotyping.

2. Material and methods

2.1. Data collection and labelling

2.1.1. Experimental site

The study area is a wheat field phenotyping platform located in Gréoux les Bains (France, 43.7° latitude North, 5.8° longitude East). Wheat was sown on November 3rd 2016 with a row spacing of 17.5 cm and a density of 300 seeds·m^{-2}. A trial of 120 microplots of 2.0 m width by 10 m long was considered. Half of the microplots was irrigated (called WW) while the other part was subjected to water stress (called WS). The 20 contrasting genotypes were replicated three times both in the WW and WS modalities and organized as an alpha-design.

A crop water balance model (Allen et al., 2005) has been used to estimate the water stress during the whole growing season. It used the measured soil water holding capacity of 143 mm to compute the actual evapotranspiration at the daily time from the rainfall and the potential evapotranspiration. The difference between the actual and maximum evapotranspiration values corresponds to the daily crop water deficit that was cumulated from emergence up to maturity (Fig. 1). This indicated that the water deficit for the WS modality started after the ear emergence stage (stage Z59). The irrigation on the WW modality were starting after this date.

2.1.2. Ground measurements of ear density

In each microplot, the ear density was measured on June the 7th 2017 after the flowering stage over three segments of 1 m length by two adjacent rows which represent a sampling area of 1.05 m^2. The first two rows located at the border of the microplots were not considered in the sampling to minimize border effects.

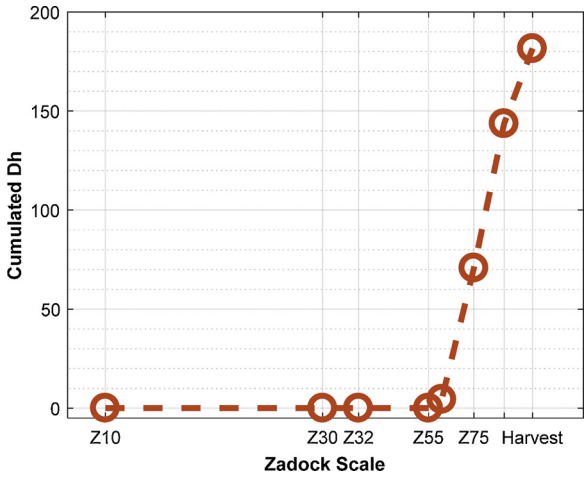

Fig. 1. Cumulated Dh for different zadocks development stages. Dh is the difference between the actual and maximum evapostranspiration values.

2.1.3. Canopy height

The height is required to define the footprint of the image and to compute the ear density by dividing the number of ears by the size of its footprint. The height was measured with a LiDAR fixed on a fully automated robot called "phénomobile" (Madec et al., 2017). The uncertainties associated to the height estimated by the LiDAR were few centimeters. More details can be found in (Madec et al., 2017).

2.1.4. Image acquisition and labelling

A Sony ILCE-6000 digital camera with a 6000 × 4000 pixels was fixed on a boom. The RGB images were taken from the nadir view direction at 2.9 m distance to the ground. For each microplot two images were recorded. The measurements were completed on June 2nd and the 16th 2017. 60 mm and 50 mm focal lengths were used respectively on June the 2nd and the 16th. This resulted in a ground sampling distance between 0.010 - 0.016 cm/pixel and a footprint area of individual images between 0.25m^2 and 0.56m^2 depending on the height of the wheat and the focal length used.

The ears were interactively labelled in all the images of the first experiment (June 2nd) resulting into 240 images (20 genotypes x 3 replicates x 2 modalities x 2 images). Between 80 and 170 ears were contained in each image. The LABELIMG (darrenl, 2017) graphical image annotation tool was used to draw the bounding boxes around each identified ear in the images (Fig. 2). The bounding boxes contain all the pixels of the ears, except when the bounding box would have to be made too large to include the awns. If possible, the boxes also contain a small portion of the stem. When comparing the results of the identification by one of the model developed, we discovered that few ears were forgotten by the operator in interactive label process. The images were thus reprocessed interactively with greater care. Finally, a total of 30,729 ears were identified after the second ear label round.

The second experiment (June the 16th) was only used to evaluate the scalability of the models when applied to another stage with different illumination conditions and camera focal length: no interactive labelling was made for this experiment.

2.1.5. Data preparation

It was not possible to train the model with the original 6000 × 4000 pixels images because of GPU memory limitation. The maximum image size acceptable for the available computer configuration () was

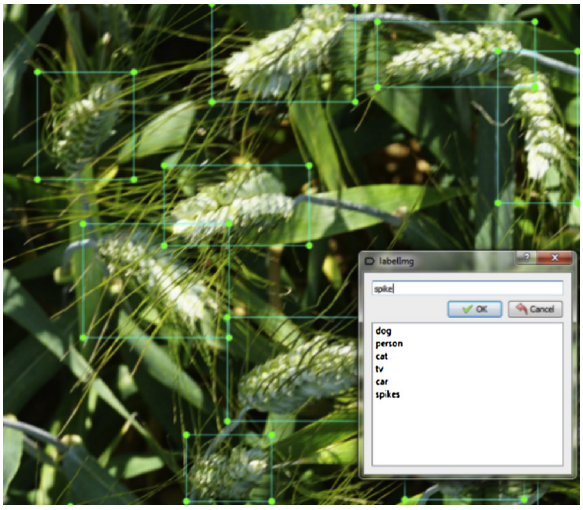

Fig. 2. Example of bounding boxes interactively drawn using the labelimg software.

Table 1
Characteristics of the several models considered in this study.

Model	Approach	Resolution factor	Resolution (mm)	Size sub-image (pixels)	Number of sub-images used for training/validation
#1	Faster-RCNN	1	0.13	500	56994/24426
#2	Faster-RCNN	2	0.26	500	12270/5258
#3	Faster-RCNN	3	0.39	500	5782/2478
#4	Faster-RCNN	4	0.52	250	12270/5258
#5	Faster-RCNN	4	0.52	500	2478/1062
#6	Faster-RCNN	6	0.78	250	5782/2478
#7	Faster-RCNN	6	0.78	500	990/424
#8	Faster-RCNN	8	1.04	250	2478/1062
#9	TasselNet	8	1.04	256	Sliding window

500 × 500 pixels. The original images where therefore split into multiple sub-images while keeping 50% overlap between the sub-images. The overlap allows to minimize problems observed on the borders when ears are only partially contained in the sub-image. Note that using smaller sub-images increases the size of the training dataset (Pound et al., 2016). To investigate the influence of the spatial resolution, several training datasets were generated by resampling the original images by a factor of 2, 3, 4, 6 and 8 (Table 1) using a bi-linear aggregation function. Note that the larger resampling factors corresponding to a degraded spatial resolution will correspond to larger footprints of the 500 × 500 pixels sub-images and therefore a limited number of training data set. To investigate this trade-off, sub-image size of 250 × 250 were also considered, which increased by a factor 4 the training dataset at the expense of more border effects. Note that no specific data-augmentation was applied to the training data base. Regarding the TasselNet method sub-image of a size of 256 were used, these sub-images were further down sample by a factor of 8.

2.2. Data processing

2.2.1. Object detection using Faster-RCNN

Object detection techniques searches first in the sub-image potential candidates. An object proposal method is thus required. A number of object proposal methods have been reviewed and compared (Hosang et al., 2015) for general object detection using the convolutional features of the full sub-image network. The Region Proposal Network (RPN) generates first a dense grid of anchor regions (candidate bounding boxes) with specified sizes and aspect ratios over the input sub-image. An anchor is assigned as positive/negative if its intersection over union (IoU) ratio with the ground truth object is greater/lower than a relatively large/small overlap threshold. The RPN made of a shallow CNN predicts a score for each anchor, which measures its probability to contain an ear. One of the advantages of this approach is that the model learns features of the background, thus removing negative location to the classification step.

The TensorFlow implementation of Faster-RCNN by the object detection API (Huang et al., 2016) was used. The RPN branch is inserted between the conv4 and conv5 blocks. The Inception-Resnet-V2 model was used here because it achieves the best accuracy among current object detectors (Huang et al., 2016). An anchor is set at each location considered by the convolution maps of the RPN layer. A set of 12 anchors with different sizes and aspect ratios were assigned at each location, following the default setting. Anchors were considered containing an ear if the IOU between their bounding boxes and those of the labelled ears were between 0.6 and 1.0. Conversely, it was considered as background if the IOU with label ears was lower than 0.175. When the IOU was in between 0.175 and 0.6, the anchors were no more considered. These hyper-parameters were corresponding to standard values. The number of proposed anchors per sub-image was fixed to 300 which is consistent with the maximum expected number of ears in a sub-image. The batch size was fixed to 1 because it saves the computation time and memory requirement while marginally impacting the performances (results not shown for the sake of brevity). Each bounding box was associated with a score value. A score threshold of 0.5 was used to decide whether a bounding box will be considered as an ear or not. To limit overlap between bounding boxes containing the same ear, an IOU threshold of 0.6 was used to select only one of the two bounding boxes (Hosang et al., 2017). The model was pretrained on the COCO dataset (The Open Images dataset, 2018). It contains 0.33 million images with 1.5 million of object instances belonging to 80 object categories. The model was finally fine trained with a learning rate of 0.0003 and a momentum of 0.9.

The results on the sub-images were then merged to count the ears over the full original image. Because of the 50% overlap between the sub-images an ear was generally detected in more than one sub-image. An overlap ratio was computed for each bounding box. It was computed as the intersection area between the two bounding boxes divided by the area of the smaller bounding box. If this ratio was larger than 0.85, the smaller bounding box was deleted.

2.2.2. Counting by regression using TasselNet

TasselNet is a recent regression-based counting approach. TasselNet learns a mapping from local visual characteristics to local image counts. The image is processed using a sliding window. The global image count is computed by summing the counts over the set of local windows. Compared to Faster R-CNN, learning TasselNet only needs dotted annotations (the center of each bounding box). Following the suggestions from (Lu et al., 2017), the Alex-like CNN model with local counts as the regression target based on L1 loss function were used here. Further, since TasselNet allows to work on relatively low-resolution images, the original image was down sampled to 1/8 of its original size and 32 × 32 pixels sub-images were considered, corresponding to 256 × 256 pixels sub-images in the original spatial resolution (Table 1). We refer readers to (Lu et al., 2017) for further details.

2.2.3. Evaluation metrics

The training and validation datasets were populated with different genotypes: 14 genotypes (168 images) out of the 20 were randomly selected for training the models. The six remaining genotypes (72 images) were used for the validation. This will allow identification of possible overfitting in the training process.

A predicted bounding box is considered correct (true positive, TP) if it overlaps more than the IOU threshold with a labelled bounding box.

Otherwise the predicted bounding box is considered as false positive (FP). When the labelled bounding box have an IOU with a predicted bounding box lower than the threshold value, it is considered as false negative (FN). The standard IOU threshold value of 0.5 was used. The precision and recall are then computed (Eq. (1)):

$$Precision = \frac{TP}{FP + TP}, Recall = \frac{TP}{FN + TP} \quad (1)$$

The score associated to each bounding box allows evaluating the trade-off between false positive and false negative. The average precision (AP@0.5IOU) (Salton and McGill, 1983) was used to quantify the detection performances. The standard average precision metrics, AP@0.5IOU, is the area under the precision-recall curve obtained for different bounding box scores. The AP@0.5IOU balances the precision and recall performances terms that may be strongly correlated. AP@0.5IOU varies between 0 (TP = 0) to 1 (FN = 0)

The ear counting performances were quantified using several metrics: root mean squared error (RMSE), the relative RMSE (rRMSE), the mean absolute error (MAE), the Bias (BIAS) and the coefficient of determination (R^2):

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{n}(t_k - c_k)^2} \quad (2)$$

$$rRMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{n}\left(\frac{t_k - c_k}{t_k}\right)^2} \quad (3)$$

$$BIAS = \frac{1}{N}\sum_{k=1}^{n}(t_k - c_k) \quad (4)$$

$$MAE = \frac{1}{N}\sum_{k=1}^{n}|t_k - c_k| \quad (5)$$

$$R^2 = 1 - \frac{\sum_{k=1}^{n}(t_k - c_k)^2}{\sum_{k=1}^{n}(t_k - \bar{t}_k)^2} \quad (6)$$

Where N denotes the number of test images, t_k and c_k are respectively the reference and estimated counts for image k, and $\bar{t}_k$ is the mean reference count.

3. Results and discussion

3.1. A resolution around 0.3 mm is needed for best performance with faster-rcnn

The time required to train each model was around 1 h (4000 iterations run on a NVIDIA GTX 1080Ti). The model performance (AP@0.5IOU computed on the validation dataset) was computed at several stages during the training process for the several combinations of spatial resolution and sub-window size (Table 2). This allows for evaluation of the quality of the training process. Results showed that AP@0.5IOU was generally converging rapidly towards a maximum value

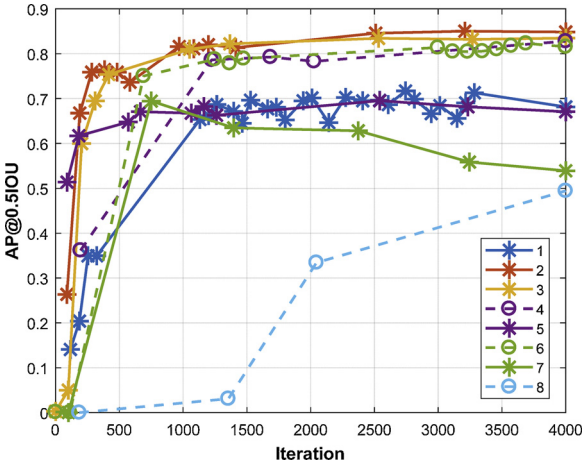

Fig. 3. AP@0.5IOU as a function of the number of iterations during the training process. The several curves correspond to the models presented in Table 2.

(Fig. 3): after 1000 iterations, most of the models reached an AP@0.5IOU close to the maximum one. This is explained by the fact that the model was pretrained and initialized using the COCO dataset. No overfitting characterized by a decrease of the AP@0.5IOU after reaching a maximum value was observed (Fig. 3), except for the model #8 that was trained on a to small number of sub-images. The maximum values of AP@0.5IOU found here (Table 2) were higher than those reported in other studies based also on the COCO dataset (Huang et al., 2016). This improvement in the performances observed in our study was mostly explained by the larger size of the training dataset used and the relatively lower complexity of the ear detection problem as compared to the categories considered in (Huang et al., 2016).

When using the original spatial resolution of the images (model #1), borders with ears partially in the sub-image were often observed (Fig. 3). This may explain why the performances degraded as compared to slightly coarser resolution (Fig. 3, Table 2). Note that the sub-image size was limited by the GPU memory. Further, the object size (Table 2) observed for model #1 with the highest spatial resolution is much larger than the typical size of objects considered in standard convolutional networks (in between 100 and 250 pixels) (Huang et al., 2016). This may pose difficulties to handle these large objects for the first step of the algorithm where regions are proposed (RPN).

The spatial resolution showed a marginal impact on the AP@0.5IOU values that kept around AP@0.5IOU = 0.9 for most models having more than 3000 sub-images used for the training, with the exception of the original resolution (model #1) wich showed strong border effects and too large bounding boxes as already discussed (Table 2). The AP@0.5IOU was therefore mostly influenced by the number of sub-images used in the training process. For the same spatial resolution (models #4 and #5 with 0.52 mm resolution, and models #6 and #7 with 0.78 mm resolution) the AP@0.5IOU was always higher when the number of

Table 2
Performances of the Faster-RCNN models considered evaluated over the 72 validation images.

Dataset	Resolution (mm)	Average object size (pixels)	Size of sub-image (pixels)	Number of sub-images	AP@0.5IOU after 4000 iterations	ear count per image	
						R^2	rRMSE
#1	0.13	221.6	500	56994	0.70	0.73	11%
#2	0.26	110.8	500	12270	0.85	0.91	5.3%
#3	0.39	73.9	500	5782	0.83	0.85	5.4%
#4	0.52	55.4	250	12270	0.83	0.83	11.2%
#5	0.52	55.4	500	2478	0.67	0.87	24.7%
#6	0.78	36.9	250	5782	0.82	0.75	11.2%
#7	0.78	36.9	500	990	0.54	0.33	38.5%
#8	1.04	27.7	250	2478	0.49	0.62	30.3%

images used for training was larger (Table 2). Even with a spatial resolution degraded down to 0.78 mm (model #6), AP@0.5IOU were only marginally decreasing as compared to 0.39 mm resolution (model #3) when the training data set is large enough (5782 sub-images). For the 0.78 mm resolution, the average size of the ear bounding box is 37 pixels, which is consistent with other studies (A closer look, 2018) and with the size considered for the proposed objects. For coarser spatial resolution, the number of available sub-images used for the training will be too small to provide robust performances. Further, even by increasing the size of the training dataset with additional labelled images, performances are expected to decrease because Faster-RCNN is known to have difficulties with small objects (A closer look, 2018).

For a more detailed evaluation of the performances of the several models presented in Table 2, the R^2 and rRMSE of the ear counting estimation for each sub-image were also computed. Those two metrics were generally in agreement with the AP@0.5IOU (Table 2). However, relatively high rRMSE and small R^2 were observed for models #4 and #6. Visual inspection of the resulting estimated bounding boxes showed that too many boxes were assigned for the same ear which is not properly considered by the AP@0.5IOU metrics. This problem corresponds to poorer efficiency of the RPN step when the size of the sub-images (250 × 250) was too small. It was thus recommended to use sub-image size larger than 250 × 250 pixels. However, it is possible that manipulating concurrently other hyper-parameters such as the number of maximum proposed bounding boxes will partly solve this limitation.

Best performances were observed for dataset #2 with a sub-image size of 500 × 500 pixels and a spatial resolution around 0.26 mm (Fig. 3). This dataset was used in the validation part of this project (Fig. 4)

A total of 8097 ears were detected with model #2 applied to the validation dataset, with 1.5% false positive (commission) and 2.9% false negative (omission). Closer inspection of the false positive cases showed that a significant part (around 40%) corresponded to actual ears that were not identified during the interactive labelling process. The Faster-RCNN model achieved thus a better ear detection than humans when properly trained. However, the model failed to detect most of occluded ears with poor lighting conditions that were also largely missed by the human labelling: the model was obviously not trained for these situations. The false positive cases were also generally associated with a lower confidence score (Fig. 5). The influence of this score will be further discussed in later in this paper.

When part of a stem was visible, model #2 was more easily detecting the ears: the stem carried therefore useful information for ear recognition. However, this situation is not the dominant one for most of the genotypes at the early stages since ears are mostly vertical and observed from nadir (Fig. 6). The model also failed to detect very large ears (Fig. 7, left). Lower AP@0.5IOU was computed when the model was facing ears with no awn. Further, the model had more difficulties with bounding boxes that had an aspect ratio different from one. The IOU ratio with the reference labelled boxes was generally smaller for ears with no awn or for bounding boxes with aspect ratio different from one. This later problem can be improved by adding anchors with a larger range of size and aspect ratio.

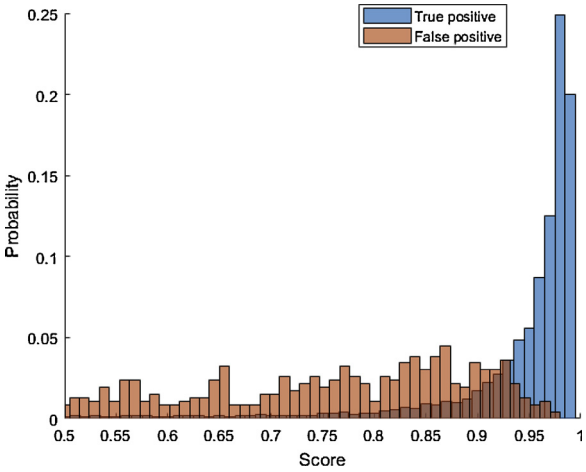

Fig. 5. Distribution of the associated scores for the true positive (blue) and the false positive (red) bounding boxes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Faster-RCNN is more robust than TasselNet

The ear counting based on TasselNet required very few hyper-parameters. The standard values proposed by (Lu et al., 2017) were used here. The comparison between TasselNet and Faster-RCNN is based on the ear density estimated from the images belonging to the validation dataset. In fact, TasselNet did not identify and locate the ears and so it is not possible to compute a confusion matrix from which the AP@0.5IOU could be derived. Further, the lack of localization step in TasselNet prevents the opportunity for exploration of other potential traits exploration at the ear level such as detecting the presence of awns, measuring the size and shapes of the ears and quantifying the flowering status. Nervertheless, TasselNet presented the advantage to identify each ear using a single point as compared to the more complex label using bounding boxes as in the Faster-RCNN approach.

Results showed good performances for both methods as evaluated over the validation dataset, with very small bias (< 5 ears) and a better rRMSE for Faster-RCNN ($\approx 5\%$) (Fig. 8). This result was expected in cases of non-crowded scenes with little overlap between objects, which was the case for ears in this study (Fig. 6): less than 1% of the interactively label bounding boxes had an IOU > 0.5. TasselNet was more efficient for relatively low spatial resolution images to evaluate the density of small object instances (< 30 pixels). TasselNet seemed therefore not exploiting all the detailed texture information required for Faster-RCNN to identify individual ears: the degraded spatial resolution (1.04 mm) of the images used for TasselNet provided better performances as compared to Faster-RCNN applied to the same spatial

Fig. 4. Example of the sub-windows used for model-#1 (left: original resolution (0.13 mm) and subimage size = 500 × 500), model #2 (middle: resolution degraded by a factor of 2 (0.26 mm) and subimage size = 500 × 500), and model #5 (right: resolution degraded by a factor of 4 (0.52 mm) and subimage size = 500 × 500).

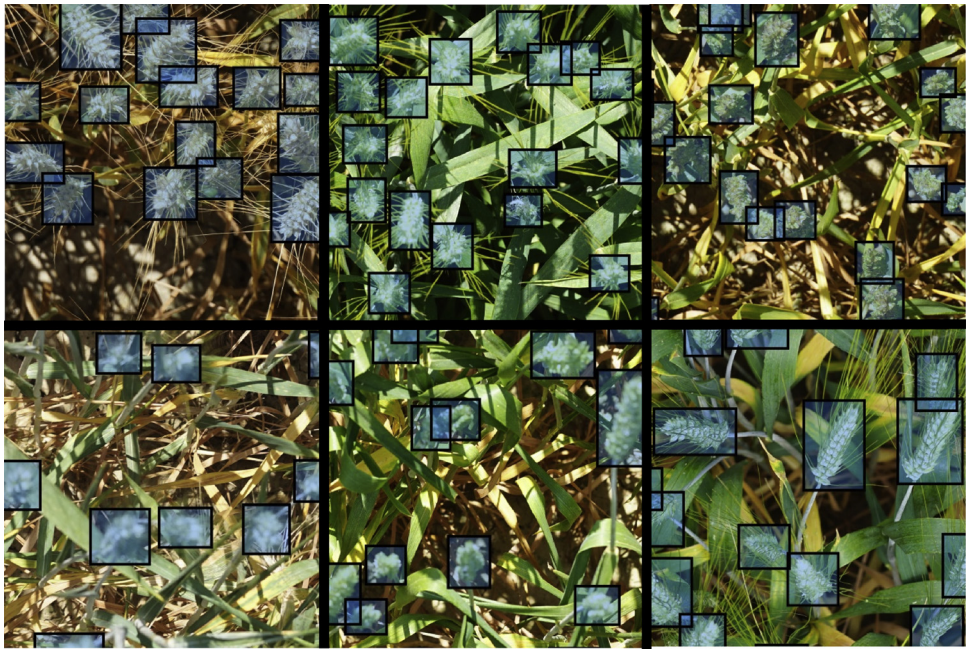

Fig. 6. Examples of sub-images (500 × 500 pixels) with ears detected by model #2.

resolution (model #8 in Table 2).

Since the number of ears is not expected to change after the flowering stage, estimations for the first date were compared with those of the second date for which no ears were labelled. This allowed evaluation of the scalability for the second date of the models calibrated on the first date. Because the images taken over each microplot were not located exactly at the same place between the two dates, performances were based on the average ear density of the microplot. It was computed from the ear count of the two images taken over each microplot, divided by their footprint area defined by plant height and camera field of view. Results showed that the application to the second date of the Faster-RCNN previously trained over the first date agreed very well with ear density estimates for the first date (Fig. 9) with however a slight underestimation of the ear density that will be further investigated in the next section.

Conversely, TasselNet showed large discrepancies between the ear density estimation of the two dates with a strong underestimation (Fig. 9). This appeared mostly related to the senescence state of the microplot, TasselNet having difficulties to detect senescent ears over senescent leaves (Fig. 10). Further, ears from the second date presented different visual aspects as compared to the first date, with in addition a slight change in the spatial resolution due to the use of a 60 mm and 50 mm focal lengths camera respectively for the first and second dates (Fig. 10). TasselNet failed to capture ears from the second date and thus generalized poorly in new scenes. The better scalability of the Faster-RCNN model may be due the fact that it was already pretrained to detect millions of object instances, exploiting more the gray-scale image pattern than the TasselNet model that seemed to be much more sensitive to the color of the objects. More investigations should be carried out to improve TasselNet scalability based on a larger training dataset with a significant fraction of images over senescent crops, or to simply transform the RGB images into gray-scale images.

Because of the limits of the TasselNet model highlighted previously, focus will be put on the Faster-RCNN model#2 trained over the first date. The slight degradation of performances of the Faster-RCNN model for the second date was further investigated. The RMSE between the ear density estimated from the RGB images and the ground measurements were computed for a range of score threshold values used to decide if a bounding box is considered containing an ear or not (Fig. 11). Results showed that the RMSE decreases with the score threshold values down

Fig. 7. Illustration of problems encountered with model #2. On the left, example of False negative (omission). On the right, example of false positive (commission). All the images have the same resolution.

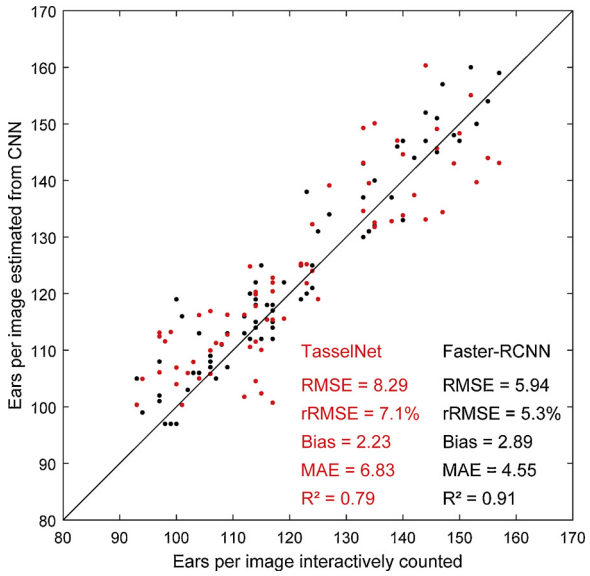

Fig. 8. Comparison between the number of ears in each image visually labelling and that estimated using either the Faster-RCNN (model #2 black dots) or the Tasselnet (model #9 red dots). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

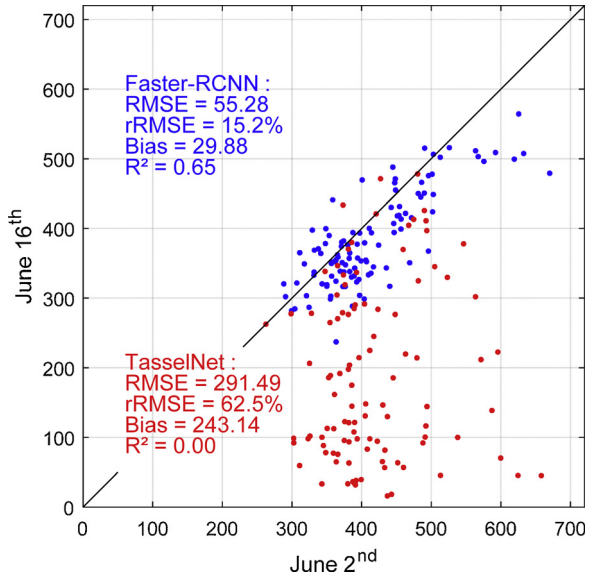

Fig. 9. Comparison between the average ear density estimated for the first (June 2nd) and the second (June 16th) dates using Faster-RCNN (blue dots) and TasselNet (red dots) models trained over the first date only. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

to a minimum around 0.7 for the first date: increasing the score threshold value limits the fraction of false positives. After this minimum value, the RMSE increased with the score since the fraction of false negative will increase. The situation is slightly different for the second date: the RMSE increased continuously with the score threshold value. Ears for the second date were slightly different from the first date used to train the model. The score threshold should therefore be relaxed to prevent rejecting too many candidate ears that were slightly different from the first date. The score threshold value initially used (0.5) appeared thus to be optimal when considering the two measurement dates: both curves were crossing for a score threshold value of 0.5 (Fig. 11).

3.3. Ear density estimates are highly heritable

The broad sense heritability (H^2) quantifying the repeatability of the ear density estimation was computed as the ratio between the genotypic variances to the total one (Holland et al., 2002). A linear mixed-effects statistical model was applied on each date to quantify the genetic variance. The 'lm4' R package was applied to our alpha plan experimental design (Bates et al., 2014). The soil water holding capacity (S) that was carefully documented was used as fixed effect in the model that writes (random terms are underlined) as:

$$Y = \mu + S + \underline{G} + \underline{L} + \underline{C} + \underline{L:C} + \varepsilon \quad (7)$$

Where Y is the ear density, $\underline{G}$ is the random effect of the genotypes, $\underline{L}$ and $\underline{C}$ are respectively, the random effects for lines and column in the alpha plan, and $\underline{L:C}$ is the random sub-block effect. μ is the fixed intercept term and ε the random residual error. Since the genotypes may express differently depending on the environmental conditions, the heritability was computed independently for the two modalities.

The high heritability values (≈85%) for the estimated ear density was observed for both measurement dates and the two modalities (Table 3). This was partly explained by the fact that contrasted genotypes were used with significant differences in the tillering capacity. The heritability is higher for the first date as compared to the second one. This can be attributed by the fact that the models were trained with images from the first date. The heritability of the WS modality is slightly higher than that observed on the WW modality. The heritability associated to the ground measurements of the ear density were in the same order of magnitude as those estimated from the RGB imagery and the faster-RCNN model. The heritability of the WW modality (H^2 = 80%) is lower than that of the WS modality (H^2 = 91%) in agreement with RGB imagery estimates. However, the heritability of the WW modality is lower than that provided by the Faster-RCNN model. This point will be investigated in the next section.

3.4. Faster-RCNN was more reliable than ground measured ear density

The ear density of the WS and WW modalities were expected to be very similar since the water stress was mostly appearing after the ear emergence stage (Fig. 1) when all ears have already emerged from the stems. The estimated ear density averaged over the three replicates were thus compared between the two modalities. The same was done for the ground measured ear densities. The best coefficient of determination (R^2) was observed with the images from the first date and the output from Faster – RCNN model #2 (Fig. 12): the ear density between the two modalities were very similar as expected with almost no bias (Bias = 0.6 ears/m^2). The same was observed over the second date of RGB image acquisition (R^2 = 0.78; Bias = 20.6 ears/m^2). Conversely, ground measured ear densities were higher in the WS modality as compared to those of the WW modality. This was not expected and should result from larger uncertainties in the ground measurements. This may also explain the low heritability of ground measured ear density found for the WW modality Table 3.

The ear density estimated with Faster-RCNN was finally compared with the ground measurements. The ear density estimated with model #2 was in relatively good agreement with the ground measurements for the WS modality and the first date (June 2nd) of RGB images acquisition. (Table 4 and Fig. 13). The scatter of points observed might be partly attributed to the relatively small sampling size used for the ground observations (1.05 m^2) and for the RGB images (about 0.6 m^2 for the first date and 1.0 m^2 for the second date). The spatial representativeness was therefore limited to get an accurate comparison between the two types of ear density that were not measured at the same place over each microplot. Increasing the number of RGB images taken on each plot would improve this aspect which should not be a major issue considering the high-throughput associated both to the image acquisition and data processing. The ear density over the WW

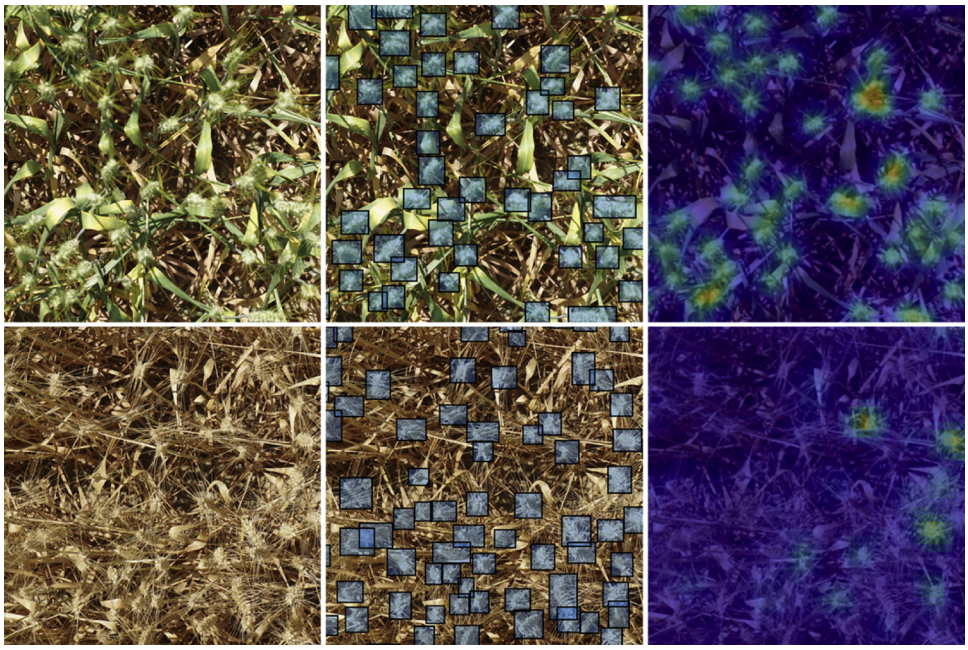

Fig. 10. Example of two images selected from the same microplot on the first and second date over which Faster-RCNN and TasselNet models were applied. The models were trained on the first date and the genotype was in the validation dataset. The top images corresponded to microplot showing little senescence, while the bottom images corresponded to microplot showing almost complete senescence. From left to right: original RGB image, detection by Faster-RCNN model #2, density maps estimated using the TasselNet model.

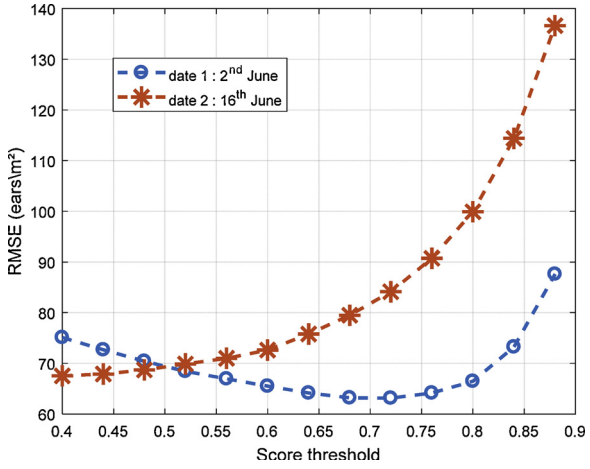

Fig. 11. RMSE of the estimated ear density as a function of the score threshold value.

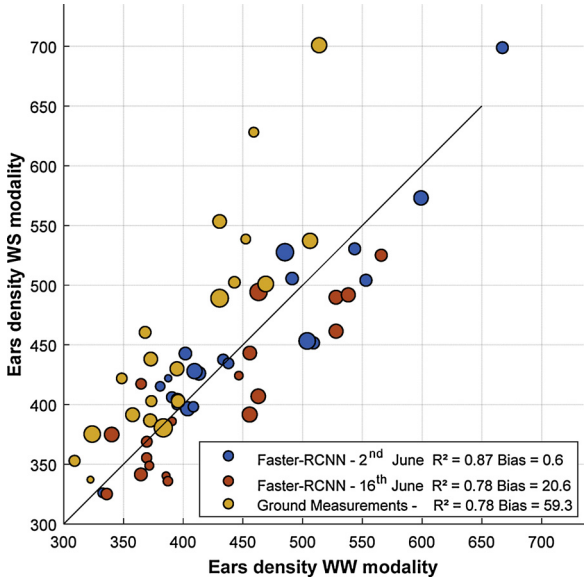

Fig. 12. Comparison between ear density estimated in the WW modality and in the WS modality (The size of the circles represented the variability of the measurements for the different replicates).

Table 3
Broad sense heritability (H^2) computed for the ear density for the ground measurements and the estimates for the two dates considered. The H^2 values are computed for the two modalities.

H^2 (%)	Date	WW	WS	ALL
Ground Measurements	June 7th	79.8	91.4	66.3
Estimates from first date	June 2nd	86.9	88.5	86.5
Estimates from second date	June 16th	82.2	82.8	76.3

modality showed a significant degradation of the agreement between estimated and measured values. A systematic under estimation of the ear density from the ground measurement in the WW modality was observed, probably due to the uncertainties attached to the ground measurements already outlined.

4. Conclusions

The main objective of this study was to evaluate the efficiency of deep learning approaches to estimate the ear density in wheat phenotyping trials using high resolution RGB images acquired at nadir.

Table 4
Performances of ear density estimated from the Faster-RCNN trained over the first date (June 2nd) and applied to the two dates for both WW (irrigated) and WS (water stress) modalities.

	June 2nd			June 16th		
	WW	WS	All	WW	WS	All
RMSE (ear/m²)	82.0	53.0	68.7	62.1	77.4	69.4
rRMSE (%)	16.4	12.1	14.4	15.1	18.9	16.9
Bias (ear/m²)	55.2	2.73	25.8	24.62	52.49	10.05
R^2	0.52	0.70	0.53	0.46	0.62	0.37

Considering the challenge of managing the large diversity in the ear and background aspects due to genotypic specificities as well as dates of observations, our results were promising (rRMSE = 5.3%) for the

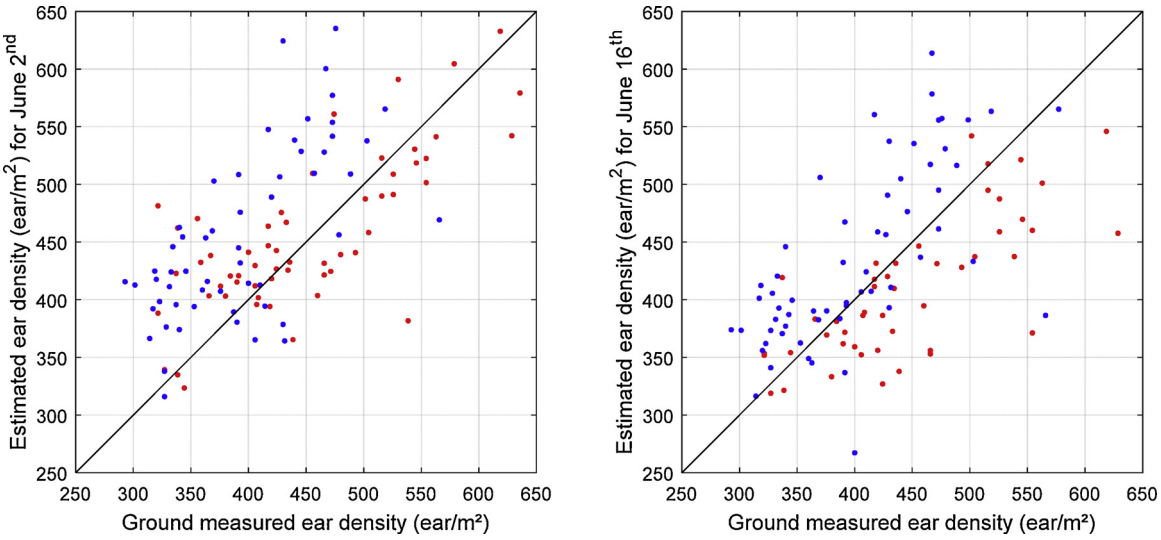

Fig. 13. Comparison between ear density estimated from the RGB images and the ground-measurements for the two modalities (red : WS; blue : WW). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Faster-RCNN model. The Faster-RCNN model were much better than the first round of interactive labeling: many ears were missed by the operator when label the ears on the screen. This forced us to improve the interactive label by reanalyzing the training and validation datasets. Faster-RCNN model was demonstrated to be more robust than counting by regression using the TasselNet model. Counting by regression such as with the TasselNet model should be more efficient for crowded scenes with objects that overlap frequently, which was not the case for ears as observed from nadir. Faster-RCNN beneficiated from the transfer learning approach based on a model pretrained on the COCO dataset. However, the gain in robustness for the Faster-RCNN model comes at the expense of larger computation requirements using GPU resources. Nevertheless, models based on object detection present the advantage of providing the basis for estimating additional traits for the ears including spatial distribution between rows, presence of awns, size, inclination or color that could be useful for breeders.

Results showed that the broad sense heritability of ear density estimates from RGB images was high and close to that computed from the direct in situ measurements. However, the ear density computed from the RGB images based on Faster-RCNN model showed only a fair agreement with the ear density measured in the field (rRMSE ≈ 15%), particularly for the WW modality that was suspected to had larger uncertainties attached to the ground counting. An improved match between the RGB estimates and direct counting in the field is expected mostly either (i) by working on the same samples which is not straightforward to achieve, or better by (ii) increasing the sampling size of both the ground sampling as well as the sampling area covered by the images. This is easy to achieve with the RGB imagery by capturing more images over each microplot, while it is expensive for the in-situ measurements.

Faster-RCNN model was demonstrated to have optimal performances for images with spatial resolution between 0.26 and 0.39 mm. Higher spatial resolution corresponded to too large bounding boxes as well as increased border effects difficult to manage. For broader resolution, the loss of textural information degraded the identification performances. Therefore, an optimal resolution around 0.3 mm would allow to use UAV observations for covering large phenotyping experiments as already demonstrated by (Jin et al., 2017) and get a very high-throughput method. Further, UAV observations cover the whole microplot, allowing a large sampling area, thus increasing the precision and heritability. Further, uncertainties attached to the knowledge of the area used to compute the density will be negligible. This was not the case in this study where the relatively small footprint of the images forced to get accurate estimation of the distance between the camera and the ear layer, with possible representativeness of the row effect since the width of the images was not necessarily a multiple of the distances between rows.

The size and the diversity of the training dataset is critical to obtain good estimation performances. A Faster-RCNN model trained over one date was demonstrated to apply well on another date with only a small degradation of performances even if the ears and the background aspects were quite different. However, improved performances are expected by applying data augmentation to artificially increase the size of the training dataset that was demonstrated to be critical, as well as the diversity of measurement conditions including orientation, adjustment of the white balance, of the spatial resolution and of the sharpness. The concept of domain adaptation should also help to tackle the domain and dataset discrepancy problem (Chen et al., 2018). Nevertheless, performance can be further improved with availability of large datasets of carefully labelled images. For this reason, we offer to the community the labelled dataset used in this study that is freely accessible at: https://github.com/simonMadec where 30,729 ears were identified in 240 images over 20 contrasting genotypes grown in two environmental conditions.

Funding

This study was supported by “Programme d’investissement d’Avenir” PHENOME (ANR-11-INBS-012) and Breedwheat (ANR-10-BTR-03) with participation of France Agrimer and “Fonds de Soutien à l’Obtention Végétale”. The work was completed within the UMT-CAPTE funded by the French ministry of Agriculture.

Acknowledgement

We thank very much Olivier Moulin, Guillaume Meloux and Magali Camous from the Arvalis experimental station in Gréoux for their kind support during the measurements.

References

A closer look, 2018. A closer look. Small Object Detection in Faster R-CNN - IEEE Conference Publication [Internet]. [cited 2018 Jan 11]. Available from: http://ieeexplore.ieee.org/abstract/document/8019550/.

Allen, R.G., Clemmens, A.J., Burt, C.M., Solomon, K., O’Halloran, T., 2005. Prediction accuracy for projectwide evapotranspiration using crop coefficients and reference evapotranspiration. J. Irrig. Drain. Eng. 131, 24–36.

Araus, J.L., Cairns, J.E., 2014. Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant. Sci. 19, 52–61.
Bates D., Mächler M., Bolker B., Walker S. Fitting linear mixed-effects models using lme4. ArXiv Prepr ArXiv14065823 [Internet]. 2014; Available from: https://arxiv.org/abs/1406.5823.
Chen Y., Li W., Sakaridis C., Dai D., Van Gool L. Domain adaptive faster r-cnn for object detection in the wild. ArXiv Prepr ArXiv180303243. 2018.
darrenl. labelImg: :metal: labelImg is a graphical image annotation tool and label object bounding boxes in images [Internet]. 2017. Available from: https://github.com/tzutalin/labelImg.
Dauphin, G.M.Y., Glorot, X., Rifai, S., Bengio, Y., Goodfellow, I., Lavoie, E., et al., 2012. Unsupervised and transfer learning challenge: a deep learning approach. PMLR [Internet] 97–110. [cited 2017 May 30] Available from: http://proceedings.mlr.press/v27/mesnil12a.html.
Deery, D., Jimenez-Berni, J., Jones, H., Sirault, X., Furbank, R., 2014. Proximal remote sensing buggies and potential applications for Field-based phenotyping. Agronomy 4, 349–379.
Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., et al., 2014. DeCAF: A deep convolutional activation feature for generic visual recognition. Icml [Internet] 647–655. Available from: http://www.jmlr.org/proceedings/papers/v32/donahue14.pdf.
Fernandez-Gallego, J.A., Kefauver, S.C., Gutiérrez, N.A., Nieto-Taladriz, M.T., Araus, J.L., 2018. Wheat ear counting in-field conditions: high throughput and low-cost approach using RGB images. Plant. Methods. 14, 22.
Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science. 313, 504–507.
Holland, J.B., Nyquist, W.E., Cervantes-Martínez, C.T., 2002. Estimating and interpreting heritability for plant breeding: an update. In: Janick, J. (Ed.), Plant Breed Rev [Internet]. John Wiley & Sons, Inc, pp. 9–112. Available from: http://onlinelibrary.wiley.com/doi/10.1002/9780470650202.ch2/summary.
Hosang, J., Benenson, R., Dollár, P., Schiele, B., 2015. What makes for effective detection proposals? IEEE Trans Pattern Anal. Mach. Intell. PAMI 2015 [Internet] Available from: http://ieeexplore.ieee.org/abstract/document/7182356/.
Hosang J., Benenson R., Schiele B. Learning non-maximum suppression. ArXiv Prepr ArXiv170502950 [Internet]. 2017; Available from: https://arxiv.org/abs/1705.02950.
Huang J., Rathod V., Sun C., Zhu M., Korattikara A., Fathi A., et al. Speed/accuracy trade-offs for modern convolutional object detectors. ArXiv161110012 Cs [Internet]. 2016; Available from: http://arxiv.org/abs/1611.10012https://github.com/tensorflow/models/tree/master/research/object_detection.
Jin, X., Liu, S., Baret, F., Hemerlé, M., Comar, A., 2017. Estimates of plant density of wheat crops at emergence from very low altitude UAV imagery. Remote. Sens. Environ. 198, 105–114.
Journaux, L., Marin, A., Cointault, F., Miteran, J., 2010. Fourier Filtering for Wheat Detection in a Context of Yield Prediction. [cited 2017 May 29]. Available from:. CIGR. http://www.csbe-scgab.ca/docs/meetings/2010/CSBE101090.pdf.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. [Internet] 1097–1105. Available from: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature. 521, 436–444.
Li, L., Zhang, Q., Huang, D., 2014. A review of imaging techniques for plant phenotyping. Sensors 14, 20078–20111.
Lu, H., Cao, Z., Xiao, Y., Zhuang, B., Shen, C., 2017. TasselNet: counting maize tassels in the wild via local counts regression network. Plant. Methods. 13, 79.
Madec, S., Baret, F., de Solan, B., Thomas, S., Dutartre, D., Jezequel, S., et al., 2017. High-throughput phenotyping of plant height: comparing unmanned aerial vehicles and Ground LiDAR estimates. Front. Plant. Sci. [Internet] 8 [cited 2018 Jan 22]. Available from: https://www.frontiersin.org/articles/10.3389/fpls.2017.02002/full.
Mohanty, S.P., Hughes, D.P., Salathé, M., 2016. Using deep learning for image-based plant disease detection. Front. Plant. Sci. [Internet] 7 Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5032846/.
Pound, M.P., Burgess, A.J., Wilson, M.H., Atkinson, J.A., Griffiths, M., Jackson, A.S., et al., 2016. Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. bioRxiv., 053033.
Pound, M.P., Atkinson, J.A., Wells, D.M., Pridmore, T.P., French, A.P., 2017. Deep learning for multi-task plant phenotyping. bioRxiv. 204552.
Salton, G., McGill, M.J., 1983. Introduction to Modern Information Retrieval [Internet]. Available from:. McGraw-Hill, New York. https://trove.nla.gov.au/work/19430022.
Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S., 2014. CNN features off-the-shelf: an astounding baseline for recognition. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop [Internet] 806–813. Available from: http://www.cv-foundation.org/openaccess/content_cvpr_workshops_2014/W15/html/Razavian_CNN_Features_Off-the-Shelf_2014_CVPR_paper.html.
Simonyan K., Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv14091556 Cs [Internet]. 2014; Available from: http://arxiv.org/abs/1409.1556.
Singh, A., Ganapathysubramanian, B., Singh, A.K., Sarkar, S., 2016. Machine learning for high-throughput stress phenotyping in plants. Trends Plant. Sci. 21, 110–124.
Szegedy C., Ioffe S., Vanhoucke V., Alemi A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. ArXiv160207261 Cs [Internet]. 2016; Available from: http://arxiv.org/abs/1602.07261.
Tardieu, F., Le Gouis, J., Lucas, P., Baret, F., Neveu, P., Pommier, C., et al., 2015. PHENOME: French plant phenomic center. EPPN Plant. Phenotyping Symp. np.
The Open Images dataset, 2018. The Open Images Dataset [Internet]. Openimages. Available from:. https://github.com/openimages/dataset.
Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks? Adv. Neural Inf. Process. Syst. [Internet] 3320–3328. Available from: http://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks.